H^{2}FA R-CNN: Holistic and Hierarchical Feature Alignment for Cross-domain Weakly Supervised Object Detection

Yunqiu Xu[1,2*] Yifan Sun[1] Zongxin Yang[3] Jiaxu Miao[3] Yi Yang[3]
[1]Baidu Research [2]ReLER, AAII, University of Technology Sydney [3]CCAI, Zhejiang University
imyunqiuxu@gmail.com sunyf15@tsinghua.org.cn {yangzongxin,jiaxumiao,yangyics}@zju.edu.cn

Abstract

Cross-domain weakly supervised object detection (CDWSOD) aims to adapt the detection model to a novel target domain with easily acquired image-level annotations. How to align the source and target domains is critical to the CDWSOD accuracy. Existing methods usually focus on partial detection components for domain alignment. In contrast, this paper considers that all the detection components are important and proposes a Holistic and Hierarchical Feature Alignment (H^{2}FA) R-CNN. H^{2}FA R-CNN enforces two image-level alignments for the backbone features, as well as two instance-level alignments for the RPN and detection head. This coarse-to-fine aligning hierarchy is in pace with the detection pipeline, i.e.*, processing the image-level feature and the instance-level features from bottom to top. Importantly, we devise a novel hybrid supervision method for learning two instance-level alignments. It enables the RPN and detection head to simultaneously receive weak/full supervision from the target/source domains. Combining all these feature alignments, H^{2}FA R-CNN effectively mitigates the gap between the source and target domains. Experimental results show that H^{2}FA R-CNN significantly improves cross-domain object detection accuracy and sets new state of the art on popular benchmarks. Code and pre-trained models are available at* https://github.com/XuYunqiu/H2FA_R-CNN.

1. Introduction

Cross-domain weakly supervised object detection (CDWSOD) is of significant value in realistic detection applications. Specifically, the training data and the testing data are sometimes under different domains (*i.e.*, the source domain and target domain, respectively), yielding a cross-domain detection scenario. To mitigate the domain shift, there are three potential solutions, *i.e.*, the supervised, unsupervised and weakly supervised approaches. The supervised approach requires additional densely annotated samples (*i.e.*, instance-level bounding boxes) on the target domain, which can be very burdensome. In contrast, the unsupervised approach [10,15,50] relieves the annotation cost, but generally achieves inferior detection accuracy. Therefore, many literature [27, 30, 43] explore the weakly supervised approach (*i.e.*, CDWSOD), which provides good trade-off between accuracy and annotation efficiency. Generally, CDWSOD improves cross-domain detection accuracy by adapting the deep model to the target domain with additional weak supervision signals (*i.e.*, the image-level annotations).

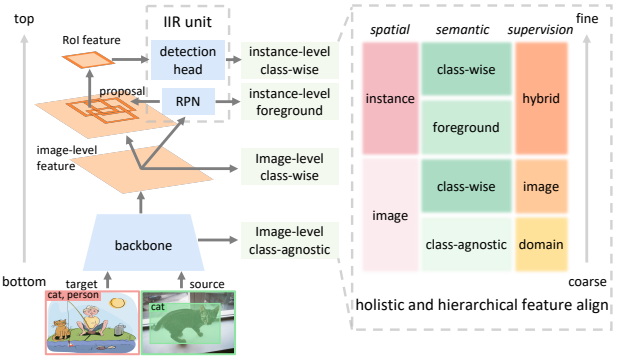

Figure 1. Our H^{2}FA R-CNN employs four feature alignments, *i.e.*, image-level (class-agnostic and class-wise) alignments and instance-level (foreground and class-wise) alignments for CDWSOD. From the viewpoints of spatial granularity, semantic granularity and the supervision signals, there is a clear hierarchy from coarse to fine, which is paced with the detection pipeline from bottom to top. The novel Instance- and Image-level Recognition (IIR) unit is based on the RPN and detection head and is compatible to both full and weak supervision signals.

We argue it is important to exploit the characteristics of the detection pipeline during the domain adaptation for CDWSOD. Specifically, a popular cross-domain detection baseline adopts the two-stage pipeline [45] and is consisted of a backbone, a region proposal network (RPN) and a detection head. While the common sense is that all these three components are critical to the detection accuracy, existing methods usually focus on partial components for do-

*Work done during an internship at Baidu Research.

main alignment. For instance, [27] aligns the backbone features and neglects aligning the RPN and detection head. Some self-training-based methods [30, 43] use instance-level pseudo labels for adaptive training and can be viewed as directly aligning the features in the detection head. In contrast to the previous literature, we believe that all these components are important for domain alignment.

Such motivated, we propose a novel CDWSOD method named Holistic and Hierarchical Feature Alignment (H^2FA) R-CNN, as illustrated in Figure 1. H^2FA R-CNN not only includes *holistic* detection components (*i.e.*, the backbone, the RPN and the detection head) for domain alignment, but also organizes these multiple alignments in a *hierarchical* sequence in pace with the detection pipeline. We explain the hierarchical sequence as below:

1) Two image-level alignments for the backbone: When the detection pipeline is within the backbone, the network processes each image as a whole. Correspondingly, we enforce two image-level alignments (a class-agnostic and a class-wise one) on the backbone features. Such a *class-agnostic* $\rightarrow$ *class-wise* sequence is paced with the fact that the backbone feature gradually develops class-wise discriminative ability from bottom to top layers.

Specifically, the class-agnostic alignment uses adversarial domain classifiers to pull close two domains, without categorizing each image. In contrast, the class-wise domain alignment employs a multi-label classification task to learn a set of class-wise prototypes (a single prototype for a respective class). During training, each prototype pulls close the features (of the corresponding class) from two domains, therefore facilitating the class-wise alignment.

2) Two instance-level alignments for the RPN and detection head: When the detection pipeline proceeds to the RPN and detection head, the network shifts to instance-level object recognition. Correspondingly, we enforce an instance-level foreground alignment for the RPN and an instance-level class-wise alignment for the detection head, respectively. Since the target domain does not provide instance-level but image-level annotations, we transform the vanilla RPN and detection head into a novel Instance- and Image-level Recognition (IIR) unit (as introduced below), which is compatible to both weak and full supervision. Such a *foreground* $\rightarrow$ *class-wise* sequence is paced with the two-stage hierarchy of the detection baseline.

Apart from the overall framework, another important characteristic of H^2FA R-CNN is the novel Instance- and Image-level Recognition (IIR) unit. IIR unit has two functions: 1) IIR preserves the original instance-level recognition function of the RPN and detection head; 2) IIR merges the outputs from the RPN and detection head for image-level recognition. Therefore, IIR can receive full/weak supervision from the source/target domains simultaneously, facilitating the desired instance-level alignments.

There is a clear hierarchy in the above aligning pipeline of image-level (class-agnostic $\rightarrow$ class-wise) $\Rightarrow$ instance-level (foreground $\rightarrow$ class-wise). In pace with the detection pipeline from bottom to top, the semantic and spatial granularity of the alignments are from coarse to fine (see Figure 1). Meanwhile, the supervision signals for learning these alignments are from weak to strong (*i.e.*, domain labels, image-level labels, and the hybrid *image-level plus instance-level* labels). We empirically show that the holistic and hierarchical characteristics are both important for H^2FA (see §4.4). Experimental results show that H^2FA R-CNN significantly improves the cross-domain object detection performance and sets new state of the art on popular benchmarks. Our main contributions can be summarized as follows:

- We propose the holistic and hierarchical feature alignment (H^2FA) R-CNN for the CDWSOD task. H^2FA R-CNN organizes two image-level and two instance-level alignments in a hierarchical manner.
- As an important component, we devise an Instance- and Image-level Recognition (IIR) unit to replace the vanilla RPN and detection head. IIR receives hybrid supervision from the source and target domains and facilitates instance-level alignments.
- We evaluate the proposed H^2FA R-CNN through comprehensive experiments. Experimental results demonstrate that H^2FA R-CNN not only achieves state-of-the-art cross-domain object detection performance, but also has the advantage of strong noise robustness.

2. Related Work

Object detection. Modern object detection methods [6, 40, 45] have achieved promising detection accuracy based on some large-scale datasets [17, 21, 41, 52]. However, deploying a well-trained detector to another novel domain may bring catastrophic performance degradation. This paper aims to mitigate the accuracy gap between detection and cross-domain detection with additional weak supervision on the target domain. We use Faster R-CNN [45] as our baseline model. Such a choice is consistent with most prior cross-domain object detection works [10, 15, 27, 30, 50, 68].

Weakly supervised object detection. Most prior weakly supervised object detection (WSOD) arts [3, 12, 29, 31, 46, 53, 54, 58, 62, 63, 73, 75, 77] focus on learning object detectors with only image-level annotations. They formulate WSOD as a multiple instance learning problem. More recently, several works [2, 5, 5, 16, 18, 26, 33, 47, 82] try boost WSOD performance with some instance-level annotations. Most of them [5, 16, 26, 33, 60, 82] focus on extending detectors to novel categories with image-level annotations. Compared with WSOD, the topic of this paper (*i.e.*, CDWSOD)

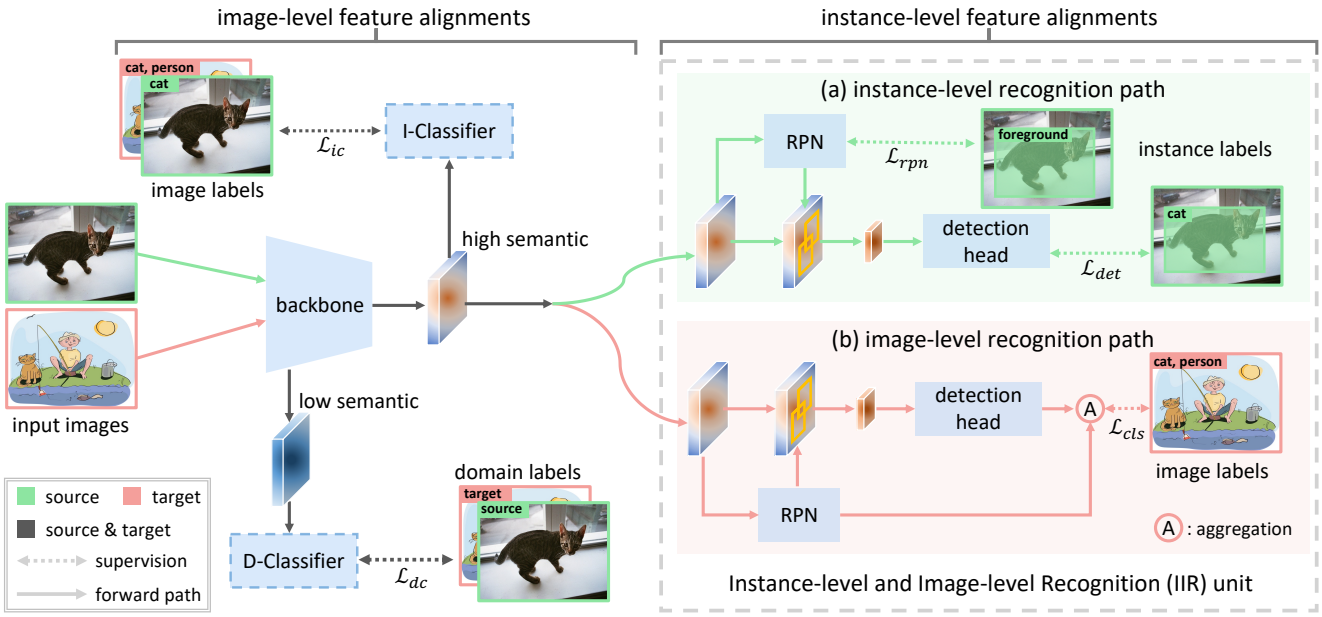

Figure 2. H^{2}FA R-CNN enforces two image-level alignments and two instance-level alignments from bottom to top. We use both the full supervision on the source domain and the weak supervision on the target domain for training H^{2}FA R-CNN. Within the backbone, we respectively use a D-Classifier and I-Classifier to enforce image-level class-agnostic and class-wise alignments. After the backbone, we further enforce two instance-level alignments using an Instance- and Image-level Recognition (IIR) unit. IIR constructs two different and parallel paths based on the RPN and detection head: 1) it uses the vanilla instance-level recognition path for detection and receives full supervision on the source domain; 2) it uses an image-level recognition path to receive weak supervision on the target domain.

is more challenging due to the severe domain shift. The proposed H^{2}FA R-CNN is featured for holistic and hierarchical feature alignment, which largely mitigates the domain shift problem.

Cross-domain object detection. Cross-domain object detection aims at detecting objects cross different domains. Most previous works solving cross-domain object detection mainly focus on UDAOD [4, 7, 15, 20, 24, 25, 42, 55, 57, 61, 65, 76, 79, 84]. Prior UDAOD methods can be roughly divided in to two groups, *i.e.*, adversarial feature alignment [9–11, 28, 48, 50, 59, 66, 68, 72, 78, 80, 81] and self-training [34, 35, 37, 44, 49]. Apart from the standard UDAOD setting, source-free and multi-source UDAOD tasks are studied in [38] and [48, 74] respectively. Moreover, domain generalization in object detection is explored in [39, 64]. Compared with UDAOD, CDWSOD [27, 30, 43] provides additional image-level annotations on the target domain and generally achieves superior accuracy. This paper well exploits the weak supervision signals on the target domain for domain alignment on four feature levels.

3. H^{2}FA R-CNN

3.1. Overview

As shown in Figure 2, H^{2}FA R-CNN adopts a two-stage detection framework [45] comprised of a backbone, a RPN, and a detection head. H^{2}FA R-CNN takes the mixture of source and target images as its input and seeks for domain alignment in the holistic and hierarchical manner. To this end, H^{2}FA R-CNN makes two changes to the baseline structure: 1) It appends extra domain classifiers (D-Classifier) and an image classifier (I-Classifier) to the backbone for image-level feature alignments. 2) It transforms the RPN and detection head into an Instance- and Image-level Recognition (IIR) unit for instance-level alignments.

When the detection pipeline is within the backbone, H^{2}FA R-CNN enforces two image-level alignments with D-Classifier and I-Classifier (see §3.2). Specifically, D-Classifier enforces image-level class-agnostic alignment on bottom-layer features, by learning to recognize the underlying domain of each image with a popular adversarial loss $\mathcal{L}_{dc}$ [50]. I-Classifier enforces image-level class-wise alignment on top-layer features, by learning multi-label classification through a binary cross-entropy loss $\mathcal{L}_{ic}$. We arrange I-Classifier (for class-wise alignment) behind D-Classifier (for class-agnostic alignment). Such arrangement is because the classification generally requires higher semantic information and thus favors the top-layer features.

After image-level alignments, H^{2}FA R-CNN further enforces instance-level alignments by transforming the RPN and detection head into an Instance- and Image-level Recognition (IIR) unit (see §3.3). The proposed IIR unit achieves two different functions via routing different paths on the shared RPN and detection head. The first path main-

tains the vanilla detection pipeline for the source domain and is supervised with popular detection losses [45], *i.e.*, $\mathcal{L}_{rpn}$ (for the RPN) and $\mathcal{L}_{det}$ (for the detection head). The second path aggregates the instance-level predictions of the RPN and detection head into image-level predictions. During training, IIR inputs the target domain features into the second path and supervises the corresponding image-level predictions with a binary cross-entropy loss $\mathcal{L}_{cls}$.

During training, H^{2}FA R-CNN aggregates all the above loss functions for an end-to-end optimization:

$$\mathcal{L} = \underbrace{\lambda_{dc}\mathcal{L}_{dc} + \lambda_{ic}\mathcal{L}_{ic}}_{\text{image-level align}} + \underbrace{\mathcal{L}_{rpn} + \mathcal{L}_{det} + \lambda_{cls}\mathcal{L}_{cls}}_{\text{instance-level align}}. \quad (1)$$

During inference, H^{2}FA R-CNN employs a standard inference pipeline as in the baseline [45]. In other words, we remove the D-Classifier, I-Classifier and restore the IIR unit into the vanilla RPN and detection head.

3.2. Image-level feature alignments

Class-agnostic alignment with domain supervision. Domain labels are freely available for all data. Using domain labels, we perform image-level class-agnostic feature alignment via adversarial training. Similar to [50], we attach two domain classifiers (D-Classifier) on the backbone. The D-Classifier try to distinguish which domain the input image belongs to. Meanwhile, the gradient reverse layers [19] reverse the gradients propagated by D-Classifier to confuse the backbone. By optimizing D-Classifier with domain supervision and the adversarial loss $\mathcal{L}_{dc}$ introduced in [50], backbone parameters gradually lose the ability to distinguish domains. Consequently, these parameters become domain irrelevant, enabling cross-domain features to be aligned in a unified space.

Class-wise alignment with weak supervision. Performing cross-domain feature alignment in a class-agnostic manner only ensures the global distributions of two domains are aligned [61]. However, it is tolerant for class-wise misalignment, which compromises the recognition accuracy. Since top layer of the backbone contains high-level semantics and already gains discriminative ability for each individual class, we further impose class-wise alignment to take the benefit of the image-level annotations for both domains.

Concretely, we add a multi-label image classifier (I-Classifier) at the top of backbone. Supervised by a binary cross-entropy loss $\mathcal{L}_{ic}$, I-Classifier learns a set of prototypes (one for each object class). During training, these prototypes pull the features from the corresponding classes towards themselves, regardless of the underlying domain. Consequently, the source and target domain features of a same class are pulled close towards each other, yielding the desired class-wise feature alignment.

We note that the I-Classifier in our H^{2}FA R-CNN is significantly different from those in previous UDAOD works [9, 71, 80], in terms of motivation and mechanism. These methods do not use the image-level classifier to directly align source and target domains. Instead, they only train the image-level classifier on the source domain and use it to enhance the effect of domain classifiers. In contrast, our I-Classifier explicitly aligns two domains along each individual class, as visualized in Figure 3b.

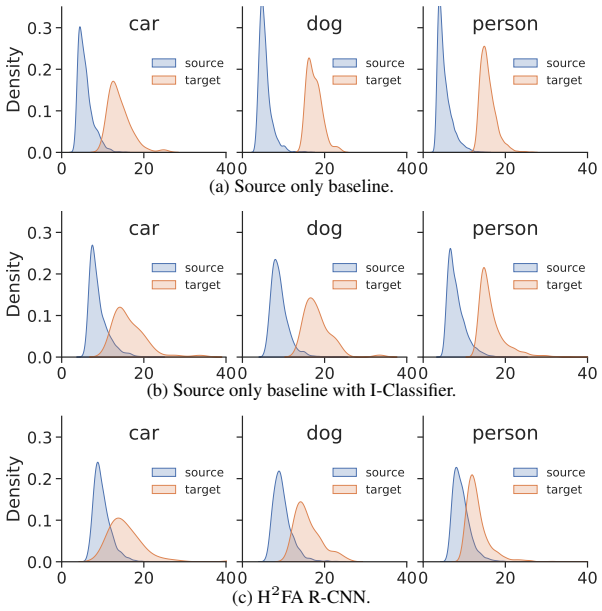

Figure 3. Within-class distributions of a *vehicle*, an *animal* and a *person* class on Watercolor [30] dataset. In (a), the model is trained only on the source domain and thus incurs significant domain gap along each class. In (b), adding the I-classifier mitigates the class-wise domain gap. In (c), H^{2}FA R-CNN combining all the feature alignments further mitigates the class-wise domain gap. More visualization examples are shown in the Appendix.

3.3. Instance-level feature alignments

Instance- and image-level recognition unit. After coarse alignment at image-level, H^{2}FA R-CNN further seeks for instance-level alignment. Enforcing instance-level alignment along each class is non-trivial, because target domain has no instance-level annotations. Prior self-training-based methods [30, 43] try to solve this problem by using pseudo labels and are vulnerable to the pseudo label noises. In contrast, H^{2}FA R-CNN directly tackles this problem with hybrid supervision (*i.e.*, combining full supervision on the source domain and weak supervision on the target domain) and thus requires no pseudo labels.

To this end, we propose a novel Instance- and Image-level Recognition (IIR) unit, which shares exactly the same modules (*i.e.*, a RPN and a detection head) and parameters for both the domains yet performs two different functions via routing different paths. For the source domain, IIR treats each proposal independently and produces instance-

level predictions as a regular detection model. For the target domain, IIR switches to an image-level recognition function which aggregates the multiple outputs of RPN and detection head into an image-level multi-label prediction and thus enables the weak supervision. We present the details as below:

• *Instance-level recognition path* adopts a standard two-stage pipeline [45] to generate instance-level predictions. The RPN first generates coarse object proposal candidates. Then, the detection head extracts instance-level features from these proposals for further refinement. As instance-level annotations of source domain are already available, we train IIR using standard object detection losses as in [45] (*i.e.*, $\mathcal{L}_{rpn}$ for the RPN, and $\mathcal{L}_{det}$ for the detection head).

• *Image-level recognition path* reuses the instance-level predictions from the RPN and detection head to generate image-level prediction. To this end, it considers representing the whole image with a few informative instances that 1) are more likely to cover complete objects, and 2) have a high probability belong to an individual class.

Let us assume for an image, the RPN predicts objectness logits $o \in \mathbb{R}^N$, while the detection head predicts the classification logits $x \in \mathbb{R}^{N \times \mathcal{C}}$ ($\mathcal{C}$ is the total number of object classes). We aggregate o and x for a single multi-class prediction, as shown in Figure 4. Specifically, we first assign the n-th objectness o_n to a specific object class according to $x_n \in \mathbb{R}^{\mathcal{C}}$: if the index for the largest-value entry of x_n is i, we assign o_n to the i-th entry and assign 0 objectness to all the other entries. In this way, we obtain a class-specific objectness matrix $\bar{o} \in \mathbb{R}^{N \times \mathcal{C}}$.

We use $\bar{o}$ and the classification logits x to generate the image-level prediction. We first extract the proposals' probabilities of belonging to each object class by a softmax along object classes (*i.e.*, softmax along row in Figure 4). Given these probability scores from multiple proposals, we use *weighted sum* to collect them into a single image-level probability. To this end, we leverage a softmax along proposals (*i.e.*, softmax along column in Figure 4) for assigning weights, as it provides normalization effect and naturally highlights the most representative proposals. Finally, the image-level predictions are obtained by accumulating all proposals. Formally, the aforementioned image-level prediction aggregation can be given as:

$$P_c = \sum_{n=1}^{N} (\sigma^{row}(x_{n,c}) \odot \sigma^{col}(\bar{o}_{n,c})), \tag{2}$$

where P_c is the predicted probability for class c (indicating whether current image contains objects of class c), $\sigma^{row}(\cdot)$ and $\sigma^{col}(\cdot)$ are softmax operations along row and column respectively, and $\odot$ is the element-wise product operation.

After obtaining the image-level multi-label predictions, we employ a binary cross-entropy loss $\mathcal{L}_{cls}$ for optimization. If class c exists in the current image, minimizing $\mathcal{L}_{cls}$ pushes P_c towards 1. In contrast, if class c is absent from the current image, minimizing $\mathcal{L}_{cls}$ pushes P_c towards 0.

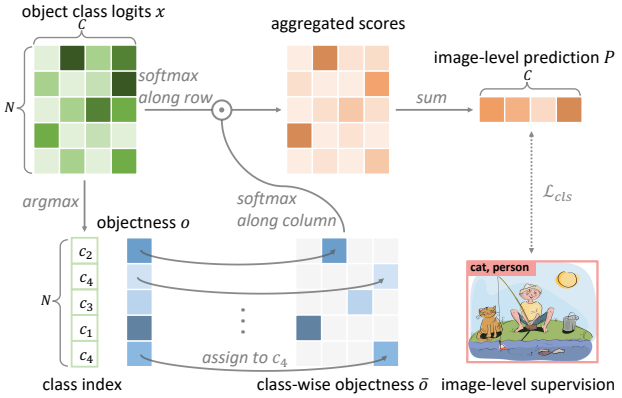

Figure 4. Pipeline of image-level prediction aggregation in image-level prediction recognition path, where $\odot$ indicates element-wise product. The object class logits are from the detection head, and the objectness logits are from the RPN.

Why IIR enforces instance-level alignments. In IIR, both the instance-level and the image-level recognition paths share the same RPN and detection head. It makes the RPN and detection head competent for recognizing the objects on both the source and target domains, and therefore mitigates the domain gap, as illustrated in Figure 3c. More concretely, we note that there is a learnable foreground prototype in the RPN. To recognize foreground on both domains, the foreground prototype aligns the objects from both domain towards itself. Similarly, the detection head contains a set of class-wise prototypes. These class-wise prototypes pull the objects of each corresponding class towards themselves, therefore aligning two domains along each class.

4. Experiments

4.1. Datasets

Following [8, 27, 30, 43, 50], we use four datasets, *i.e.*, PASCAL VOC (VOC) [17], Clipart, Watercolor and Comic [30] for evaluation. We use the general object detection benchmark VOC as the source domain, and use the rest three artistic painting datasets as the target domains. According to CDWSOD, the source domain provides instance-level annotations for training, while the target domain provides only image-level annotations.

The `trainval` split of VOC 0712 is used as source domain training data, which contains ∼16.5k real-world images of 20 object categories. Clipart has a `train` split and a `test` split. Each split contains 500 images of 20 object categories. Following prior arts [8, 27, 50], both splits are used for training and testing (we termed such adaptation task as $\text{Clipart}_{\texttt{all}}$). Meanwhile, as in [15, 30, 43], we also train using only the `train` split and evaluate on the `test` split (we termed such adaptation task as $\text{Clipart}_{\texttt{test}}$). Both Watercolor and Comic contain 2k images with 6 classes.

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
source-only	23.9	45.2	26.2	21.3	33.4	44.2	25.8	18.4	37.9	19.8	27.2	12.5	24.6	45.4	30.2	41.1	9.1	17.1	45.8	35.4	29.2
PCL [58]	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
EDRN [54]	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
SWDA [50]	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
HTD [8]	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
IIOD [67]	41.5	52.7	34.5	28.1	43.7	58.5	41.8	15.3	40.1	54.4	26.7	28.5	37.7	75.4	63.7	48.7	16.5	30.8	54.5	48.7	42.1
I^3Net [9]	30.0	67.0	32.5	21.8	29.2	62.5	41.3	11.6	37.1	39.4	27.4	19.3	25.0	67.4	55.2	42.9	19.5	36.2	50.7	39.3	37.8
DBGL [7]	28.5	52.3	34.3	32.8	38.6	66.4	38.2	25.3	39.9	47.4	23.9	17.9	38.9	78.3	61.2	51.7	26.2	28.9	56.8	44.5	41.6
DT+PL [30]	50.1	**75.0**	37.0	38.7	58.1	83.4	50.1	38.0	55.2	67.3	51.1	34.8	49.8	89.9	60.2	63.4	28.8	42.4	62.6	70.9	55.3
ICCM [27]	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
H^2FA R-CNN	**58.1**	73.0	**56.8**	**50.4**	**61.2**	**98.6**	**69.5**	**57.8**	**66.4**	**77.1**	**56.1**	**84.1**	**64.3**	**100.0**	**78.1**	**78.2**	43.5	**65.4**	**77.3**	**79.7**	**69.8**

Table 1. Mean AP performance (%) on $\text{Clipart}_{\texttt{all}}$.

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
source-only	23.3	52.5	22.2	26.1	34.4	46.5	28.2	13.6	43.3	15.9	38.9	3.0	29.4	48.1	36.7	44.7	14.3	5.5	38.6	24.9	29.5
WSDDN [3]	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 [32]	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
DM [36]	28.5	63.2	24.5	42.4	47.9	43.1	37.5	9.1	47.0	46.7	26.8	24.9	48.1	78.7	63.0	45.0	21.3	36.1	52.3	53.4	41.8
ATF [25]	41.9	67.0	27.4	36.4	41.0	48.5	42.0	13.1	39.2	75.1	33.4	7.9	41.2	56.2	61.4	50.6	42.0	25.0	53.1	39.1	42.1
UMT [15]	39.6	59.1	32.4	35.0	45.1	61.9	48.4	7.5	46.0	67.6	21.4	29.5	48.2	75.9	70.5	56.7	25.9	28.9	39.4	43.6	44.1
DT+PL [30]	**51.6**	**84.0**	30.0	41.1	52.3	82.0	50.2	19.0	51.8	58.3	41.3	14.6	47.0	86.2	61.9	**58.6**	24.9	22.5	47.4	52.8	48.9
PLGE [43]	43.4	52.5	29.4	40.1	30.4	71.9	**54.9**	3.6	**52.4**	73.8	**53.5**	24.0	**54.8**	**89.1**	65.1	40.5	32.3	**33.8**	45.4	**61.0**	47.6
H^2FA R-CNN	38.5	70.6	**38.9**	**47.4**	**59.6**	**83.5**	47.0	**29.3**	51.5	**76.3**	44.4	**48.1**	47.3	79.2	**75.7**	54.4	**53.9**	32.0	**56.6**	51.1	**55.3**
oracle	55.2	78.3	51.1	58.1	60.7	58.4	61.5	27.3	60.9	71.7	60.5	40.7	56.9	82.5	82.8	65.9	49.2	46.1	59.7	58.1	59.3

Table 2. Mean AP performance (%) on $\text{Clipart}_{\texttt{test}}$.

We use 1k images from the `train` split for training, and 1k images from the `test` split for testing.

4.2. Implementation details

Our proposed method is implemented and evaluated using `Detectron2` [69] and `PaddleDetection` [1]. The base framework is a two-stage detector Faster R-CNN [45] with RoIAlign [22], following [8, 15, 27, 50, 68]. ImageNet [14] pre-trained ResNet-101 [23] is utilized as our network backbone in all experiments, unless otherwise specified. We use a mini-batch size of 8 (4 images per domain) in 2 GPUs, and an initial learning rate is set to 0.005. The loss weights λ_{dc}, λ_{ic} and λ_{cls} are empirically set to 1, 0.1 and 1. Other hyper-parameters are the default setups in `Detectron2`, and we do not tune them ad hoc. For the $\text{Clipart1k}_{\texttt{all}}$, we train for 36k iterations with the learning rate multiplied by 0.1 at 24k and 32k iterations. For the rest three data splits, we train for 24k iterations with the learning rate multiplied by 0.1 at 16k and 21.5k iterations.

4.3. Main results

We compare H^2FA R-CNN against several baselines and state-of-the-art methods, including: 1) source-only baseline trained on the source domain with instance-level labels; 2) WSOD methods [3, 32, 54, 58] trained on the target domain with only image-level labels; 3) UDAOD methods [7, 15, 25, 50] trained with fully-labeled source domain and unlabeled target domain; 4) CDWSOD methods [27,30,43] trained with fully-labeled source domain and image-level labeled target domain; 5) oracle model trained with fully-labeled source and fully-labeled target domain. For fair comparison, we re-implement the Faster-RCNN variant of DT+PL [30] with ResNet-101 and use the processed intermediate data released by the authors[1].

	bike	bird	car	cat	dog	person	mean
source-only	77.6	39.0	46.7	21.5	16.2	47.5	41.4
PCL [58]	6.7	28.8	20.2	9.5	5.4	27.4	16.3
EDRN [54]	5.2	29.3	15.3	1.4	0.9	34.9	14.5
SWDA [50]	82.3	55.9	46.5	32.7	35.5	66.7	53.3
HTD [8]	69.2	49.5	49.5	34.9	30.8	61.2	49.2
ATF [25]	78.8	59.9	47.9	41.0	34.8	66.9	54.9
MCAR [80]	87.9	52.1	51.8	41.6	33.8	68.8	56.0
IIOD [67]	**95.8**	54.3	48.3	42.4	35.1	65.8	56.9
UMT [15]	88.2	55.3	51.7	39.8	43.6	69.9	58.1
I^3Net [9]	81.1	49.3	46.2	35.0	31.9	65.7	51.5
VDD [68]	90.0	56.6	49.2	39.5	38.8	65.3	56.6
DBGL [7]	83.1	49.3	50.6	39.8	38.7	51.3	53.8
DT+PL [30]	81.0	49.5	39.5	32.3	28.4	62.4	48.8
ICCM [27]	86.6	**64.2**	52.6	32.4	41.2	67.4	57.4
PLGE [43]	73.7	56.1	50.6	42.5	41.8	**74.6**	56.5
H^2FA R-CNN	88.6	52.4	**53.6**	**46.4**	**44.5**	73.8	**59.9**
oracle	73.3	65.5	57.3	45.7	37.3	80.5	59.9

Table 3. Mean AP performance (%) on Watercolor.

$\text{Clipart}_{\texttt{all}}$. Table 1 reports the cross-domain detection results on $\text{Clipart}_{\texttt{all}}$, where the target-domain training data and testing data are the same. We observe that H^2FA R-CNN achieves 69.8% mAP, surpassing all the compared approaches. Remarkably, H^2FA R-CNN outperforms previous state-of-the-art CDWSOD method [30] by 14.5%.

$\text{Clipart}_{\texttt{test}}$. Fewer target-domain training and testing data are available in $\text{Clipart}_{\texttt{test}}$. The comparison with previous arts is shown in Table 2. H^2FA R-CNN brings significant improvement (from 29.5% to 55.3% mAP) over the source-only model. Compared with the previous state-of-the-art [30], H^2FA R-CNN also shows 6.4% improvement in terms of mAP.

[1] https://github.com/naoto0804/cross-domain-detection/tree/master/datasets

	bike	bird	car	cat	dog	person	mean
source-only	43.2	10.7	24.1	9.1	11.7	20.9	19.9
PCL [58]	1.2	0.4	8.9	2.9	2.3	15.6	5.2
EDRN [54]	1.6	0.5	13.2	7.2	2.5	13.2	6.4
SWDA [50]	30.3	19.6	28.8	15.2	24.9	46.9	27.6
HTD [8]	35.4	14.8	26.6	13.7	26.9	40.0	26.2
MCAR [80]	47.9	20.5	37.4	20.6	24.5	50.2	33.5
I^3Net [9]	47.5	19.9	33.2	11.4	19.4	49.1	30.1
DBGL [7]	35.6	20.3	33.9	16.4	26.6	45.3	29.7
DT+PL [30]	53.0	23.7	34.4	27.4	27.2	44.0	35.0
ICCM [27]	50.6	23.3	35.4	32.3	33.8	47.1	37.1
PLGE [43]	55.0	21.2	40.0	35.1	37.9	60.9	41.7
H^2FA R-CNN	**55.3**	**26.6**	**45.9**	**38.1**	**45.6**	**66.8**	**46.4**
oracle	61.9	38.9	50.8	48.9	45.2	76.6	53.7

Table 4. Mean AP performance (%) on Comic.

Watercolor. Table 3 summarizes the comparison on Watercolor. While previous state-of-the-art method already approaches close the oracle accuracy, H^2FA R-CNN still outperforms the most competitive UMT [15] and ICCM [27] by $\sim 2\%$ mAP.

Comic. As shown in Table 4, Comic benchmark is a very challenging benchmark, where the source-only model only achieves 19.9% mAP. H^2FA R-CNN exceeds all the previous methods and largely closes the gap to the oracle model. Concretely, H^2FA R-CNN obtains 46.4% mAP, surpassing the second place PLGE [43] by 4.7%.

4.4. Ablation study

Table 5 investigates the characteristics of H^2FA R-CNN through ablation on four benchmarks. For brevity, we use A1, A2, A3, and A4 to indicate image-level class-agnostic, image-level class-wise, instance-level foreground and instance-level class-wise alignments from bottom to top, respectively. We draw three important observations:

Each individual alignment is beneficial. Comparing methods in Lines *(b)-(e)* against the source-only baseline in Line *(a)*, we observe that most individual alignment brings more or less improvement. For example, on Comic, the A1 to A4 alignments improve the source-only baseline by $+13.4\%$, $+11.9\%$, $+7.2\%$ and $+14.6\%$ mAP, respectively. Although A4 by itself sometimes decreases the source-only baseline (*e.g.*, -2.2% on $\text{Clipart}_{\texttt{test}}$), adding A4 over "A1+A2+A3" brings consistent improvement. The reason for A4 sometimes decreasing the baseline is the deteriorated hierarchy, as to be analyzed later.

The holistic alignment is important. Comparing Line *(i)* (the holistic alignments) against Lines *(a)-(e)*, we find that employing holistic alignments achieves the largest improvement. Specifically, the holistic improvement in Line *(i)* surpasses any individual improvement by a large margin.

The hierarchical alignment is important. While Line *(i)* (adding A4 upon "A1+A2+A3" in Line *(g)*) brings consistent improvement on all these benchmarks, we clearly observe the effect of A4 based on the source-only baseline is

	A1	A2	A3	A4	$\text{Clipart}_{\texttt{all}}$	$\text{Clipart}_{\texttt{test}}$	Watercolor	Comic
(a)					29.2	29.5	41.4	19.9
(b)	✓				37.1 (+7.9)	39.5 (+10.0)	49.6 (+8.2)	33.3 (+13.4)
(c)		✓			39.0 (+9.8)	33.0 (+3.5)	53.8 (+12.4)	31.8 (+11.9)
(d)			✓		50.8 (+21.6)	39.2 (+9.7)	42.0 (+0.6)	27.1 (+7.2)
(e)				✓	30.8 (+1.6)	27.3 (-2.2)	53.4 (+12.0)	34.5 (+14.6)
(f)	✓	✓			48.0 (+18.8)	44.2 (+14.7)	53.3 (+11.9)	39.6 (+19.7)
(g)	✓	✓	✓		63.1 (+33.9)	50.3 (+20.8)	49.3 (+7.9)	41.6 (+21.7)
(h)			✓	✓	59.1 (+29.9)	37.8 (+8.3)	55.0 (+13.6)	38.3 (+18.4)
(i)	✓	✓	✓	✓	69.8 (+40.6)	55.3 (+25.8)	59.9 (+18.5)	46.4 (+26.5)

Table 5. Effectiveness of different feature alignments in H^2FA R-CNN, where mean AP performance (%) over all classes is reported. A1-A4 denote the four different types of feature alignments from bottom to top, where A1 and A2 are image-level alignments and A3 and A4 are instance-level alignments. A more detailed ablation is provided in the Appendix.

	extra	bike	bird	car	cat	dog	person	mean
DT+PL [30]		81.0	49.5	39.5	32.3	28.4	62.4	48.8
H^2FA R-CNN		88.6	52.4	**53.6**	46.4	44.5	73.8	59.9
DT+PL [30]	✓	89.1	48.7	47.4	39.2	33.5	64.1	53.7
H^2FA R-CNN	✓	**90.2**	**57.5**	49.8	**48.0**	**53.0**	**77.2**	**62.6**

(a) Mean AP performance (%) on Watercolor

	extra	bike	bird	car	cat	dog	person	mean
DT+PL [30]		53.0	23.7	34.4	27.4	27.2	44.0	35.0
H^2FA R-CNN		55.3	26.6	45.9	38.1	45.6	**66.8**	46.4
DT+PL [30]	✓	**60.7**	28.8	38.6	37.9	33.5	51.0	41.7
H^2FA R-CNN	✓	60.2	**36.6**	**47.6**	**59.6**	**48.7**	65.3	**53.0**

(b) Mean AP performance (%) on Comic

Table 6. Comparison based on extra noisy target training data.

quite unstable. In other words, A4 relies on the coarse alignments of "A1+A2+A3" as its prerequisite to maintain consistent improvement. Without coarse alignments, the fine-grained A4 can be unstable. For example, on $\text{Clipart}_{\texttt{all}}$, the improvement is slight ($+1.6\%$ mAP). On $\text{Clipart}_{\texttt{test}}$, A4 even decreases on source-only baseline (-2.2% mAP). It indicates that without coarse alignments at the bottom, the fine-grained alignment at top can be unstable or even deteriorates the aligning effect.

4.5. Further empirical analysis

Evaluation on noisy target data. We evaluate H^2FA R-CNN under noisy target domain, where the image-level labels have considerable noises. To this end, we employ the `extra` splits of Watercolor and Comic datasets for extra $\sim$15.8k and $\sim$50.8k noisy training samples. The results are respectively summarized in Tables 6a and 6b, from which we draw two observations.

Firstly, using additional (noisy) training samples on the target domain further improves H^2FA R-CNN. For example, on Watercolor, H^2FA R-CNN achieves 2.7% mAP improvement with extra data. Secondly, compared with DT+PL [30], H^2FA R-CNN shows consistent and massive improvement, both with and without extra noisy data.

Generalization to similar domains. We evaluate H^2FA R-CNN under the small domain gap scenario (*i.e.*, adaptation from CityScapes [13] to Foggy Cityscapes [51]). Foggy Cityscapes is a dataset rendered from Cityscapes and sim-

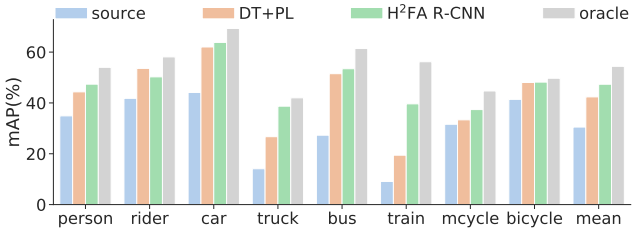

Figure 5. Mean AP (%) performance on Foggy Cityscapes.

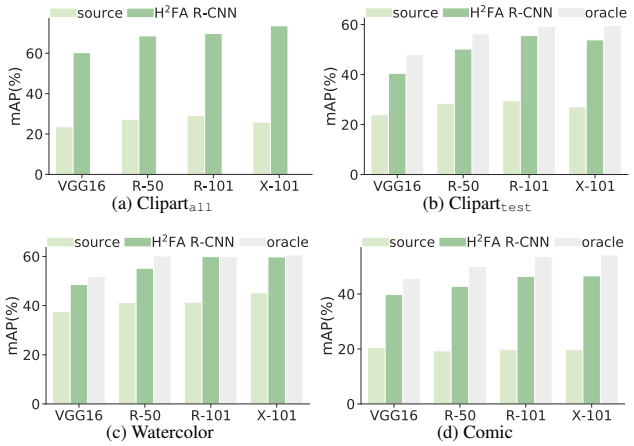

Figure 6. H^{2}FA R-CNN with various backbones. The *source* in legends denotes the source-only lower-bound. We omit the oracle for Clipart$_{all}$, because its training and testing set are the same.

ulates the foggy scenes. As shown in Figure 5, H^{2}FA R-CNN achieves significant improvement on all the categories (*e.g.*, $+19.7\%$ mAP on car), compared with the source-only baseline. It surpasses a recent state-of-the-art CDWSOD method [30] by a clear margin, as well. We thus infer that H^{2}FA R-CNN has strong generalization capacity towards similar domains.

Generalization to different backbones. We investigate different backbones including VGG16 [56], ResNet-50, ResNet-101 [23] and ResNeXt-101 [70] for H^{2}FA R-CNN. As illustrated in Figure 6, H^{2}FA R-CNN achieves consistent and considerable improvement over multiple backbones on all the four benchmarks. For the largest ResNeXt-101 backbone, the improvement of both H^{2}FA R-CNN and oracle is relatively small. This is probably because the small-scale training set on the target domain (*e.g.*, only 500 training samples for Clipart$_{test}$) becomes the bottleneck.

Analysis on training and inference efficiency. Figure 7 depicts the trade-off of the performance and time consumption. Compared with the source-only baseline, H^{2}FA brings significant improvement ($\sim$ 40% mAP), while incurring $\sim$ 3 hours additional training time. Compared with another CDWSOD method DT+PL [30], H^{2}FA achieves significant better training efficiency.

We infer the efficiency of training H^{2}FA R-CNN is due

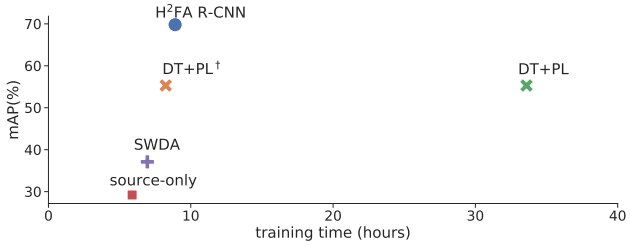

Figure 7. Comparison on training time and detection accuracy on Clipart$_{all}$. All models are trained with ResNet-101 backbone on 2 NVIDIA V100 GPUs. For DT+PL†, we only count the time for its detector training and exclude the time for Cycle GAN training.

to its end-to-end training manner. Specifically, H^{2}FA R-CNN is directly trained on the mixture of source and target domain without warm-up training. Therefore, although the feature alignments in H^{2}FA R-CNN adds several extra modules for training, the overall consumption is relatively low. In contrast, self-training-based CDWSOD methods [30, 43] typically require a warm-up training on the source domain, and sometimes requires extra for style-transfer using a dataset-specific CycleGAN [83].

During inference, we may directly remove all the additional components for feature alignments, making the structure exactly the same as Faster R-CNN [45]. In other words, H^{2}FA R-CNN brings no computational burden for inference, compared with the Faster R-CNN baseline.

5. Conclusion

This paper proposes a Holistic and Hierarchical Feature Alignment (H^{2}FA) R-CNN for cross-domain weakly supervised object detection (CDWSOD). H^{2}FA R-CNN enforces multiple domain alignments on all the critical detection components and organizes them in a hierarchy in pace with the detection pipeline. Apart from two image-level alignments for the backbone, H^{2}FA R-CNN imposes two instance-level alignments for the RPN and detection head. Such instance-level alignments are challenging and are jointly learned in a novel hybrid supervision manner. Comprehension experiments confirm that H^{2}FA R-CNN significantly improves CDWSOD, and sets new state of the art on popular benchmarks.

Limitations. Currently, all existing CDWSOD methods, including our method, assume that the target domain classes largely overlap with the source domain classes. Under practical scenarios, it is quite possible that the target domain has abundant classes that are novel to the source domain. How to further utilize these novel target domain classes deserves future exploration.

Acknowledgement. We thank all anonymous reviewers and area chairs for their constructive suggestions. The work of Yunqiu Xu was supported in part by the Chinese Scholarship Council under Grant 201908500109.

References

[1] PaddlePaddle Authors. PaddleDetection, object detection and instance segmentation toolkit based on PaddlePaddle. https://github.com/PaddlePaddle/PaddleDetection, 2019.

[2] Carlo Biffi, Steven McDonagh, Philip Torr, Aleš Leonardis, and Sarah Parisot. Many-shot from low-shot: Learning to annotate using mixed supervision for object detection. In *ECCV*, 2020.

[3] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016.

[4] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *CVPR*, 2019.

[5] Tianyue Cao, Lianyu Du, Xiaoyun Zhang, Siheng Chen, Ya Zhang, and Yan-Feng Wang. CaT: Weakly supervised object detection with category transfer. In *ICCV*, 2021.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[7] Chaoqi Chen, Jiongcheng Li, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. Dual bipartite graph learning: A general approach for domain adaptive object detection. In *ICCV*, 2021.

[8] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *CVPR*, 2020.

[9] Chaoqi Chen, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. I^3Net: Implicit instance-invariant network for adapting one-stage object detectors. In *CVPR*, 2021.

[10] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive Faster R-CNN for object detection in the wild. In *CVPR*, 2018.

[11] Yuhua Chen, Haoran Wang, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Scale-aware domain adaptive Faster R-CNN. *IJCV*, 2021.

[12] Ze Chen, Zhihang Fu, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. SLV: Spatial likelihood voting for weakly supervised object detection. In *CVPR*, 2020.

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[15] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, 2021.

[16] Bowen Dong, Zitong Huang, Yuelin Guo, Qilong Wang, Zhenxing Niu, and Wangmeng Zuo. Boosting weakly supervised object detection via learning bounding box adjusters. In *ICCV*, 2021.

[17] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 2010.

[18] Linpu Fang, Hang Xu, Zhili Liu, Sarah Parisot, and Zhenguo Li. EHSOD: CAM-guided end-to-end hybrid-supervised object detection with cascade refinement. In *AAAI*, 2020.

[19] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.

[20] Qiqi Gu, Qianyu Zhou, Minghao Xu, Zhengyang Feng, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. PIT: Position-invariant transform for cross-FoV domain adaptation. In *ICCV*, 2021.

[21] Agrim Gupta, Piotr Dollár, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019.

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[24] Zhenwei He and Lei Zhang. Multi-adversarial Faster-RCNN for unrestricted object detection. In *ICCV*, 2019.

[25] Zhenwei He and Lei Zhang. Domain adaptive object detection via asymmetric tri-way Faster-RCNN. In *ECCV*, 2020.

[26] Judy Hoffman, Sergio Guadarrama, Eric S Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. LSDA: Large scale detection through adaptation. *NeurIPS*, 2014.

[27] Luwei Hou, Yu Zhang, Kui Fu, and Jia Li. Informative and consistent correspondence mining for cross-domain weakly supervised object detection. In *CVPR*, 2021.

[28] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *ECCV*, 2020.

[29] Zeyi Huang, Yang Zou, Vijayakumar Bhagavatula, and Dong Huang. Comprehensive attention self-distillation for weakly-supervised object detection. In *NeurIPS*, 2020.

[30] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, 2018.

[31] Qifei Jia, Shikui Wei, Tao Ruan, Yufeng Zhao, and Yao Zhao. GradingNet: Towards providing reliable supervisions for weakly supervised object detection by grading the box candidates. In *AAAI*, 2021.

[32] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *ECCV*, 2016.

[33] Siddhesh Khandelwal, Raghav Goyal, and Leonid Sigal. UniT: Unified knowledge transfer for any-shot object detection and segmentation. In *CVPR*, 2021.

[34] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. In *ICCV*, 2019.

[35] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *ICCV*, 2019.

[36] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *CVPR*, 2019.

[37] Shuai Li, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Category dictionary guided unsupervised domain adaptation for object detection. In *AAAI*, 2021.

[38] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang. A free lunch for unsupervised domain adaptive object detection without source data. In *AAAI*, 2021.

[39] Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-invariant disentangled network for generalizable object detection. In *ICCV*, 2021.

[40] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

[41] 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 *ECCV*, 2014.

[42] Feng Liu, Xiaosong Zhang, Fang Wan, Xiangyang Ji, and Qixiang Ye. Domain contrast for domain adaptive object detection. *IEEE TCSVT*, 2021.

[43] Shengxiong Ouyang, Xinglu Wang, Kejie Lyu, and Yingming Li. Pseudo-label generation-evaluation framework for cross domain weakly supervised object detection. In *ICIP*, 2021.

[44] Rindra Ramamonjison, Amin Banitalebi-Dehkordi, Xinyu Kang, Xiaolong Bai, and Yong Zhang. SimROD: A simple adaptation method for robust object detection. In *ICCV*, 2021.

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015.

[46] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G. Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *CVPR*, 2020.

[47] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Alexander G. Schwing, and Jan Kautz. UFO^2: A unified framework towards omni-supervised object detection. In *ECCV*, 2020.

[48] Farzaneh Rezaeianaran, Rakshith Shetty, Rahaf Aljundi, Daniel Olmeda Reino, Shanshan Zhang, and Bernt Schiele. Seeking similarities over differences: Similarity-based domain alignment for adaptive object detection. In *ICCV*, 2021.

[49] Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In *CVPR*, 2019.

[50] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, 2019.

[51] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 2018.

[52] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019.

[53] Yunhang Shen, Rongrong Ji, Zhiwei Chen, Yongjian Wu, and Feiyue Huang. UWSOD: Toward fully-supervised-level capacity weakly supervised object detection. *NeurIPS*, 2020.

[54] Yunhang Shen, Rongrong Ji, Yan Wang, Zhiwei Chen, Feng Zheng, Feiyue Huang, and Yunsheng Wu. Enabling deep residual networks for weakly supervised object detection. In *ECCV*, 2020.

[55] Zhiqiang Shen, Mingyang Huang, Jianping Shi, Zechun Liu, Harsh Maheshwari, Yutong Zheng, Xiangyang Xue, Marios Savvides, and Thomas S Huang. CDTD: A large-scale cross-domain benchmark for instance-level image-to-image translation and domain adaptive object detection. *IJCV*, 2021.

[56] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[57] Peng Su, Kun Wang, Xingyu Zeng, Shixiang Tang, Dapeng Chen, Di Qiu, and Xiaogang Wang. Adapting object detectors with conditional domain normalization. In *ECCV*, 2020.

[58] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. PCL: Proposal cluster learning for weakly supervised object detection. *IEEE TPAMI*, 2018.

[59] Kun Tian, Chenghao Zhang, Ying Wang, Shiming Xiang, and Chunhong Pan. Knowledge mining and transferring for domain adaptive object detection. In *ICCV*, 2021.

[60] Jasper Uijlings, Stefan Popov, and Vittorio Ferrari. Revisiting knowledge transfer for training object class detectors. In *CVPR*, 2018.

[61] Vibashan VS, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. MeGA-CDA: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *CVPR*, 2021.

[62] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-MIL: Continuation multiple instance learning for weakly supervised object detection. In *CVPR*, 2019.

[63] Fang Wan, Pengxu Wei, Zhenjun Han, Jianbin Jiao, and Qixiang Ye. Min-entropy latent model for weakly supervised object detection. *IEEE TPAMI*, 2019.

[64] Xin Wang, Thomas E Huang, Benlin Liu, Fisher Yu, Xiaolong Wang, Joseph E Gonzalez, and Trevor Darrell. Robust object detection via instance-level temporal cycle confusion. In *ICCV*, 2021.

[65] Yu Wang, Rui Zhang, Shuo Zhang, Miao Li, YangYang Xia, XiShan Zhang, and ShaoLi Liu. Domain-specific suppression for adaptive object detection. In *CVPR*, 2021.

[66] Wang Wen, Cao Yang, Zhang Jing, He Fengxiang, Zha Zheng-Jun, Wen Yonggang, and Tao Dacheng. Exploring sequence feature alignment for domain adaptive detection transformers. In *ACM MM*, 2021.

[67] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Instance-invariant domain adaptive object detection via progressive disentanglement. *IEEE TPAMI*, 2021.

[68] Aming Wu, Rui Liu, Yahong Han, Linchao Zhu, and Yi Yang. Vector-decomposed disentanglement for domain-invariant object detection. In *ICCV*, 2021.

[69] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[70] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.

[71] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *CVPR*, 2020.

[72] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *CVPR*, 2020.

[73] Yunqiu Xu, Chunluan Zhou, Xin Yu, Bin Xiao, and Yi Yang. Pyramidal multiple instance detection network with mask guided self-correction for weakly supervised object detection. *IEEE TIP*, 2021.

[74] Xingxu Yao, Sicheng Zhao, Pengfei Xu, and Jufeng Yang. Multi-source domain adaptation for object detection. In *ICCV*, 2021.

[75] Yufei Yin, Jiajun Deng, Wengang Zhou, and Houqiang Li. Instance mining with class feature banks for weakly supervised object detection. In *AAAI*, 2021.

[76] Bo Zhang, Tao Chen, Bin Wang, Xiaofeng Wu, Liming Zhang, and Jiayuan Fan. Densely semantic enhancement for domain adaptive region-free detectors. *IEEE TCSVT*, 2021.

[77] Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: A survey. *IEEE TPAMI*, 2021.

[78] Yixin Zhang, Zilei Wang, and Yushi Mao. RPN prototype alignment for domain adaptive object detector. In *CVPR*, 2021.

[79] Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin. Collaborative training between region proposal localization and classification for domain adaptive object detection. In *ECCV*, 2020.

[80] Zhen Zhao, Yuhong Guo, Haifeng Shen, and Jieping Ye. Adaptive object detection with dual multi-label prediction. In *ECCV*, 2020.

[81] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *CVPR*, 2020.

[82] Yuanyi Zhong, Jianfeng Wang, Jian Peng, and Lei Zhang. Boosting weakly supervised object detection with progressive knowledge transfer. In *ECCV*, 2020.

[83] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

[84] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *CVPR*, 2019.