# Category aware and Scale aware Unsupervised Domain Adaptive Object Detection

Jayeon Yoo[1] Chaerin Kong[1] JunHoo Lee[1] Jooeun Kim[2]
Department of Intelligence and Information [1] Department of Data Science[2]
Seoul National University
{jayeon.yoo, veztylord, mrjunoo, kje980714}@snu.ac.kr

## Abstract

*Training object detectors typically requires large scale dataset with heavy annotations, but their performance significantly degrades even under relatively mild domain shifts. As collecting new sets of labeled data for object detection can be very costly, unsupervised approach for domain adaptation has drawn enormous attention from the community. One of the most widely used methods for unsupervised domain adaptation is to align the features of the source and target domain to have the same distribution through adversarial training, which intends to learn generalizable features to deceive the auxiliary domain classifiers. Based on the observation that object-level feature distribution heavily depends on the object category and its size, we propose a novel class-aware and scale-aware conditional feature alignment framework that actively incorporates class and scale information into the adversarial learning process. Extensive evaluations on two of the most widely used benchmarks for domain adaptation, i.e., CS2Foggy and Sim10k2CS, demonstrate the effectiveness of our method in FCOS based domain adaptive object detection.*

## 1. Introduction

These days, deep learning has shown great performance in various computer vision tasks. However, if the distribution of test data is far from the distribution of training data, the model performance is greatly degraded. In a real environment, the domain of data that the model needs to operate can be very diverse, so this performance degradation problem is very critical. For example, the model trained with the images which were taken in daytime will perform poorly on the images taken in nighttime. To solve this problem, the model must be retrained with a lot of data with the changed distribution whenever the data distribution changes. However labeling data to retrain the model is very expensive. In particular, labeling is a very labor-intensive process for object detection which requires the bounding boxes and the classes of each instances in images. Unsupervised Domain Adaptation (UDA) provides an efficient solution to this domain-shift problem by training the model to be domain invariant using labeled source domain data and unlabeled target domain data.

To deal with UDA for object detection, many methods have studied in three sections: adversarial learning, image translation, and self-training. Starting with [3], most of the studies have focused on aligning the feature distribution of source and target domain globally or class-wise using a domain discriminator based on the theoretical analysis of DANN. Several other works [2, 12] have allowed the model to learn from the target domain with ground truth labels by translating the source domain images into the target domain style and training the model with the translated images. These methods translate images using CycleGAN and Fourier transform, which has a limitation in that original target domain images and the translated target-style images still have inconsistency. Recently, self-training based methods have been proposed. In [17], the model which is trained on the source domain generates pseudo labels for the unlabeled target domain and retrain the model with the pseudo labeled target domain data. Other works [14] suggested a method in which a student model is trained to follow predictions from a more consistent teacher network using the Mean Teacher framework widely used for semi-supervised learning.

Fig. 2 shows the TSNE of backbone features located at the center of objects among the backbone features of FCOS. In Fig. 2a, each color refers to difference classes. In the case of Cityscapes dataset, there are total 8 classes. Although the distribution is not very clear, it can be seen that features corresponding to classes such as person, car, rider, and bicycle show different distributions. Fig. 2b shows the feature distribution in another perspective, the scale of bounding boxes. The color means the scale of bounding boxes: the darker the color, the larger the scale of bounding boxes, and


Figure 1: These visual differences between domains, *i.e.,* domain gaps, severely harm the performance of conventional object detectors.

vice versa. It shows that the backbone features have the distribution according to the size of the corresponding objects. The tendency is more pronounced when visualizing a single category, such as the class car in Fig. 2c. Although the features are from the same category (e.g. car), they have difference distributions according to the size of the objects. The feature of larger object are closer to the larger one than the features of smaller objects. For object detection which requires not only classification but also bounding box regression, especially for FCOS which predicts the values of left, top, right, bottom of bounding boxes, the features are gathered between same classes and are distributed along the size of the objects. Paying attention to this analysis, we will not only align features of source and target domain with the same class, but also align features conditionally according to the size of the object.

Hence, in this paper, we propose a novel framework for feature alignment with object category-aware and scale-aware conditioning, which can be directly incorporated upon aforementioned FCOS detector. To this end, we train a convolutional domain classifier along with our detector, to which the outer product of the feature vector and our model's prediction is provided as the feature map. This, in essence, is equivalent to intentionally forming a larger dimensional discriminator embedding space roughly partitioned by the class and scale conditions, where the model prediction vector is expected to assign the corresponding feature to the appropriate ghetto. Since it is not straightforward to train a conditional discriminator with continuous conditions such as object scale, we opt for the simplest choice of binning the scale values and using the resultant one-hot vectors instead. With this formulation, we can efficiently model conditional input with a single strong discriminator, opening up the doors for class- and scale-aware feature alignment in a very simple yet effective manner. Our contributions can be summarized as:

- We demonstrate the distributional discrepancy of FCOS features with differing object scales.
- We introduce a novel framework for adversarial feature alignment that conditions on object category and scale in a highly efficient manner.
- We show the effectiveness of our proposed method through empirical evaluations on two widely used domain adaptation benchmarks, CS2Foggy and Sim10k2CS.

## 2. Related Works

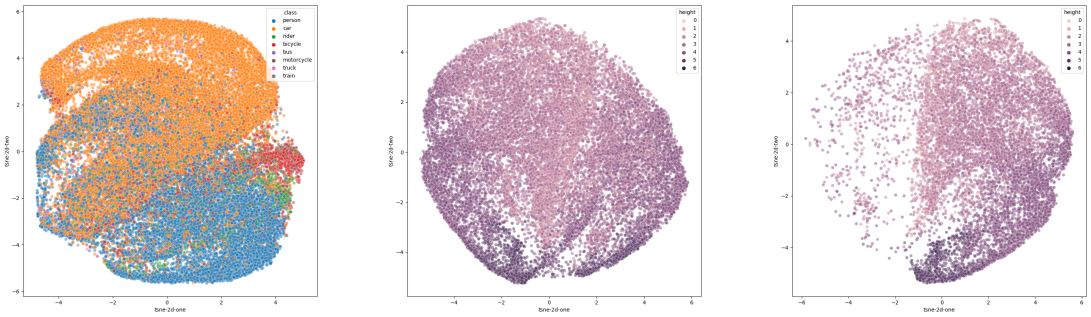
### 2.1. Domain Adaptive Object Detection

Numerous methods regarding object detection and domain adaptation were developed, but it was only until recently that domain adaptive object detection models with various approach arose.

**Adversarial Learning** Chen *et al.* [3] first proposed domain adaptive object detector by extending the idea of domain adaptation to existing object detection models. Based on Faster R-CNN [24], they used two domain adaptation components for global image-level and local instance-level each. Adversarial learning was adopted in order to train the domain classifiers while minimizing the domain distance. This work was elaborated [25] by giving more focus on local instance-level alignment and applying weak global alignment.

**Image Translation** Pixel-level domain adaptation using image translation used intermediate domain in CycleGAN [33] for domain adaptation. Hsu *et al.* [11] attempt to bridge the gap between source and target domains with an intermediate domains. With this approach, domain adaptation is decomposed into two subtasks, which is to first align the source domain to the intermediate domain, and then the intermediate domain progressively learns the target domain. Chen *et al.* [2] aims to focus more on feature representations of object detectors. To enhance the existing adversarial-based methods, the potential contradiction between transferability and discriminability were hierarchically harmonized.

**Conditional Align** Motivated by conditional generative adversarial networks [21], conditional domain adversarial networks [19] were proposed to condition discriminative information conveyed in the classifier predictions in adversarial adaptation models. CDANs adopted multilinear conditioning to capture domain-specific feature representations and predictions. Further, domain discriminators are conditioned on the uncertainty of classifier predictions, prioritizing the discriminator on easy-to-transfer examples. Recent works in conditional unsupervised domain adaptation are based on the ideas proposed in [19] and tend to focus on class-wise alignment. Inspired from [19], Zhao *et al.*[32] utilizes the multi-label prediction probability to perform conditional global feature alignment. Li *et al.*[18] elaborates on the idea of conditional alignment as a collaborative class conditional generative adversarial net to bypass the dependence on the source data. With the concept of collaboration between the generator and the prediction model



(a) Distribution of features according to the classes.

(b) Distribution of features according to box heights.

(c) Distribution of features corresponding to car according to box heights.

Figure 2: TSNE of backbone features of the source domain. The color means (a) the classes (b, c) the scale of the height of GT bounding boxes, which means that the darker the color, the larger the height value.

without the source data, [18] applies weight constraint to encourage similarity to the source model for indirect supervision. The idea of class conditional domain adaptation that attempted to remove pseudo-label bias from the existing models was proposed by Jiang *et al.*[13], who presented a sampling-based implicit alignment approach, where sample selection is implicitly guided by the pseudo-labels. MeGA-CDA [30] was also proposed to address negative transfer of features during class-agnostic domain alignment. This method trains category-wise discriminators and generates memory-guided category-specific attention map for target features in which class information is not available.

### 2.2. Improving Robustness

Despite its broad terminological practice, approaches designed to improve robustness typically aim for smooth function approximations. [31, 1, 29] leverage mixup based data augmentation to encourage linear behaviors in-between observed train examples. Modifications on the objective function have also been made. [20] formulates optimization scheme more robust to diverse adversarial attacks, while [5] proposes to seek local minima with flat neighborhood landscape. Empirical studies show that pretraining [8], self-supervised learning [9] and training models with shape bias [6] improve model robustness. On distribution shift, adversarial training has also been shown beneficial for transfer performance [28].

Robustness is an important property for object detection models as well. [16] learns object detector more robust to distribution shift by using noisy labels. [12] demonstrates the effect of adversarial training for robust domain adaptation. Recently, [23] proposes domain-centric data augmentation and effective knowledge distillation that show impressive results on diverse benchmarks.

## 3. Method

We propose a method of category-aware and object scale-aware adversarial feature aligning for FCOS object detector. Section 3.1 briefly describes FCOS [27], one-stage detector we use, to explain our concept more clearly. Section 3.2 describes conditional adversarial aligning method which is a simple and effective for category-aware feature aligning by exploiting the backbone features and category predictions simultaneously. Section 3.3 applies conditional adversarial alignment to the object scale to align backbone features according to the object scale. Finally, Section 3.4 summarizes the overall objectives.

### 3.1. Preliminary: FCOS

FCOS is a one-stage detector that predicts object categories and bounding boxes directly from feature maps without a region proposal network. FCOS configures feature pyramids of five levels using backbone features and is trained to predict object categories, centerness, and left, top, right, and bottom distances to the closest bounding box from each location in the feature map as shown in Fig. 4. In the inference stage, the more accurately the features located at the center of the object predict the result, so the object is detected based on the prediction of the feature with high classification confidence and centerness.

We observe that the backbone features are distributed according to not only object categories but also object sizes especially for FCOS because it predicts the left, top, right, and bottom distance to the bounding boxes from the current location. It is different from that Faster-RCNN, a representative two-stage detector, regress four values to correct the proposal which is generated from a Region Proposal Network. Focusing on this point, we align the features of the source and the target domains in category-aware and scale-

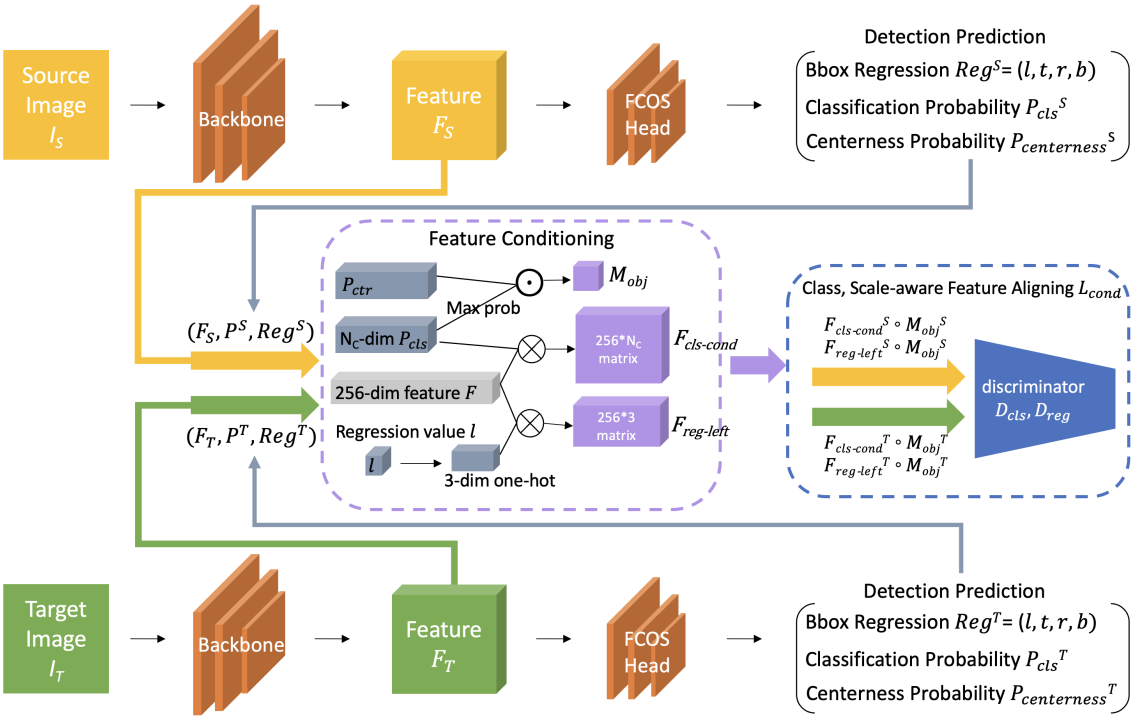



Figure 3: Overview of our model

aware based on the prediction of classification and regression.

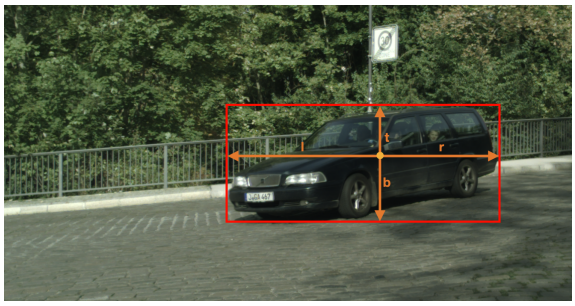



Figure 4: FCOS predicts object categories, centerness, and left, top, right, and bottom distances to the bounding box from each location in the feature map.

### 3.2. Category-aware feature alignment

In order for the object detector trained in the source domain to adapt well in the target domain and perform object classification and bounding box regression well, features corresponding to foreground objects in the source domain image must have a distribution similar to that of the target domain. Therefore, in order to focus on aligning features corresponding to object other than background in the feature maps, we generate a mask $M_{obj}$ by multiplying classification confidence and centerness score $P_{centerness}$as in Eq. 1. We use the maximum probability value among classes at each location to obtain classification confidence. The higher the class confidence and centerness score, the higher the probability of being a foreground object, so by applying the mask to the feature maps, we can obtain foreground object features.

$$M_{obj} = \max_{C}(P_{cls}) \odot P_{centerness} \tag{1}$$

Unlike EPM [10], which entirely aligns features corresponding to foreground objects, we align those features to have the same distribution by category through the conditional adversarial alignment method. Let $F^{u,v}_{backbone} \in \mathbb{R}^{dim_f}$ be the feature vector located at $(u,v)$-th position in the backbone feature maps and $P^{u,v}_{cls} \in \mathbb{R}^{N_C}$ is a classification probability vector of FCOS head at the same location $(u,v)$ when $N_C$ is the number of categories. By flattening the matrix, which is the result of the outer product of the original feature vector $F^{u,v}_{backbone}$ and the class probability vector $P^{u,v}_{cls}$, we can obtain a new feature vector $F^{u,v}_{cls\text{-}cond} \in \mathbb{R}^{dim_f \times N_C}$ conditioned on classification prediction. This new feature map $F_{cls\text{-}cond}$ would be fed into the discriminator with mask $M_{obj}$. Conditioning through outer product has the effect of increasing the dimension while allowing the feature to have different subspaces according to the class probability. For example, suppose an object detection problem that classifies objects into three categories. Sup-

pose that the backbone feature vector $f_1 \in \mathbb{R}^d$ located at the $(u_1, v_1)$-th position is masked in with high classification confidence and high centerness, and the corresponding classification probability $P_1$ is (0, 0.99, 0.01). The dimension of the conditioned feature $f_{cls\text{-}cond\ 1}$ flattened the matrix made of the outer product would be three times the original feature dimension. Since the probability of the second class is the highest, the second part (i.e. $(d+1, d+2, ..., 2d)$-th element) of that conditioned feature $f_{cls\text{-}cond\ 1}$ would be similar to the original feature vector $f_1$, but the other elements would be almost zero. As such, features with the highest probability of the second class have feature values in the second part of the conditioned feature, which is increased in dimension, and are trained to have a similar distribution of sources and targets in this subspace. On the other hand, features with the greatest probability of the first or third class have a feature value in the first ($(1, 2, ..., d)$-th element) or third part ($(2d+1, 2d+2, ..., 3d)$-th element) when the dimension increases, resulting in a feature distribution on different subspaces. However, in the early stages of learning, even if the classification probability has high confidence, it may not be accurate, so we use an equal probability value $\frac{1}{N_C}$ for all classes for conditioning, and then gradually use the prediction value $P_{cls}$ as iteration increases as Eq. 2. In all experiments, we set $Iter$ to 6000 which is half of the first learning rate decay point and $q$ to 0, initially starting with even conditioning for all classes and increase the usage rate of predictions linearly, allowing only prediction value to be used after 6000 iteration.

$$\begin{aligned} \alpha &= max(1 - \frac{iter_{cur}}{Iter}, q) \\ F^{u,v}_{cls\text{-}cond} &= F^{u,v}_{backbone} \otimes ((1-\alpha)P^{u,v}_{cls} + \alpha\frac{1}{N_C}\mathbb{1}) \end{aligned} \tag{2}$$

The discriminator $D_{cls}$ and backbone is trained as shown in Eq. 3. The commonly used gradient reverse layer is applied to train backbone to embed features so that features cannot be distinguished from which domain they come from. As a result, we can align the feature distribution of source and target domain in a category-aware manner by using the high confident classification prediction.

$$\begin{aligned} \mathcal{L}_{cls}(I_S, I_T) = \sum_{u,v} d\log(D_{cls}((F^S_{cls\text{-}cond} \odot M^S_{obj})^{u,v})) + \\ (1-d)\log(1 - D_{cls}((F^T_{cls\text{-}cond} \odot M^T_{obj})^{u,v})) \end{aligned} \tag{3}$$

### 3.3. Scale-aware feature alignment

We will use the prediction value of bounding box regression for conditioning similar to Sec.3.2 to align the feature according to the object scale. However, in the case of bounding box regression, unlike classification, there is a big difference in predicting it as a continuous value. We simply bin the continuous regression value into three classes and change it to a categorical prediction to conduct conditioning.

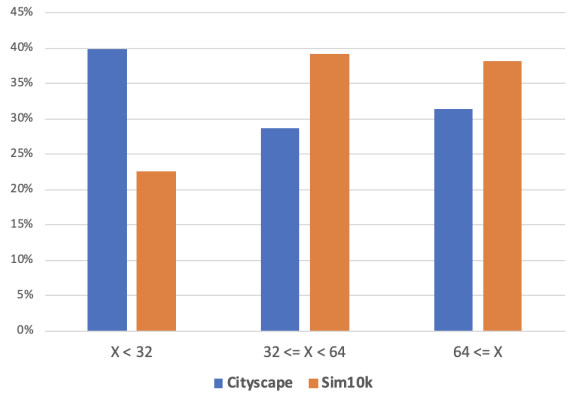



Figure 5: The distribution of the width of the ground truth bounding boxes of all objects in Cityscape and Sim10k dataset. It is divided into three categories (the width is less than 32; more than 32 and less than 64; more than 64) and binning is performed.

Based on the distribution of the width value of the gt binding box of all objects in the dataset (Fig.5), we determine the binning threshold heruristically so that the object does not tilt too much to a specific bin. FCOS head predicts four values of $(l, t, r, b)$ at each location of the feature map and we use only $l$ (left) and $t$ (top) value among these regression values. This is because $l$ (left) and $r$ (right), $t$ (top) and $b$ (bottom) have similar tendencies since we focus on features with high centerness by multiplying mask $M_{obj}$ like Sec.3.2. By binning the regression value $l$ as shown in Eq.4, we can obtain one-hot vectors of three categories for left regression value. We apply the same binning method to top regression value $t$.

$$P^{u,v}_{reg\text{-}left} = \begin{cases} (1,0,0) & \text{if } l^{u,v} < 16 \\ (0,1,0) & \text{if } 16 \le l^{u,v} < 32 \\ (0,0,1) & \text{if } 32 \le l^{u,v} \end{cases} \tag{4}$$

Since the continuous regression value is changed to a one-hot category value, we can apply the feature conditioning in Sec.3.2 equally as in Eq.5.

$$\begin{aligned} F^{u,v}_{reg\text{-}cond} &= F^{u,v}_{backbone} \otimes ((1-\alpha)P^{u,v}_{reg} + \alpha\frac{1}{3}\mathbb{1}) \\ \mathcal{L}_{reg}(I_S, I_T) &= \sum_{u,v} d\log(D_{reg}((F^S_{reg\text{-}cond} \odot M^S_{obj})^{u,v})) + \\ & (1-d)\log(1 - D_{reg}((F^T_{reg\text{-}cond} \odot M^T_{obj})^{u,v})) \end{aligned} \tag{5}$$

Like Sec.3.2, as the alpha value $\alpha$ gradually decreases from 1 to 0, and at the beginning when the prediction is inaccurate, all object scales are conditioned equally, and as

learning progresses, each scale is hard-conditioned. As a result, we can align the features of the source and target domain in object scale-aware manner.

### 3.4. Overall objective

Using labeled source domain data, FCOS backbone and FCOS head are trained for object detection tasks consisting of object classification and bounding box regression. The object detection loss is represented as $L_{det}$.

$$\mathcal{L}_{det}(I_S) = \mathcal{L}_{det\text{-}cls} + \mathcal{L}_{det\text{-}reg} \tag{6}$$

Following EPM [10] which is the first domain adaptive object detection method based on FCOS, we also align feature distribution globally. It is the same as the most basic structure of the adversarial aligning method by feeding the entire backbone feature into the discriminator to distinguish whether it is from a source domain or a target domain. We can narrow the overall domain gap in the image-level feature distribution between the source domain and the target domain by global aligning, $L_{global}$.

$$\begin{aligned}\mathcal{L}_{global}(I_S, I_T) = \sum_{u,v} d\log(D_{global}(F^{u,v}_{backbone})) \\ +(1-d)\log(1 - D_{global}(F^{u,v}_{backbone}))\end{aligned} \tag{7}$$

Finally, the total loss is the addition of detection loss $\mathcal{L}_{det}$, global aligning loss $\mathcal{L}_{global}$, category-aware aligning loss $\mathcal{L}_{cls}$ and sclae-aware aligning loss $\mathcal{L}_{reg}$ as in Eq. 8.

$$\begin{aligned}\mathcal{L}_{total} = \mathcal{L}_{det}(I_S) + \lambda_{global}\mathcal{L}_{global}(I_S, I_T) \\ +\lambda_{cond}(\mathcal{L}_{cls}(I_S, I_T) + \mathcal{L}_{reg}(I_S, I_T))\end{aligned} \tag{8}$$

## 4. Experimental Results

### 4.1. Implementation Details

We conduct experiments based on FCOS using VGG16 as the backbone. We use an Image-Net pretrained model and reduce the overall domain gap using only object detection loss and global alignment at the beginning of training following EPM [10]. Then, we train the model for 20000 iteration with weight decay 1e-4, initial learning rate 0.02 for Cityscape to Foggy Cityscape and 0.01 for Sim10k to Cityscape, respectively. We decay the learning rate at 12000 and 18000 iteration by the rate of one-tenth. During training, $\lambda_{global}$ and $\lambda_{cond}$ are fixed to 0.01 and 0.1, respectively. We set the weight for the Gradient Reversal Layer (GRL) to 0.02 for global aligning and 0.2 for our conditional aligning. Also, in order to reduce the effects of incorrect predictions with high confidence in the early stages of learning, we set $Iter$ to 6000 which is half of the first learning rate decay point and $q$ to 0. Therefore, after 6000 iterations, only the predicted probability is used for conditioning features. For Cityscapes to Foggy Cityscapes benchmark set, both category-aware and scale-aware feature alignment are applied, but for Sim10k to Cityscapes, only scale-aware feature alignment is applied because there is only one category, car, in that benchmark set. Input image is resized to 800 for short side, and 1333 to long side.

### 4.2. Datasets

We conduct experiments for two scenarios as described in Sec. 4.1. One is the adverse weather driving adaptation scenario and the other is learning from synthetic data scenario.

**Driving in Adverse Weather** Cityscapes [4] consists of clear city domain images under driving scenarios, summing to 2975 and 500 images for training and validation, respectively. There are 8 categories, *i.e.,* person, rider, car, truck, bus, train, motorcycle and bicycle. And Foggy Cityscapes [26] are made of synthetic dataset which transferred original Cityscapes images into foggy domain. We used Cityscapes data as the source domain, and Foggy Cityscapes as the target to simulate domain shift caused by the weather condition.

**Learning from synthetic data** Sim10k [15] consists of 10,000 synthesized city-domain images with their corresponding bounding box annotations. In this experiment, we set Sim10k as the source domain and Cityscapes as the target domain. So this adaptation tests the model whether it can adapt well from synthesized domain to real world domain. Only the class *car* is considered.

### 4.3. Overall Performance

We denote category-aware feature aligning as *Category*, and scale-aware feature aligning as *Bbox-width* and *Bbox-height* for conditioning on $l$ (left) and $t$ (top) value, respectively. The backbone architecture is VGG-16. Source Only refers to the case where the model is trained only on the source domain data without domain adaptation. Oracle refers to the model which learned from target label in supervised setting, juxtaposed to demonstrate our adaptation ability in a relative manner.

**Driving in Adverse Weather** We conduct experiments with category-aware and scale-aware feature aligning, as well as with a method of applying both of them. All of our methods outperforms other methods significantly regardless the type detector as shown in table 1. The number of EPM is taken from the original paper, which set the initial learning rate to 0.005 and train without learning decay. EPM* is the results of [10] with the same initial learning rate of 0.02 and scheduling as ours to make a fair comparison. The results show that each single module, category-aware alignment and scale-aware alignment alone, are effective enough to increase $\text{mAP}^r_{0.5}$ by 2.8 and 2.7, respectively,

Table 1: Results of Cityscape to Foggy Cityscape. * denotes the results of our re-implementation.

| Method | Detector | person | rider | car | truck | bus | train | mbike | bicycle | $mAP^r_{0.5}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Source Only | Faster-RCNN | 17.8 | 23.6 | 27.1 | 11.9 | 23.8 | 9.1 | 14.4 | 22.8 | 18.8 |
| DAF [3] | | 25.0 | 31.0 | 40.5 | 22.1 | 35.3 | 20.2 | 20.0 | 27.1 | 27.6 |
| SC-DA [34] | | 33.5 | 38.0 | 48.5 | 26.5 | 39.0 | 23.3 | 28.0 | 33.6 | 33.8 |
| MAF [7] | | 28.2 | 39.5 | 43.9 | 23.8 | 39.9 | 33.3 | 29.2 | 33.9 | 34.0 |
| SW-DA[25] | | 29.9 | 42.3 | 43.5 | 24.5 | 36.2 | 32.6 | 30.0 | 35.3 | 34.3 |
| DAM [17] | | 30.8 | 40.5 | 44.3 | 27.2 | 38.4 | 34.5 | 28.4 | 32.2 | 34.6 |
| MeGA-CDA [30] | | 37.7 | 49.0 | 52.4 | 25.4 | 49.2 | 46.9 | 34.5 | 39.0 | 41.8 |
| Oracle | | 37.2 | 48.2 | 52.7 | 35.2 | 52.2 | 48.5 | 35.3 | 38.8 | 43.5 |
| Source Only | FCOS | 30.2 | 27.4 | 34.2 | 6.8 | 18.0 | 2.7 | 14.4 | 29.3 | 20.4 |
| EPM [10] | | 41.9 | 38.7 | 56.7 | 22.6 | 41.5 | 26.8 | 24.6 | 35.5 | 36.0 |
| EPM* [10] | | 44.9 | 44.4 | 60.6 | 26.5 | 45.5 | 28.9 | 30.6 | 37.5 | 39.9 |
| Synergy [22] | | 45.1 | 47.4 | 59.4 | 24.5 | 50.0 | 25.7 | 26.0 | **38.7** | 39.6 |
| Ours (Category) | | **47.2** | 44.7 | 61.9 | **27.6** | **48.9** | 38.6 | 31.7 | 38.6 | 42.4 |
| Ours (Bbox-width) | | 46.7 | 45.3 | **62.6** | 25.9 | 47.7 | 39.6 | 32.3 | 38.5 | 42.3 |
| Ours (Category & Bbox-width) | | 46.4 | **45.8** | 62.0 | 26.7 | 48.5 | **45.3** | **34.9** | 37.9 | **43.4** |
| Oracle | | 49.6 | 47.5 | 67.2 | 31.3 | 52.2 | 42.1 | 32.9 | 41.7 | 45.6 |

Table 2: Results of Sim10k to Cityscapes. * denotes the results of our re-implementation. No class conditional alignments are used since only car classes are used.

| Method | Detector | $mAP^r_{0.5}$ |
|---|---|---|
| Source only | Faster-RCNN | 34.3 |
| DAF [34] | | 39.0 |
| MAF [7] | | 41.1 |
| MeGA-CDA [30] | | 44.8 |
| SW-DA [25] | | 42.3 |
| SC-DA [34] | | 43.0 |
| Oracle | | 69.7 |
| Source Only | FCOS | 40.4 |
| EPM [10] | | 49.0 |
| EPM* [10] | | 50.0 |
| Synergy[22] | | 51.8 |
| Ours (Bbox-width) | | 53.7 |
| Ours (Bbox-height) | | **53.9** |
| Oracle | | 72.7 |

compared to basleine EPM*. In addition, since the object category and object scale are complementary conditions, it can be seen that when the two conditioning are applied simultaneously, $mAP^r_{0.5}$ increases by 3.8 and is only 1.8 less than oracle performance. Considering that oracle with FCOS is slightly better than oracle with Faster-RCNN, ours shows better performance than other Faster-RCNN based methods.

**Learning from synthesis data** Table 2 shows the results on Sim10k to Cityscape benchmark set. Since dataset has only one category, Model considered Bbox-scale only. We conduct experiments on conditioning width and height of bounding boxes. Both binning methods result significant improvements compared to the baseline [10]. We obtained the best results among models. Improvement of performance was similar whether it is binned with Height or Width. But binning with width recorded slightly better than the other.

### 4.4. Ablation Study

Our discriminator takes the outer product of model prediction and the corresponding feature as the input feature map, which results in both higher dimensional input and discriminator feature space. We hypothesize that this alone could improve the overall performance, hence conduct ablation studies on the precise impact of conditioning on the model prediction as shown in Tab. 4. In Tab. 4, when $q = 1$, $\alpha$ is always 1 during training, and as a result, the features are conditioned on only uniform constant vector, *i.e.,* $\frac{1}{N}\mathbb{1}$ where $N$ refers to the number of classes/bins and $\mathbb{1}$ represents a length-$N$ vector with all elements equal to 1. It is the same as simply increasing the dimension of the existing features without conditioning on either category-aware or object scale-aware. We observe overall boosts in performance due to larger input and model size, but further gains can consistently be obtained when model prediction is used, as it effectively guides the input feature vectors to the appropriate regions in the enlarged discriminator latent space. Thus, we empirically confirm that our proposed method utilizes conditional information in a highly efficient manner, outperforming simple baselines with equivalent capacity.

## 5. Conclusion

We present a novel unsupervised domain adaptation framework with object category-aware and scale-aware feature alignment that builds upon the single stage object de-

Table 3: Results of Sim10k to Cityscapes. * denotes the results of our re-implementation. No class conditional alignments are used since only car classes are used.

| Dataset | Method | mAP | $mAP^r_{0.5}$ | $mAP^r_{0.75}$ | $mAP^r_S$ | $mAP^r_M$ | $mAP^r_L$ |
|---|---|---|---|---|---|---|---|
| Cityscapes → Foggy Cityscapes | Source only | 10.8 | 20.4 | 10.0 | 1.2 | 11.3 | 28.4 |
| | Ours (Category) | 23.5 | 42.4 | 21.6 | 3.7 | 22.2 | 43.5 |
| | Ours (Bbox-width) | 24.0 | 42.3 | 23.0 | 3.9 | 21.9 | 43.9 |
| | Ours (Category & Bbox-width) | 24.0 | 43.4 | 23.0 | 4.0 | 21.7 | 44.5 |
| | Oracle | 25.4 | 45.6 | 23.4 | 5.5 | 24.2 | 44.8 |
| Sim10k → Cityscapes | Source only | 12.8 | 40.4 | 2.9 | 4.3 | 13.1 | 27.0 |
| | Ours (Bbox-width) | 31.4 | 53.7 | 31.3 | 8.7 | 29.0 | 60.9 |
| | Ours (Bbox-width & Bbox-height) | 29.9 | 53.0 | 28.7 | 8.1 | 32.0 | 59.2 |
| | Oracle | 48.7 | 72.7 | 51.3 | 18.9 | 55.5 | 81.4 |

Table 4: Comparison of $mAP^r_{0.5}$ according to the change in $q$ value, which is in $\alpha = max(1 - \frac{iter_{cur}}{Iter}, q)$ of Eq.2

| Dataset | Conditioning | $q = 1$ | $q = 0$ (ours) |
|---|---|---|---|
| Cityscapes → Foggy Cityscapes | Category | 41.8 | 42.4 |
| | Bbox-width | 41.2 | 42.3 |
| Sim10k → Cityscapes | Bbox-width | 52.6 | 53.8 |
| | Bbox-height | 52.6 | 53.9 |

tector, FCOS. In our framework, we utilize a strong domain classifier trained on the outer product of the feature vector and class/scale prediction vectors for adversarial feature alignment. We empirically show the effectiveness of our proposed method through thorough evaluations on two widely used domain adaptation benchmarks, CS2Foggy and Sim10k2CS. We hope our work fertilizes active research in the field of unsupervised domain adaptation.

## References


[1] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019. 3

[2] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors, 2020. 1, 2

[3] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild, 2018. 1, 2, 7

[4] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6

[5] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 3

[6] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 3

[7] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6668–6677, 2019. 7

[8] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721. PMLR, 2019. 3

[9] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *arXiv preprint arXiv:1906.12340*, 2019. 3

[10] 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 *European Conference on Computer Vision*, pages 733–748. Springer, 2020. 4, 6, 7

[11] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection, 2019. 2

[12] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Rda: Robust domain adaptation via fourier adversarial attacking. *arXiv preprint arXiv:2106.02874*, 2021. 1, 3

[13] Xiang Jiang, Qicheng Lao, Stan Matwin, and Mohammad Havaei. Implicit class-conditioned domain alignment for unsupervised domain adaptation, 2020. 3

[14] Deng Jinhong, Li Wen, Chen Yuhua, and Duan Lixin. Unbiased mean teacher for cross-domain object detection. 2021. 1

[15] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983*, 2016. 6

[16] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain

adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 480–490, 2019. 3

[17] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection, 2019. 1, 7

[18] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 3

[19] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional adversarial domain adaptation, 2018. 2

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 3

[21] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014. 2

[22] Muhammad Akhtar Munir, Muhammad Haris Khan, M. Saquib Sarfraz, and Mohsen Ali. Synergizing between self-training and adversarial learning for domain adaptive object detection. *CoRR*, abs/2110.00249, 2021. 7

[23] Rindra Ramamonjison, Amin Banitalebi-Dehkordi, Xinyu Kang, Xiaolong Bai, and Yong Zhang. Simrod: A simple adaptation method for robust object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3570–3579, 2021. 3

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 2

[25] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection, 2019. 2, 7

[26] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018. 6

[27] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 3

[28] Francisco Utrera, Evan Kravitz, N Benjamin Erichson, Rajiv Khanna, and Michael W Mahoney. Adversarially-trained deep nets transfer better: Illustration on image classification. In *International Conference on Learning Representations*, 2020. 3

[29] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019. 3

[30] 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, 2021. 3, 7

[31] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 3

[32] Zhen Zhao, Yuhong Guo, Haifeng Shen, and Jieping Ye. Adaptive object detection with dual multi-label prediction, 2020. 2

[33] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020. 2

[34] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 687–696, 2019. 7