Informative and Consistent Correspondence Mining for Cross-Domain Weakly Supervised Object Detection

Luwei Hou[*1,3], Yu Zhang[*†3], Kui Fu[1], Jia Li[†1,2]
[1]State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China
[2]Peng Cheng Laboratory, Shenzhen, China [3]SenseTime Research
{houluwei,zhangyu1}@sensetime.com {kuifu,jiali}@buaa.edu.cn

Abstract

*Cross-domain weakly supervised object detection aims to adapt object-level knowledge from a fully labeled source domain dataset (*i.e.*, with object bounding boxes) to train object detectors for target domains that are weakly labeled (*i.e.*, with image-level tags). Instead of domain-level distribution matching, as popularly adopted in the literature, we propose to learn pixel-wise cross-domain correspondences for more precise knowledge transfer. It is realized through a novel cross-domain co-attention scheme trained as region competition. In this scheme, the cross-domain correspondence module seeks for informative features on the target domain image, which if warped to the source domain image, could best explain its annotations. Meanwhile, a collaborative mask generator competes to mask out the relevant target image region to make the remaining features uninformative. Such competitive learning strives to correlate the full foreground in cross-domain image pairs, revealing the accurate object extent in target domain. To alleviate the ambiguity of inter-domain correspondence learning, a domain-cycle consistency regularizer is further proposed to leverage the more reliable intra-domain correspondence. The proposed approach achieves consistent improvements over existing approaches by a considerable margin, demonstrated by the experiments on various datasets.*

1. Introduction

With decades of efforts made in improving feature representations [7, 20, 13], learning architectures [9, 28, 36] and large-scale datasets [8, 31, 23], performance of modern object detectors has been raised to a brand new level. Nevertheless, generalizing existing detection models to novel unseen domains still remains an issue, as the models are often biased to dataset-specific patterns rather than data-invariant "common knowledge". It often takes huge efforts to collect the well-annotated training data of the novel target domains from scratch to feed the data-hungry modern architectures.

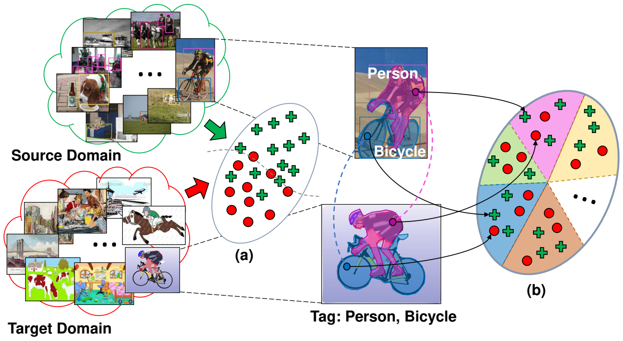

Figure 1. Motivation of the proposed approach to address cross-domain object detection. (a) Conventional approaches project images from different domains into a unified feature space such that a domain classifier cannot easily separate them. (b) Assuming the target domain is weakly labeled, we explicitly establish pixel-wise correspondence among the semantic regions of cross-domain images and form semantic clusters in feature space for accurate, localized domain transfer. Best viewed in color.

Recently this issue is addressed via *unsupervised domain adaptation* [3, 10], that transfers task knowledge from a richly annotated source domain to poorly annotated target domains. Yet, domain discrepancy lie in various frequency patterns [45], visual styles [21] and class distributions [40], making it nontrivial to align knowledge transfer to the desired task objective. Heuristic strategies (*e.g.* adaptive weighting [48, 42, 18], foreground mining [14, 43, 2], self-supervision [30, 19]) were proposed to guide the domain transfer to focus more on detecting foreground objects. However, with the lack of direct knowledge about the

[*] Equal contribution. Part of this work is done during Luwei's internship with SenseTime Research.

[†] Correspondence should be addressed to Jia Li (jiali@buaa.edu.cn) and Yu Zhang (zhangyulb@gmail.com).

object distributions in target domain, it is difficult to achieve accurate object-level domain transfer.

To address this issue, a relaxed setting was proposed by Inoue *et al.* [15], that the target domain images are assumed to be tagged with semantic labels. This novel setting, called *cross-domain weakly supervised object detection*, gives access to direct knowledge of the target domain with minimal annotation cost. As such, foreground/background ambiguity of object localization within target domain was greatly reduced. However, knowledge adaptation is still conducted globally at domain level, while local feature alignment and knowledge transfer that could be mined from the weak annotations in the target domain, is less explored.

In this paper, we embrace the weakly supervised setting of [15] and propose a novel approach for cross-domain object detection. As shown in Fig. 1, different from previous works that perform knowledge transfer at domain-level, our approach explicitly establishes pixel-wise semantic correspondences in each pair of cross-domain images for accurate local knowledge transfer. As the target domain is weakly labeled, the core idea is to divide each image into semantic clusters in weakly supervised manner that can well explain the region annotations of source domain image, under the cross-domain warping indicated by their correspondences. Specifically, a cross-domain co-attention module is trained to seek for informative features on the target domain image so as to well reconstruct the annotations of the corresponding source domain image. At the same time, a jointly trained mask generator competes to mask out the relevant target image region, to make the remaining correspondences uninformative. Such competitive process facilitates the correlation of full extent of the underlying objects across domains. To reduce the ambiguity of cross-domain matching, we further propose a novel domain-cycle consistency regularizer to leverage intra-domain correspondence as robust self-supervision. The proposed approach sets new state-of-the-art results, improves over previous works consistently by $4\% \sim 6\%$ in mean average precision on 3 datasets.

We highlight the following contributions. 1) We propose a novel approach formulated as region competition, capable of establishing explicit pixel-wise semantic correspondences across domains and enabling accurate local knowledge transfer. 2) We introduce the cycle consistency regularizer to cross-domain object detection, which provides robust and cost-free self-supervision by leveraging both inter and intra-domain cues. 3) We conduct extensive experiments to evaluate the proposed modules, showing notably and consistently improved results on three benchmarks.

2. Related Works

Object detection owns a long story in computer vision research since its emerge. It has benefited greatly by recent advance in deep learning architectures [28, 36, 22, 12]. Their success also attributes to the development of large-scale, manually annotated datasets [8, 31, 23]. However, manually annotating a large number of images usually costs a lot and is not scalable. This largely hinders the application of the state-of-the-art detectors in unseen domains or datasets. There are two ideas proposed to solve this issue.

Weakly supervised object detection is a potential solution that assumes labour-friendly, yet weaker forms of manual annotations (*e.g.* class labels), while inferring potential object locations using the "collective knowledge" provided by the joint distribution of weak labels in the dataset. The crucial challenge is to recover the full object extent, not only the discriminative parts indicated by classification activations [50, 33]. It was achieved by hiding the current discriminative parts then seeking for next ones, while the hiding strategy was set randomly [35], via attentional dropout [5], adversarial dropout [29], or complementary learning [46]. Full localization was also guided by foreground expansion [47], weighted region voting [4, 37], network optimization [34] and geometrical priors [26]. However, without knowledge to "general objects", localizing the full extent could be difficult due to the diversity of object deformations, poses, viewpoints and background appearance.

Cross-domain object detection is an alternate solution that assumes large-scale manually annotated datasets available and studies how to generalize the knowledge of source domain to novel datasets. It was achieved with adversarial feature alignment [3, 38], so that a domain classifier cannot easily distinguish between the features coming from different domains. Yet, the distribution gap of images from different domains is typically multi-modal, lying in various aspects such as frequency patterns [45], visual styles [21] and class distributions [40]. Regularizations were proposed to guide the alignment process, via adaptive region weighting [2, 18, 49], foreground mining [53, 43, 30, 14], heuristic class sampling [19, 42, 32]. In [15], a weakly supervised setting of cross-domain object detection is proposed, with novel datasets and baselines introduced. In this work we attempt to provide deeper investigations to this approach.

3. Proposed Approach

3.1. Overview

For the proposed task we assume two datasets $\mathcal{S}$ and $\mathcal{T}$, from the source and target domains, respectively. For each image in $\mathcal{S}$, objects coming from a predefined set of semantic classes $\mathbb{C}$ are fully annotated with bounding boxes and the class labels. We are interested in training detectors that generalize to $\mathcal{T}$, where only presence of object classes from $\mathbb{C}$ are annotated for each image.

Characteristically we follow recent detector adaptation pipeline [3, 2] that applies the two-stage Faster-RCNN [28] detector and adversarial learning, which is summarized in

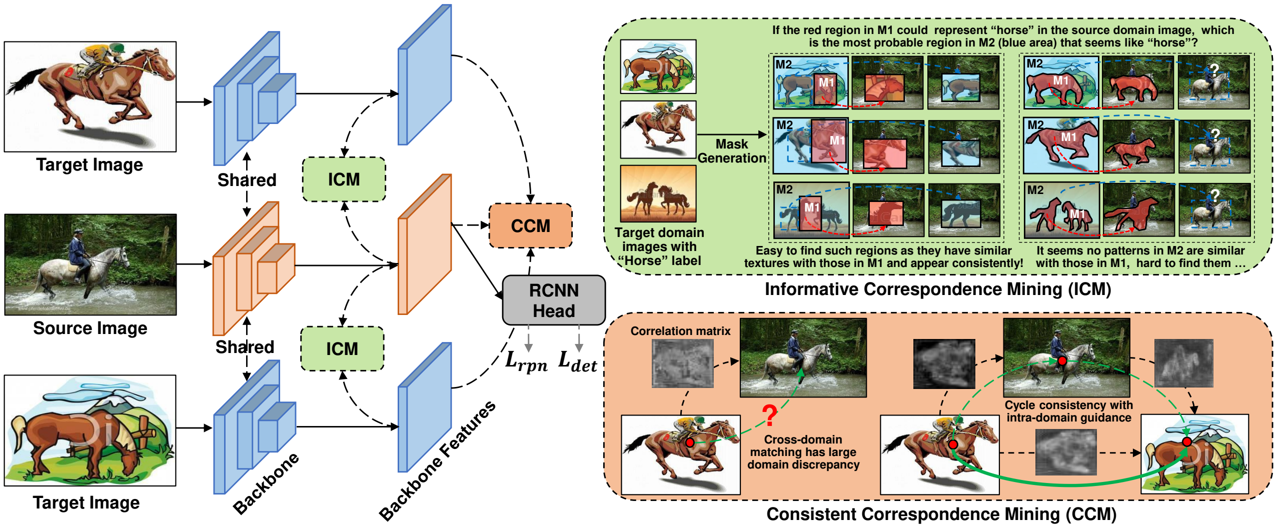
Target Image
Shared
ICM
Source Image
Shared
ICM
Target Image
Backbone
Backbone Features
CCM
RCNN Head
L_{rpn} L_{det}
If the red region in M1 could represent "horse" in the source domain image, which is the most probable region in M2 (blue area) that seems like "horse"?
Mask Generation
Target domain images with "Horse" label
M2
M1
?
Easy to find such regions as they have similar textures with those in M1 and appear consistently!
It seems no patterns in M2 are similar with those in M1, hard to find them …
Informative Correspondence Mining (ICM)
Correlation matrix
?
Cross-domain matching has large domain discrepancy
Cycle consistency with intra-domain guidance
Consistent Correspondence Mining (CCM)

Figure 2. Pipeline of our approach. During training we sample two images from the target domain and one from the source domain. This triplet is passed to the backbone of Faster-RCNN, yielding feature maps to be fed into the Region Proposal Networks (RPN). We enforce regularizations on these feature maps with Informative Correspondence Mining and Consistent Correspondence Mining. These feature maps are then forwarded to the RCNN head (details omitted due to the limit of space) for classification and bounding box regression.

Fig. 2. At each training step, pairs of source and target domain images (denoted with $\mathcal{I}_\mathcal{S}$ and $\mathcal{I}_\mathcal{T}$) are sampled. Since $\mathcal{I}_\mathcal{S}$ is fully annotated, the associated semantic labels $\mathbb{Y}_\mathcal{S} = \{\mathbf{y}_\mathcal{B} \in \{0,1\}^{1\times(|\mathbb{C}|+1)} | \mathcal{B} \in \mathbb{B}\}$ and bounding box coordinates $\mathbb{P}_\mathcal{S} = \{\mathbf{p}_\mathcal{B} \in \mathbb{R}^{|\mathbb{B}|\times 4} | \mathcal{B} \in \mathbb{B}\}$ are available, where symbols $\mathbb{R}$ and $\mathbb{B}$ denote the real number field and the set of groundtruth object bounding boxes, respectively. Detector training is fully supervised for source domain:

$$L_\mathcal{S}\left(I_\mathcal{S}, \mathbb{Y}_\mathcal{S}, \mathbb{P}_\mathcal{S}\right) = L_{rpn}\left(I_\mathcal{S}, \mathbb{P}_\mathcal{S}\right) + L_{det}\left(I_\mathcal{S}, \mathbb{Y}_\mathcal{S}, \mathbb{P}_\mathcal{S}\right), \tag{1}$$

where L_{rpn} and L_{det} summarize the loss functions for training Region Proposal Networks (RPN) and bounding box prediction heads. Their further details are referred to [28].

To bridge the domain gap between $\mathcal{S}$ and $\mathcal{T}$, various previous works (*e.g.* [18, 32, 49, 2]) propose to impose adversarial alignment on the features of $\mathcal{I}_\mathcal{S}$ and $\mathcal{I}_\mathcal{T}$:

$$L_\mathcal{T}\left(I_\mathcal{S}, I_\mathcal{T}\right) = \mathbb{E}\left[\log \mathcal{D}\left(\mathbf{f}_\mathcal{S}\right)\right] + \mathbb{E}\left[\log\left(1 - \mathcal{D}\left(\mathbf{f}_\mathcal{T}\right)\right)\right], \tag{2}$$

where $\mathbf{f}_\mathcal{S}$ and $\mathbf{f}_\mathcal{T}$ are feature maps extracted from $\mathcal{I}_\mathcal{S}$ and $\mathcal{I}_\mathcal{T}$, before being fed into the RPN. Adversarial training projects them into aligned feature space so that a discriminator $\mathcal{D}$ cannot distinguish between their domains. However, criterion (2) favors minimizing the most discriminative variance between domains, which may have weak effect on transferring desired instance-level knowledge. To solve this issue, we introduce Informative Correspondence Mining (ICM) and Consistent Correspondence Mining (CCM) as explicit semantic regularizations.

3.2. Informative Correspondence Mining

Formulation. We explicitly impose semantically consistent matching constraints between $\mathbf{f}_\mathcal{S}$ and $\mathbf{f}_\mathcal{T}$, *i.e.* a source image region belonging to a certain class should be matched to the target image area occupied by the same class. For each image region $\mathcal{R}$ from $\mathbb{R}_\mathcal{S}$, *i.e.*, the set of all possible sub-image regions on $\mathbf{f}_\mathcal{S}$, we assume a correspondence field that searches for the matched area on the target feature maps $\mathbf{f}_\mathcal{T}$. Such field is generally expressible with a column vector $\mathbf{w}_\mathcal{R} \succeq \mathbf{0}$, $\mathbf{w}_\mathcal{R}^\top \mathbf{1} = 1$, where each element in $\mathbf{w}_\mathcal{R}$ denotes the soft activation of a certain pixel in $\mathbf{f}_\mathcal{T}$. In this manner, representation of the (warped) matched region can write as $\mathbf{w}_\mathcal{R}^\top \mathbf{f}_\mathcal{T}$. Suppose $\mathbb{C}_{\mathcal{S}\cap\mathcal{T}} \subseteq \mathbb{C} \cup \{\mathcal{C}_0\}$ ($\mathcal{C}_0$ denotes the background) be the shared classes between $\mathcal{I}_\mathcal{S}$ and $\mathcal{I}_\mathcal{T}$, which are ready to compute given image labels in both domains. As we cannot observe region-level annotations in target domain, we propose to divide $\mathbf{f}_\mathcal{T}$ via *weakly supervised clustering*, generating consistent semantic partition of $\mathbf{f}_\mathcal{T}$ that can best explain the source domain annotations.

To this end, we assume that the target domain image has a non-overlapping partition $\{\Omega_\mathcal{C}\}_{\mathcal{C}\in\mathbb{C}_{\mathcal{S}\cap\mathcal{T}}}$, $\forall \mathcal{C}_1, \mathcal{C}_2 \in \mathbb{C}_{\mathcal{S}\cap\mathcal{T}}$, $\mathcal{C}_1 \neq \mathcal{C}_2$, $\Omega_{\mathcal{C}_1} \cap \Omega_{\mathcal{C}_2} = \emptyset$, and $\cup_{\mathcal{C}\in\mathbb{C}_{\mathcal{S}\cap\mathcal{T}}} \Omega_\mathcal{C} = \mathcal{I}_\mathcal{T}$. For each source image region $\mathcal{R}$, let $\mathbf{w}_\mathcal{R}^\mathcal{C}$ denote the correspondence field of $\mathcal{R}$, but restricted to searching regions only in $\Omega_\mathcal{C}$, *i.e.* elements of $\mathbf{w}_\mathcal{R}^\mathcal{C}$ are zeros for pixels outside $\Omega_\mathcal{C}$. To derive the optimal partition we propose an unsupervised criterion, which we draw inspiration from Fig. 2. Since the class of source image region $\mathcal{R}$ is known (the class assignment rule will be elaborated later), suppose that $\mathcal{R}$ belongs to class $\mathcal{C}_\mathcal{R}$, and the correspondence $\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}}$ matches $\mathcal{R}$ to a target region $\mathcal{A}$ in $\Omega_{\mathcal{C}_\mathcal{R}}$ with a particular pattern (*e.g.* the horse head in Fig. 2). If $\Omega_{\mathcal{C}_\mathcal{R}}$ does not cover the full object but only a proportion of it, it is highly possible that there exists $\mathcal{B} \subseteq \Omega_{\mathcal{C}_-}, \mathcal{C}_- \neq \mathcal{C}_\mathcal{R}$, such that $\mathcal{B}$ possesses a pattern with similar features with those of $\Omega_{\mathcal{C}_\mathcal{R}}$, while consistently appeared in many images with the same shared label (*e.g.* the horse body in Fig. 2). If it happens, a good instantiation of $\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}$ is to correlate $\mathcal{R}$ with $\mathcal{B}$ to explain the class $\mathcal{C}_\mathcal{R}$. If $\Omega_{\mathcal{C}_\mathcal{R}}$ covers the full extent of $\mathcal{C}_\mathcal{R}$, then it is difficult to find such consistent patterns in $\Omega_{\mathcal{C}\_}$. In this case, knowing the match in $\Omega_{\mathcal{C}_\mathcal{R}}$ tells us little about knowing where to match in $\Omega_{\mathcal{C}\_}$. Such optimal condition could be formally modelled by minimizing the concept of mutual information:

$$\min_{\boldsymbol{\Omega}} \sum_{\substack{\mathcal{C}_-\in\mathbb{C}_{\mathcal{S}\cap\mathcal{T}}\\ \mathcal{C}_-\neq\mathcal{C}_\mathcal{R}}} I\left(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}, \mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}} | \mathbb{I} = (\mathcal{I}_\mathcal{S}, \mathcal{I}_\mathcal{T})\right), \tag{3}$$

where the mutual information

$$I(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}, \mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}}|\mathbb{I}) = H(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbb{I}) - H(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}}, \mathbb{I}), \tag{4}$$

and $H(\cdot)$ denotes the entropy. We assume uniform distribution on $P(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbb{I})$ regardless of $\mathcal{R}$ and $\mathbb{I}$, and its entropy becomes a constant. The only remaining term that matters requires solving the posterior $P(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}}, \mathbb{I})$, which, however, is difficult to directly compute.

Variational approximation. We make use of source domain annotations as immediate random variables, approximating posterior $P(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}}, \mathbb{I})$ with a factorizable one:

$$Q(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}}, \mathbb{I}) = \int P(\mathbf{a}_\mathcal{R}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}}, \mathbb{I}) P(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbf{a}_\mathcal{R}, \mathbb{I}) d\mathbf{a}_\mathcal{R}, \tag{5}$$

where $\mathbf{a}_\mathcal{R}$ denotes the annotation of source region $\mathcal{R}$. It could be shown[1] that using $\mathcal{Q}$ as the surrogate posterior, the problem (3) could be simplified to

$$\begin{gathered}\max_{\boldsymbol{\Omega}} A(\mathcal{S},\mathcal{T}) = \frac{1}{Z_A}\sum_{\mathbb{I}}\sum_{\mathcal{R}\in\mathbb{R}_\mathcal{S}}\sum_{\mathcal{C}_-} H\left(\mathbf{a}_\mathcal{R}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}, \mathbb{I}\right), \\ \text{s.t. } \mathbf{a}_\mathcal{R} = \arg\max_{\mathbf{a}} P\left(\mathbf{a}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}}, \mathbb{I}\right),\end{gathered} \tag{6}$$

where Z_A is a normalization constant. Note the constraint in (6) requires $\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}}$ to find correct semantic pattern to explain $\mathcal{C}_\mathcal{R}$, while its objective is making such explanation difficult for any remaining region, which meets our intuition.

Note that so far we assume that the learned semantic correspondence $\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}$ is perfect enough to correctly localize any semantic region informed by $\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}}$. It naturally gives rise to a min-max interpretation: while current partition $\boldsymbol{\Omega}$ should raise the entropy (or uncertainty) to find good correspondences in $\Omega_{\mathcal{C}\_}$, the correspondence field $\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}$ should be assumed as powerful as possible to find such match and minimize the entropy. Further relaxing the constraints in (6), we arrive at the following problem

$$\min_{\mathbf{w}}\left(\lambda N(\mathcal{S},\mathcal{T}) + \max_{\boldsymbol{\Omega}} A(\mathcal{S},\mathcal{T})\right), \tag{7}$$

where $\mathcal{N}(\mathcal{S},\mathcal{T}) = -\frac{1}{Z_N}\sum_\mathbb{I}\sum_{\mathcal{R}\in\mathbb{R}_\mathcal{S}}\log P(\mathbf{a}_\mathcal{R}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}}, \mathbb{I})$ denotes the negative log-posterior, and $\lambda > 0$ is a tolerance parameter and set to 1 in our experiments.

Implementation. The simplified objective (7) only involves solving the posterior $P(\mathbf{a}_\mathcal{R}|\mathbf{w}_\mathcal{R}^\mathcal{C}, \mathbb{I})$. We define

$$P\left(\mathbf{a}_\mathcal{R}|\mathbf{w}_\mathcal{R}^\mathcal{C}, \mathbb{I}\right) = P\left(\mathbf{y}_\mathcal{R}|\mathbf{w}_\mathcal{R}^\mathcal{C}, \mathbb{I}\right) P\left(\mathbf{o}_\mathcal{R}|\mathbf{w}_\mathcal{R}^\mathcal{C}, \mathbb{I}\right), \tag{8}$$

where $\mathbf{a}_\mathcal{R} = (\mathbf{y}_\mathcal{R}, \mathbf{o}_\mathcal{R})$ and $\mathbf{y}_\mathcal{R}$ consists of the annotated class and position of region $\mathcal{R}$, respectively. We set $\mathbf{y}_\mathcal{R}$ to the one-hot class vector of the closest groundtruth bounding box if their intersection-over-union overlap exceeds 0.7. If $\mathcal{R}$ has unconvincing overlap with any groundtruth (below 0.3), then $\mathbf{y}_\mathcal{R}$ is set to background. The position annotation $\mathbf{o}_\mathcal{R} \in \mathbb{R}^{1\times 4}$ is defined only for positive regions, as the offset vector towards the closest bounding box.

We parameterize the posteriors (8) with deep neural networks. It is achieved by first obtaining the warped features $\mathbf{f}_\mathcal{R}^\mathcal{C} = \left(\mathbf{w}_\mathcal{R}^\mathcal{C}\right)^\mathrm{T}\mathbf{f}_\mathcal{T}$, then feeding them into two separate fully connected layers $\mathcal{F}_c(\cdot)$ and $\mathcal{F}_o(\cdot)$ to produce the logits. We make use of the common softmax and Laplacian distributions to define the posteriors:

$$\begin{aligned} &P\left(\mathbf{y}_\mathcal{R}^\mathcal{C}|\mathbf{w}_\mathcal{R}^\mathcal{C}, \mathbb{I}\right) \sim \text{softmax}\left(\mathbf{y}_\mathcal{R}^\mathcal{C}, \mathcal{F}_c\left(\mathbf{f}_\mathcal{R}^\mathcal{C}\right)\right), \\ &P\left(\mathbf{o}_\mathcal{R}^\mathcal{C}|\mathbf{w}_\mathcal{R}^\mathcal{C}, \mathbb{I}\right) \sim \exp\left(-\frac{\|\mathbf{o}_\mathcal{R}^\mathcal{C} - \mathcal{F}_o\left(\mathbf{f}_\mathcal{R}^\mathcal{C}\right)\|_1}{\sigma_o^2}\right). \end{aligned} \tag{9}$$

The semantic correspondence field $\mathbf{w}_\mathcal{R}^\mathcal{C}$ and image partition $\boldsymbol{\Omega}$ lie in exponential solution space, whose complexity is also addressed with neural networks. Since $\mathbf{f}_\mathcal{T}$ is expected to convey rich semantic information, our partition generator $\mathcal{G}: \ \mathbf{f}_\mathcal{T} \rightarrow (0,1)^{HW\times|\mathbb{C}|}$ is simply a single convolutional layer followed by channel-wise softmax normalization. Computing the correspondence field $\mathbf{w}_\mathcal{R}^\mathcal{C}$ for each region $\mathcal{R}$ and class $\mathcal{C}$ individually would be expensive. We resort to computing cross-image co-attention followed by pooling the correspondence vectors in each region:

$$\begin{aligned} &\mathbf{w}_\mathcal{R}^\mathcal{C} = \text{avgpool}\left(\text{norm}\left(\mathbf{m}\left(\Omega_\mathcal{C}\right)\odot\kappa\left(\mathbf{f}_\mathcal{S},\mathbf{f}_\mathcal{T}\right)\right), \mathcal{R}\right), \\ &\text{where } \kappa\left(\mathbf{f}_\mathcal{S},\mathbf{f}_\mathcal{T}\right) = \text{softmax}\left(\mathbf{f}_\mathcal{S}^\mathrm{T}\mathbf{W}\mathbf{f}_\mathcal{T}\right). \end{aligned} \tag{10}$$

Here $\mathbf{W}$ is a learnable weight matrix. With row-wise softmax normalization, each row of $\kappa(\cdot,\cdot)$ represents the affinities from a certain location on $\mathbf{f}_\mathcal{S}$ to all the locations on $\mathbf{f}_\mathcal{T}$. Using the region mask $\mathbf{m}(\Omega_\mathcal{C})$ output from the generator $\mathcal{G}$, we restrict affinity computation valid in $\Omega_\mathcal{C}$ ($\odot$ denotes the point-wise multiplication with channel broadcasting). The affinities in $\Omega_\mathcal{C}$ are renormalized and finally pooled in the region $\mathcal{R}$ to obtain $\mathbf{w}_\mathcal{R}^\mathcal{C}$. This formula allows us to process hundreds of regions with negligible cost in parallel.

Discussion. A prevailing strategy proposed for weakly supervised object localization is region-based dropout [35, 5, 29], which iteratively erases the most discriminative regions and pushes the classifiers to find the next informative

[1] Please check our supplementary material for detailed derivation.

ones. ICM could be deemed as “correspondence dropout”, as it adversarially drops the information on the target image to make the correspondence searching harder. Another difference is that ICM is defined over local regions, using local instead of global classification cues. ICM also correlates with region competition [44, 6], as expanding, for example, $\Omega_{\mathcal{C}_0}$, would gain the information it contains and decrease the uncertainty to explain the source image regions for class $\mathcal{C}_0$, yet go the opposite way for other $\mathcal{C}$s. Thus, terms in (7) naturally balance each other and avoid trivial solution, acting as class proportion regularization [11, 16].

3.3. Consistent Correspondence Mining

While inter-domain semantic correspondence could be learned with weakly supervised clustering, we show another mining source derived from the cross-domain cycle consistency. As shown in Fig. 2, the high-level idea is to leverage more robust intra-domain matching as guidance, as the semantic representations of such two images have no domain discrepancy. It provides a strong signal to regularize inter-domain correspondence by forming a matching cycle.

To this end, to form the smallest valid cycle we sample a triplet of images $\mathcal{I}_{\mathcal{S}}$, $\mathcal{I}_{\mathcal{T}_1}$ and $\mathcal{I}_{\mathcal{T}_2}$, where $\mathcal{I}_{\mathcal{S}}$ is from the source domain and the other two are from the target domain. We introduce the cycle consistency regularizer [39, 51, 41, 52] to cross-domain object detection: warping the semantic features of $\mathcal{I}_{\mathcal{T}_2}$ to the coordinate frame of $\mathcal{I}_{\mathcal{T}_1}$ should be equivalent to first warping it to the coordinate frame of $\mathcal{I}_{\mathcal{S}}$ then that of $\mathcal{I}_{\mathcal{T}_1}$. This effectively prevents erroneous matching that is inconsistent to propagate. Formally, for a sampled triplet $\mathbb{J} = (\mathcal{I}_{\mathcal{S}}, \mathcal{I}_{\mathcal{T}_1}, \mathcal{I}_{\mathcal{T}_2})$ we minimize

$$C\left(\mathcal{S}, \mathcal{T}\right) = \frac{1}{Z_C} \sum_{\mathbb{J}} \mathbf{R}_{\mathbb{J}} \left\| \mathbf{K}_{\mathcal{T}_1 \leftarrow \mathcal{T}_2} - \mathbf{K}_{\mathcal{T}_1 \leftarrow \mathcal{S}} \mathbf{K}_{\mathcal{S} \leftarrow \mathcal{T}_2} \right\|_2^2, \tag{11}$$

where $\mathbf{K}_{\mathcal{B} \leftarrow \mathcal{A}} = \kappa(\mathbf{f}_{\mathcal{A}}, \mathbf{f}_{\mathcal{B}})\mathbf{f}_{\mathcal{B}}$ and $\kappa\left(\cdot, \cdot\right)$ denotes the cross-attention matrix defined in (10). The matrix $\mathbf{R}_{\mathbb{J}}$ has the same spatial dimensions with those of $\mathbf{K}_{\mathcal{B} \leftarrow \mathcal{A}}$, and quantifies the “transferablility” at each position. For instance, if $\mathcal{I}_{\mathcal{T}_1}$ and $\mathcal{I}_{\mathcal{T}_2}$ share a class that is absent in $\mathcal{I}_{\mathcal{S}}$, we cannot expect to reconstruct the warping $\mathcal{T}_1 \leftarrow \mathcal{T}_2$ faithfully everywhere using the immediate warpings $\mathcal{T}_1 \leftarrow \mathcal{S}$ and $\mathcal{S} \leftarrow \mathcal{T}_2$.

We measure the transferability of $\mathcal{B} \leftarrow \mathcal{A}$ for the ith feature pixel of $\mathcal{A}$, *i.e.* $\mathbf{f}_{\mathcal{A}}^{(i)}$, as the uncertainty of searching for correspondence on $\mathbf{f}_{\mathcal{B}}$, where $\mathcal{B}$ is assumed from the source domain. This is achieved by firstly computing normalized affinities $\mathbf{p}_{\mathcal{A},\mathcal{B}}^{(i)} = \mathrm{softmax}((\mathbf{f}_{\mathcal{A}}^{(i)})^{\mathrm{T}}\mathbf{W}\mathbf{f}_{\mathcal{B}})$. As $\mathcal{B}$ is annotated, we can generate per-class affinities by max-pooling $\mathbf{p}_{\mathcal{A},\mathcal{B}}^{(i)}$'s elements within the semantic area defined by the coverage of groundtruth object bounding boxes for each class. Assume it gives us (normalized) class affinities $\mathbf{c}_{\mathcal{A},\mathcal{B}}^{(i)} \in (0,1)^{1\times(|\mathbb{C}|+1)}$, $\|\mathbf{c}_{\mathcal{A},\mathcal{B}}^{(i)}\|_1 = 1$, where its component is simply set to a small constant near zero if the corresponding class is absent in $\mathcal{B}$. The transferability is then defined as $r_{\mathcal{A},\mathcal{B}}^{(i)} = \exp(-H(\mathbf{c}_{\mathcal{A},\mathcal{B}}^{(i)}))$, where $H(\cdot)$ is the entropy. Intuitively, if $\mathbf{f}_{\mathcal{A}}^{(i)}$ is a confident match, $\mathbf{c}_{\mathcal{A}}^{(i)}$ tends to have peaks, leading to low uncertainty (high transferability).

Let $\mathbf{r}_{\mathcal{A},\mathcal{B}}$ denote the matrix collecting $r_{\mathcal{A},\mathcal{B}}^{(i)}$s of all pixels. As $\mathcal{B}$ is assumed a source image, $\mathbf{R}_{\mathbb{J}}$ is accumulated as:

$$\mathbf{R}_{\mathbb{J}} = \mathbf{r}_{\mathcal{T}_1,\mathcal{S}} \odot \left(\mathbf{K}_{\mathcal{T}_1 \leftarrow \mathcal{T}_2} \mathbf{r}_{\mathcal{T}_2,\mathcal{S}}\right). \tag{12}$$

Note that during training $\mathbf{R}_{\mathbb{J}}$ is detached from optimization and precomputed using the estimations from the trained model of latest epoch, to prevent gradient instability issues.

3.4. Implementation Details

Our full training objective is

$$\min_{\theta_0} \left(L_{\mathcal{S}} + \alpha(\mathcal{N} + \max_{\theta_{\mathbf{\Omega}}} A) + \beta C \right), \tag{13}$$

where $L_{\mathcal{S}}$ is the source domain loss terms defined in (1), α and β are balancing weights, and θ_Ω and θ_0 are training parameters of the partition generator $\mathcal{G}$ and the remaining network, respectively. We perform adversarial training by ascending the gradients of $\mathcal{G}$ via gradient reversal [10]. Note that ICM and CCM only affect training, which outputs an adapted detector that directly applies to target domain.

Class-agnostic ICM. The original multi-class version of ICM repeats correspondence learning for every class, which would be slow if the number of classes is large. A walkaround is to unify all the object classes as “foreground”, rendering it a binary foreground/background setting. In practice, we do not observe performance degeneration for this class-agnostic setting, yet saving training time significantly.

4. Experiments

4.1. Experimental Settings

Datasets. We follow [15, 19] to organize the evaluation data, where the *trainval* sets of Pascal-VOC 2007 and Pascal-VOC 2012 [8] are treated as source domain, while the Clipart1k, Watercolor2k, and Comic2k datasets [15] as target domains. There are 16551 real-world photos from 20 object classes to form the source domain. The target domain images are unrealistic, *e.g.* with cartoon or painting styles. Clipart1k consists of 1000 images from 20 object classes, while Watercolor2k and Comic2k both contain 2000 images from 6 classes. These classes are all included by the 20 classes of the source domain. We follow previous train/test split: for Clipart1k all its 1000 images are used for training/evaluation, while for Watercolor2k and Comic2k, there are 1000 for training and another 1000 for evaluation.

State-of-the-art methods. We compare our approach with 9 recent works with released results and/or codes, organized into 3 groups: 1) Weakly Supervised (WS) Group including WSDDN [1], CLNet [17], EDRN [34], PCL [37],

Table 1. Average Precisions (AP) and mean AP on Clipart1k. Bold highlights the top place while underline the second place.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Source only	35.6	52.5	24.3	23.0	20.0	43.9	32.8	10.7	30.6	11.7	13.8	6	36.8	45.9	48.7	41.9	16.5	7.3	22.9	32	27.8
WL Group																					
WSDDN [1]	1.6	3.6	0.6	2.3	0.1	11.7	4.5	0.0	3.2	0.1	2.8	2.3	0.9	0.1	14.4	16.0	4.5	0.7	1.2	18.3	4.4
CLNet [17]	3.2	22.3	2.2	0.7	4.6	4.8	17.5	0.2	4.8	1.6	6.4	0.6	4.7	0.6	12.5	13.1	14.1	4.1	8.0	29.7	7.8
EDRN [34]	2.7	13.5	1.2	4.2	1.8	10.3	25.7	0.4	8.4	0.3	3.2	2.7	1.1	0.7	29.4	17.2	5.2	1.6	2.9	19.1	7.6
PCL [37]	3.4	10.6	2.3	1.7	5.2	3.4	23.3	1.2	5.6	0.4	7.8	3.7	5.6	0.3	24.5	19.7	11.9	3.6	9.2	25.4	8.4
UDA Group																					
ADDA [38]	20.1	50.2	20.5	23.6	11.4	40.5	34.9	2.3	39.7	22.3	27.1	10.4	31.7	53.6	46.6	32.1	18.0	21.1	23.6	18.3	27.4
SWDA [32]	26.2	48.5	32.6	33.7	38.5	54.3	37.1	18.6	34.8	58.3	17.0	12.5	33.8	65.5	61.6	**52.0**	9.3	24.9	54.1	49.1	38.1
STABR [19]	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
HTD [2]	33.6	58.9	34.0	23.4	**45.6**	57.0	39.8	12.0	39.7	51.3	21.1	20.1	39.1	72.8	63.0	43.1	19.3	30.1	50.2	**51.8**	40.3
CDWS Group																					
CDWSDA [15]	32.0	40.9	29.5	29.3	32.0	**84.7**	38.2	12.4	24.3	54.8	24.7	15.4	36.1	72.1	51.0	41.9	19.0	18.5	47.2	21.4	36.3
Proposed	**39.8**	**66.7**	**37.2**	**42.5**	43.3	48.1	**48.1**	**21.3**	**46.5**	**73.0**	29.0	**29.8**	**57.3**	**78.6**	**67.8**	48.7	**46.3**	19.3	42.8	48.5	**46.7**

Table 2. Average Precisions (AP) and mean AP on Watercolor2k. Bold highlights the top place while underline the second place.

Method	bike	bird	car	cat	dog	person	mAP
Source only	68.8	46.8	37.2	32.7	21.3	60.7	44.6
WL Group							
WSDDN [1]	1.5	26.0	14.6	0.4	0.5	33.3	12.7
CLNet [17]	4.5	27.9	19.6	14.3	6.4	31.4	17.4
EDRN [34]	5.2	29.3	15.3	1.4	0.9	34.9	14.5
PCL [37]	6.7	28.8	20.2	9.5	5.4	27.4	16.3
UDA Group							
ADDA [38]	79.9	49.5	39.5	**35.3**	29.4	65.1	49.8
SWDA [32]	82.3	55.9	46.5	32.7	35.5	66.7	53.3
STABR [19]	75.6	45.8	49.3	34.1	30.3	64.1	49.4
HTD [2]	69.2	49.5	49.5	34.9	30.8	61.2	49.2
CDWS Group							
CDWSDA [15]	68.6	46.6	37.7	35.2	36.0	62.5	47.8
Proposed	**86.6**	**64.2**	**52.6**	32.4	**41.2**	**67.4**	**57.4**

Table 3. Average Precisions (AP) and mean AP on Comic2k. Bold highlights the top place while underline the second place.

Method	bike	bird	car	cat	dog	person	mAP
Source only	28.8	13.5	18.6	14.8	15.9	33.9	20.9
WL Group							
WSDDN [1]	1.5	0.1	11.9	6.9	1.4	12.1	5.6
CLNet [17]	0.0	0.0	2.0	4.7	1.2	14.9	3.8
EDRN [34]	1.6	0.5	13.2	7.2	2.5	13.2	6.4
PCL [37]	1.2	0.4	8.9	2.9	2.3	15.6	5.2
UDA Group							
ADDA [38]	39.5	9.8	17.2	12.7	20.4	43.3	23.8
SWDA [32]	30.3	19.6	28.8	15.2	24.9	46.9	27.6
STABR [19]	**50.6**	13.6	31.0	7.5	16.4	41.4	26.8
HTD [2]	35.4	14.8	26.6	13.7	26.9	40.0	26.2
CDWS Group							
CDWSDA [15]	47.0	21.1	30.1	29.0	29.6	40.6	32.9
Proposed	**50.6**	**23.3**	**35.4**	**32.3**	**33.8**	**47.1**	**37.1**

which directly apply to the target domain without adaptation; 2) Unsupervised Domain Adaptation (UDA) Group including ADDA [38], SWDA [32], STABR [19], HTD [2], which assumes labeled source domain and unlabeled target domain; and Cross-Domain Weakly Supervised (CDWS) Group, including CDWSDA [15] and our approach, assuming weakly labeled target domain. Note that CDWSDA reports results using the SSD300 variant [24], which deviates from [2, 32] and ours that adopt Faster-RCNN as backbone. For fair comparison, we implement its Faster-RCNN variant using the processed immediate data released by authors. Our reimplementation is guaranteed to produce higher numbers on Clipart1k than those reported by the authors.

Training details. If not explained, we set the parameters α and β in (7) with 0.001 and 0.01, respectively. Their influences are also analysed in Sect. 4.3. The newly added layers (*e.g.* the learnable weights and classifiers in ICM and CCM) are all initialized with Normal distribution with zero mean and 0.01 standard deviation. We perform $70k$ training steps for Clipart1k and $140k$ for Watercolor2k/Comic2k. At each training step a mini-batch of 3 images (one from the source domain and two from the target domain) is sampled. We use SGD optimizer with momentum, with an initial learning rate 0.001 decayed by 10x at the $50k$th step.

4.2. Comparisons with State-of-the-Art Methods

Performance comparisons with the state-of-the-art methods on Clipart1k, Watercolor2k and Comic2k are summarized in Table 1, 2 and 3, respectively. The proposed approach achieves the leading results for all the datasets, improving over previous results by $4\% \sim 6\%$ in mAP. Conversely, the WS group fails achieving reasonable results because of the large appearance and style diversity of the target datasets, which is also reported in [15]. By introduc-

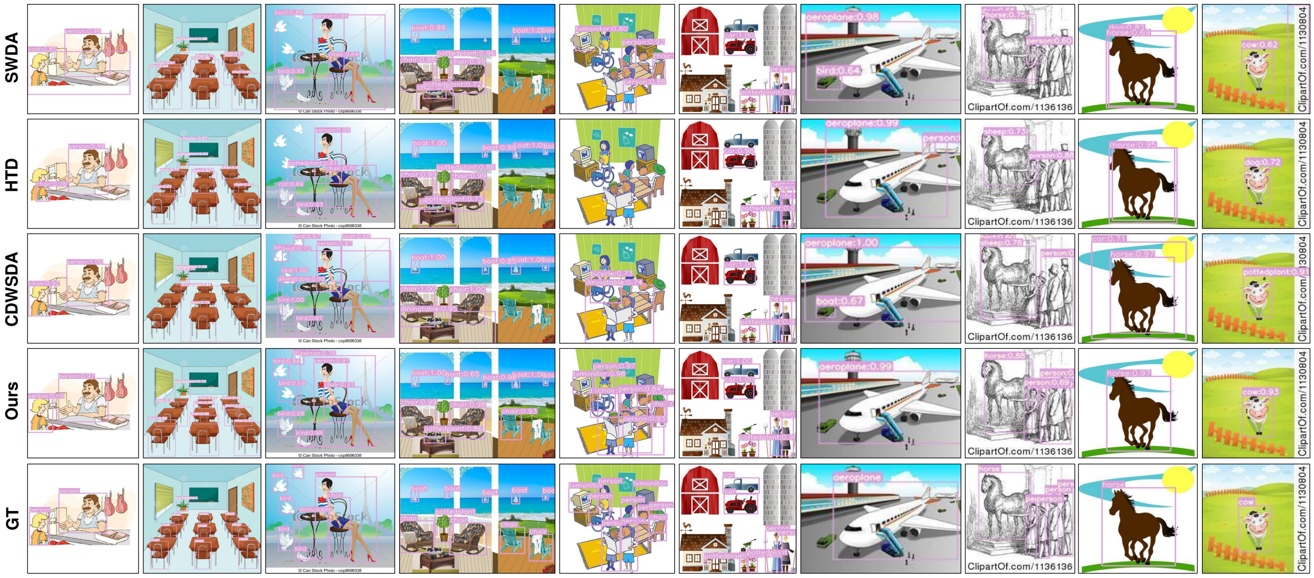

Figure 3. Representative results generated by different approaches (visualized in different rows). Best viewed with zoom in.

Table 4. Contributions to the final mAP by different components, evaluated on the Clipart1k dataset.

Source	ICM			CCM	mAP
	w/o adv.	full	w/o reg.		
✓					27.8
✓	✓				44.3
✓		✓			45.0
✓	✓			✓	45.7
✓			✓	✓	45.5
✓		✓		✓	46.7

ing the weak knowledge of target domain, performance of CDWSDA [15] is significantly better than that of the UDA group on certain categories, yet meanwhile degenerate for several others (*e.g.* chair), partly caused by the inaccuracy of pseudo labels when training target domain detectors. As for the proposed approach, dominant improvements are observed on several major classes (*e.g.* animal categories) that are richly spread in the datasets yet difficult to handle by previous works, due to largely non-rigid deformations. Our approach explicitly groups the objects of diverse appearance into semantic clusters, making their representations similar in feature space, gaining robustness to handling such appearance diversity. However, our approach does not achieve meaningful improvements for minor poputation categories, such as *train* and *tv*).

Fig. 3 shows a qualitative comparison of 4 reprentative approaches: SWDA [32], HTD [2], CDWSDA [15] and the proposed. The proposed approach generates better results in case of multiple objects (2nd and 3rd columns), cluttered scenes (5th and 8th columns), and has fewer false positives and missing detections (1st, 7∼10th columns). More visual comparisons could be found in our supplementary material.

4.3. Performance Analysis

The proposed approach contains several novel modules, whose effectiveness is evaluated via a series of experiments:

Ablation study. We quantitatively analyse the contributions of different components in Table 4. The performance is limited with only source domain training. Using the proposed ICM (without adversarial correspondence dropout), it improves the mAP remarkably by 16.5%, demonstrating the advantage of pixel-wise knowledge transfer. Adversarial mask generation further improves it by 0.7%. Including the CCM module brings with 1.4% improvement. We also analyse the necessity of region offset regression in the proposed ICM module (eq. (8)), similar with bounding box regression in Faster RCNN [28]. Surprisingly, excluding it causes a notable drop of 1.2% mAP. It indicates that explaining not only the class labels, but also the object positions of source domain is beneficial for the ICM module.

Analysis of error reduction. We further analyse the types of detection error reduced by the proposed ICM and CCM modules. We consider the baseline approach for unsupervised domain adaptation, *i.e.* source domain training plus inter-domain adversarial feature alignment, the state-of-the-art weakly supervised approach CDWSDA [15], the proposed approach equipped with the ICM module, and its full version (ICM + CCM). For each approach, we collect detection results in descending order of their scores in the whole Clipart1k dataset, and count the percentage of detections of different types as explained in Fig. 4. The top-left figure of Fig. 4 shows that both CDWSDA and ICM improves over the UDA baseline dramatically in classification and localization accuracy, while ICM beats CDWSDA by a large margin. The top-right and bottom-left figures in-

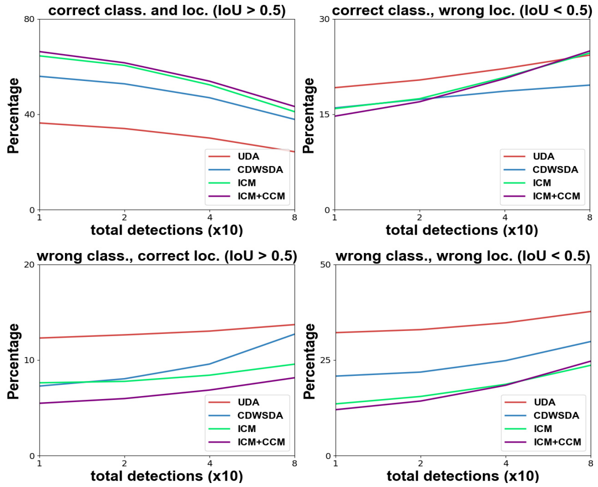

Figure 4. Percentage of detections within each type as a function of the number of detections. Top-left: detections with correct classification and localization. Top-right: classification is correct, but localization is weak ($0.1 < \text{IoU} < 0.5$). Bottom-left: wrong classification, but correct localization (IoU with at least one object exceeds 0.5). Bottom-right: detections with wrong labels and localization ($\text{IoU} < 0.1$). Note that higher percentage is preferred for only the top-left figure, as it counts for true positive detections.

dicate that ICM mainly reduce classification errors of correctly localized regions, partly because ICM does not have pseudo labeling process that may introduce undesired labeling noise. Finally, it is clearly shown that CCM has consistent positive effect in reducing all kinds of detection errors.

Visualizing the effect of ICM and CCM. Fig. 5 visualizes how ICM and CCM affects the results. To this end, we treat the pixels within the bounding boxes of the person and bicycle in Fig. 5 (a) as seeds, computing the matched positions on the target domain image according to their co-attention matrix (κ defined in (10)). For each matched position, we predict a bounding box by feeding its feature to the RCNN head, and accumulate all these bounding boxes as the coverage of the matched regions of source seeds on the target image. Note that seeds are weighted using a spatial Gaussian to suppress the contributions of background seeds. The results illustrate that without adversarial masking in ICM, only parts of object are matched in target image to explain the annotated source classes. Adversarial masking render the correspondences uniformly spread along the full objects. It also shows that excluding CCM makes the match of the person erroneously located on the background, while including it results into more accurate localization.

Parameter tuning results. Fig. 6 summarizes the sensitivity analysis of the parameters α and β in (7). The results show that the performance tend to drop if both parameters are set small. Otherwise, the performance is stably floating around $46.1\% \sim 46.7\%$, while variations are both caused by parameter change and randomness during training.

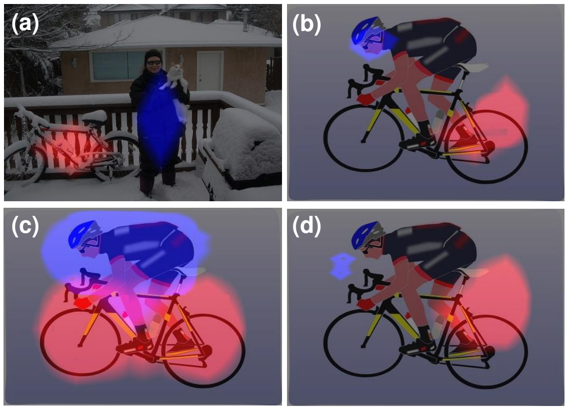

Figure 5. Visualizing the effect of Informative Correspondence Mining (ICM) and Consistent Correspondence Mining (CCM). (a) Seeds on the source domain image, weighted with a spatial Gaussian. (b) (c) (d): Visualization of the distribution of matched regions, corresponding to: (b) with naive ICM, but without adversarial masking; (c) the full ICM module; (d) without CCM.

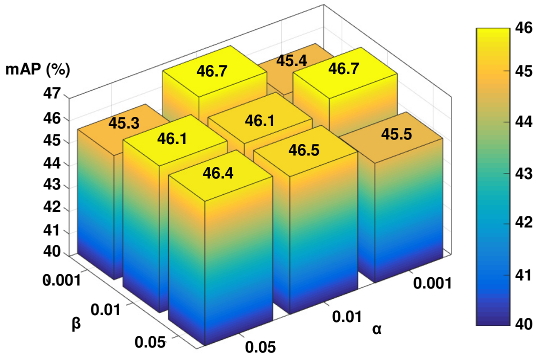

Figure 6. Mean average precision as a function of parameters α and β defined in Eqn. (7), evaluated on the Clipart1k dataset.

5. Conclusion and Limitation

In this work we propose two novel modules, Informative Correspondence Mining (ICM) and Consistent Correspondence Mining (CCM), to address cross-domain weakly supervised object detection. ICM finds informative cross-domain correspondences for local semantics transfer, while CCM incorporates cycle learning as consistency regularizer. New state-of-the-art results are set on three benchmarks.

Limitation. A drawback or our approach is the imbalanced sampling issue on cross-domain triplets, as extensively discussed in various tasks [25, 27]. Due to the uniform sampling, our approach does not treat the categories with different population in balancing way. This is why for minor categories the proposed approach does not achieve results as good as those of the major ones. Further explorations of sampling mechanisms are left as future work.

Acknowledgement. This work is supported by grants from National Natural Science Foundation of China 61922006, CAAI-Huawei MindSpore Open Fund, and Baidu academic collaboration program.

References

[1] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, pages 2846–2854, 2016. 5, 6

[2] C.i Chen, Z. Zheng, X. Ding, Y. Huang, and Q. Dou. Harmonizing transferability and discriminability for adapting object detectors. In *CVPR*, pages 8866–8875, 2020. 1, 2, 3, 6, 7

[3] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool. Domain adaptive faster R-CNN for object detection in the wild. In *CVPR*, pages 3339–3348, 2018. 1, 2

[4] Z. Chen, Z. Fu, R. Jiang, Y. Chen, and X.-S. Hua. SLV: spatial likelihood voting for weakly supervised object detection. In *CVPR*, pages 12992–13001, 2020. 2

[5] J. Choe and H. Shim. Attention-based dropout layer for weakly supervised object localization. In *CVPR*, pages 2219–2228, 2019. 2, 4

[6] D. Cremers and S. Soatto. Motion competition: A variational approach to piecewise parametric motion segmentation. *IJCV*, 62(3):249–265, 2004. 5

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 1

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010. 1, 2, 5

[9] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010. 1

[10] Y. Ganin and V. S. Lempitsky. Unsupervised domain adaptation by backpropagation. In F. R. Bach and D. M. Blei, editors, *ICML*, volume 37, pages 1180–1189, 2015. 1, 5

[11] R. Gomes, A. Krause, and P. Perona. Discriminative clustering by regularized information maximization. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *NeurIPS*, pages 775–783, 2010. 5

[12] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask R-CNN. *TPAMI*, 42(2):386–397, 2020. 2

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[14] C.-C. Hsu, Y.-H. Tsai, Y.-Y. Lin, and M.-H. Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *ECCV*, 2020. 1, 2

[15] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, pages 5001–5009, 2018. 2, 5, 6, 7

[16] A. Joulin, F. R. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, pages 1943–1950, 2010. 5

[17] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *ECCV*, volume 9909, pages 350–365, 2016. 5, 6

[18] M. Khodabandeh, A. Vahdat, M. Ranjbar, and W. G. Macready. A robust learning approach to domain adaptive object detection. In *ICCV*, pages 480–490, 2019. 1, 2, 3

[19] S. Kim, J. Choi, T. Kim, and C. Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *ICCV*, pages 6091–6100, 2019. 1, 2, 5, 6

[20] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1106–1114, 2012. 1

[21] Y. Li, L. Yuan, and N. Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, pages 6936–6945, 2019. 1, 2

[22] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *TPAMI*, 42(2):318–327, 2020. 2

[23] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *ECCV*, volume 8693, pages 740–755, 2014. 1, 2

[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg. SSD: single shot multibox detector. In *ECCV*, volume 9905, pages 21–37, 2016. 6

[25] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, pages 2537–2546, 2019. 8

[26] W. Lu, X. Jia, W. Xie, L. Shen, Y. Zhou, and J. Duan. Geometry constrained weakly supervised object localization. In *ECCV*, 2020. 2

[27] Q. Qian, L. Shang, B. Sun, J. Hu, T. Tacoma, H. Li, and R. Jin. Softtriple loss: Deep metric learning without triplet sampling. In *CVPR*, pages 6449–6457, 2019. 8

[28] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *TPAMI*, 39(6):1137–1149, 2017. 1, 2, 3, 7

[29] Z. Ren, Z. Yu, X. Yang, M.-Y. Liu, Y. J. Lee, A. G. Schwing, and J. Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *CVPR*, pages 10595–10604, 2020. 2, 4

[30] A. RoyChowdhury, P. Chakrabarty, A. Singh, S. Jin, H. Jiang, L. Cao, and E. G. Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In *CVPR*, pages 780–790, 2019. 1, 2

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z.g Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 1, 2

[32] K. Saito, Y. Ushiku, T. Harada, and K. Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019. 2, 3, 6, 7

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *IJCV*, 128(2):336–359, 2020. 2

[34] Y. Shen, R. Ji, Y. Wang, Z. Chen, F. Zheng, F. Huang, and Y. Wu. Enabling deep residual networks for weakly supervised object detection. In *ECCV*, volume 12353, pages 118–136, 2020. 2, 5, 6

[35] K. K. Singh and Y. J. Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, pages 3544–3553, 2017. 2, 4

[36] M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, volume 97, pages 6105–6114, 2019. 1, 2

[37] P. Tang, X. Wang, S.Bai, W. Shen, X. Bai, W. Liu, and A. L. Yuille. PCL: proposal cluster learning for weakly supervised object detection. *TPAMI*, 42(1):176–191, 2020. 2, 5, 6

[38] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 2962–2971, 2017. 2, 6

[39] F. Wang, Q. Huang, and L. J. Guibas. Image co-segmentation via consistent functional maps. In *ICCV*, pages 849–856, 2013. 5

[40] J. Wang, Y. Chen, S. Hao, W. Feng, and Z. Shen. Balanced distribution adaptation for transfer learning. In *ICDM*, pages 1129–1134, 2017. 1, 2

[41] X. Wang, A. Jabri, and A. A. Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, pages 2566–2576, 2019. 5

[42] C.-D. Xu, X.-R. Zhao, X. Jin, and X.-S. Wei. Exploring categorical regularization for domain adaptive object detection. In *CVPR*, pages 11721–11730, 2020. 1, 2

[43] M. Xu, H. Wang, B. Ni, Q. Tian, and W. Zhang. Cross-domain detection via graph-induced prototype alignment. In *CVPR*, pages 12352–12361. IEEE, 2020. 1, 2

[44] Y. Yang, A. Loquercio, D. Scaramuzza, and S. Soatto. Unsupervised moving object detection via contextual information separation. In *CVPR*, pages 879–888, 2019. 5

[45] Y. Yang and S. Soatto. FDA: fourier domain adaptation for semantic segmentation. In *CVPR*, pages 4084–4094, 2020. 1, 2

[46] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang. Adversarial complementary learning for weakly supervised object localization. In *CVPR*, pages 1325–1334, 2018. 2

[47] X. Zhang, Y. Wei, G. Kang, Y. Yang, and T. S. Huang. Self-produced guidance for weakly-supervised object localization. In *ECCV*, volume 11216, pages 610–625, 2018. 2

[48] G. Zhao, G. Li, R. Xu, and L. Lin. Collaborative training between region proposal localization and classification for domain adaptive object detection. In *ECCV*, 2020. 1

[49] Y. Zheng, D. Huang, S. Liu, and Y. Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *CVPR*, pages 13763–13772. 2, 3

[50] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 2

[51] T. Zhou, P. Krähenbühl, M. Aubry, Q.-X. Huang, and A. A. Efros. Learning dense correspondence via 3d-guided cycle consistency. In *CVPR*, pages 117–126, 2016. 5

[52] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2242–2251, 2017. 5

[53] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin. Adapting object detectors via selective cross-domain alignment. In *CVPR*, pages 687–696. 2