Frequency Spectrum Augmentation Consistency for Domain Adaptive Object Detection

Rui Liu  
Tianjin University  
ruiliu@tju.edu.cn  

Yahong Han  
Tianjin University  
yahong@tju.edu.cn  

Yaowei Wang  
PengCheng Laboratory  
wangyw@pcl.ac.cn  

Qi Tian  
Huawei Cloud & AI  
tian.qil@huawei.com  

Abstract

Domain adaptive object detection (DAOD) aims to improve the generalization ability of detectors when the training and test data are from different domains. Considering the significant domain gap, some typical methods, e.g., CycleGAN-based methods, adopt the intermediate domain to bridge the source and target domains progressively. However, the CycleGAN-based intermediate domain lacks the pix- or instance-level supervision for object detection, which leads to semantic differences. To address this problem, in this paper, we introduce a Frequency Spectrum Augmentation Consistency (FSAC) framework with four different low-frequency filter operations. In this way, we can obtain a series of augmented data as the intermediate domain. Concretely, we propose a two-stage optimization framework. In the first stage, we utilize all the original and augmented source data to train an object detector. In the second stage, augmented source and target data with pseudo labels are adopted to perform the self-training for prediction consistency. And a teacher model optimized using Mean Teacher is used to further revise the pseudo labels. In the experiment, we evaluate our method on the single- and compound-target DAOD separately, which demonstrate the effectiveness of our method. And our code is available at https://github.com/ruiliu-code/FSAC.

1. Introduction

There has been impressive advances in object detection [11, 13, 14, 26, 27, 29, 33, 34], with the development of detection frameworks and related datasets [7, 9, 21, 37, 59]. However, these CNN-based methods often suffer from performance degradation when the distribution of training and test data is different. To mitigate this problem called domain shift, domain adaptive object detection (DAOD) [6] has been proposed, which aims at bridging the domain gap between the training/source and test/target domain, e.g., different weather conditions and different painting styles. We use the daytime-sunny images from BDD100K as the original data, and foggy, dusk-rainy, night-rainy as the target, respectively.

To solve the issue in DAOD, many previous works [6, 16, 17, 36, 49, 52, 53, 61, 64] attempt to reduce the domain gap based on the features. And most of them adopt the adversarial training mechanism to align the different level feature distributions. In consideration of the significant gap between these domains, some methods [3, 8, 18] employ the intermediate domain as a transition to fill the gap, which is generated by CycleGAN [63]. However, we can find the semantic differences between the original and synthetic images in Fig. 1, e.g., there are false car logos, noise, and traffic lights in the right part of the three rows, respectively. Since it has only image-level semantic supervision in the CycleGAN but lacks the pix-level or instance-level supervision. And adding these constraints will be complicated. In addition, recent research [5] shows that the standard GAN fails to learn the high-frequency information of real data.

Therefore, to address the above problem, we look for a new way to generate new images associated with the source/target domain. Motivated by the frequency spectrum analysis in the traditional image processing [12], the phase angle of an image can be used to reconstruct the shape features as shown in Fig. 2, which are necessary for object detection. And there is the intensity information in the...
amplitude spectrum that we care about. As shown in Fig. 3, we count the amplitude components of the images from different domains. We can find the images from different domains have different amplitude components. And there is also a significant difference among the inter-domain images especially in the low frequency, but negligible among intra-domain images and the high-frequency parts. From Fig. 3, there exists domain-specific information in the low-frequency amplitude spectrum, which is basically consistent with the previous works [19, 55, 56].

Based on the above works, we explore the frequency spectrum augmentation with four different low-frequency filter operations without additional parameters (see Fig. 4). Apparently, there is domain-specific information in the augmented images. And how to use these domain-related information in the augmented data for DAOD is important. On the other hand, considering the unlabeled data in the target domain, we need to make full use of the information in the target domain. Therefore, we decide to adopt the prediction consistency on original and augmented data inspired by the semi-supervised learning [32, 40, 41]. The purpose of prediction consistency is to extract the features, which are independent of the different data augmentation. And the proposed frequency spectrum augmentation is associated with domain-specific information. Thus, augmentation and prediction consistency is employed for DAOD.

To be specific, there are two training stages in our framework. In the first stage, we use all the original and augmented source data to train an object detector (e.g., Faster R-CNN). Existing studies [46, 58] have shown that frequency-based augmentation can improve the generalization ability of the model. Until the model converges, we perform the second stage. We inference the original source and target data on the pre-trained model of the first stage. The resulting pseudo labels are filtered by Non-Maximum Suppression (NMS) and confidence threshold. For the prediction consistency, we adopt both the augmented source and target data with pseudo labels to perform the self-training. Significantly, to further revise the pseudo labels for target data, we use the Mean Teacher [43] paradigm on a teacher model. And the parameters of the teacher model are updated by exponential moving average for each epoch of the student model. Experimental results show that the teacher model can obtain more accurate pseudo labels. Then, they are employed in the self-training for target data. The two-stage training could promote our model to learn the domain-independent features, which is beneficial for improving the generalization ability for DAOD.

In the experiment, we first evaluate our method on the single- and compound-target [30, 50] cases. Next, we compare the performance changes under different amounts of augmented data in the first training stage. The significant performance gain over baselines illustrates the effectiveness of our augmentation method. And our code will be released.

Our contributions can be summarized as follows: (1) We introduce a novel frequency spectrum augmentation for object detection. And the experimental results show the pre-trained model with good cross-domain detection performance can be obtained only using this augmentation. (2) Different from traditional adversarial training, we present a domain-related augmentation consistency and self-training framework for DAOD, which alleviates the domain shift with simple and effective strategies. (3) Our new model achieves the state-of-the-art performances on the single- and compound-target cases, which demonstrates the effectiveness of the proposed method.

2. Related Work

Domain Adaptive Object Detection. Most previous methods [6, 17, 35, 36, 45, 52, 57, 60, 61] adopt the GRL [10]
for adversarial training to align the global or region level distributions of the source and target domains. Concretely, Chen et al. [6] utilized the image- and instance-level feature distributions alignment to reduce the domain discrepancy. Saito et al. [36] proposed to align global features weakly and local features strongly. Some methods [49, 53] explored the instance-level feature alignment to improve the ability of DAOD. Considering the significant domain gap, some works [3, 18] introduced an intermediate domain for progressive feature alignment. Deng et al. [8] adopted the CycleGAN and Mean Teacher model to eliminate the model bias. For the feature disentanglement, there are some methods [25, 47, 50] extracting the domain-invariant features for domain shift. Although these methods have been demonstrated to be effective, we proposed a new observation with domain specific augmentation consistency for DAOD.

**Fourier-based Domain Adaptation.** Yang et al. [56] proposed the Fourier domain adaptation by swapping the low-frequency spectrum between the source and target domain. And the phase component of the Fourier Transform contains the main semantic content [55]. Inspired by these works, some methods [28, 54] adopted the spectrum swapping for domain generalization. And Fourier-based data augmentation can avoid overfitting to low-level statistics. Domain randomization [19] is also used for domain generalization by keeping domain-invariant frequency components and randomizing domain-variant ones. Huang et al. [20] used the Fourier adversarial attacking for robust domain adaptation. Different from the above methods, we extend the frequency spectrum augmentation in four forms and combine the consistency-based self-training for DAOD.

**Semi-Supervised Object Detection.** Semi-supervised learning [32, 40, 41] aims to improve the predictive ability by unlabeled data. Sohn et al. [41] adopted the pseudo labels of unlabeled data for self-training on object detection. Jeong et al. [22] applied a modified interpolation regularization for object detection. Chen et al. [3] proposed a point annotated setting as weakly semi-supervised object detection using the normalized object points. Tang et al. [42] introduced the light-weighted data ensemble for the teacher model to generate the reliable pseudo labels. Wang et al. [48] considered the difficulty levels of unlabeled images and used them in different phases to address the noisy labels. Instant-Teaching [62] used a co-rectify scheme with weak-strong data augmentations to improve the quality of pseudo annotations. Our method is partially motivated by semi-supervised learning.

### 3. Frequency Spectrum Augmentation

For an RGB image $x \in \mathbb{R}^{H \times W \times 3}$, the Fourier transformation of $i$-th channel $F(x_i)$ is formulated as:

$$F(x_i)(u, v) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x_i(h, w) e^{-j2\pi(uh/H + vw/W)},$$

and the inverse Fourier transformation $F^{-1}$ is defined as:

$$x_i(h, w) = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} F(x_i)(u, v) e^{j2\pi(uh/H + vw/W)},$$

and both of them can be implemented by the FFT algorithm [31] in practice. The frequency (amplitude) and phase spectrum can be calculated respectively as:

$$|F(x_i)(u, v)| = |R^2(x_i)(u, v) + I^2(x_i)(u, v)|^{1/2},$$

$$\phi(x_i)(u, v) = \arctan \left( \frac{I(x_i)(u, v)}{R(x_i)(u, v)} \right),$$

where $R(x_i)(u, v)$ and $I(x_i)(u, v)$ denote the real and imaginary part of $F(x_i)$. Therefore, we could obtain the Fourier transformation of the RGB image $x$ by channel-wise concat.

As shown in Fig. 3, we can find the frequency spectrum distributions of images from source and target domain respectively may be different. And the images from the same domain have a similar frequency spectrum. We also calculate the average frequency spectrum distribution of each domain and reach a similar conclusion by comparison. On the other hand, the energy of an image is mainly concentrated in the low-frequency part, which is discovered from Fig. 3. As described in [12, 56], the low-frequency part means the slowly changing part of the image, which is related to domain-specific information, such as color and style. And the high-frequency part means the sharply changing part, e.g., texture and edge. Therefore, we try to perform some operations on the low-frequency part for domain-related augmentation.

We introduce four kinds of frequency spectrum augmentation in this section. We first shift the discrete Fourier transform to center the frequency spectrum with low-frequency origin. Then, we define an ideal high-pass filter as:

$$B(u, v) = \begin{cases} 
1 & \text{if } |u| \geq \alpha H \text{ and } |v| \geq \alpha W, \\
0 & \text{else}, 
\end{cases}$$

where $(u, v)$ means the coordinates in the centered frequency spectrum $(H \times W)$ and $\alpha$ means the “cut off” ratio.
Similarly, we also define an ideal low-pass filter as:

$$D(u, v) = \begin{cases} 1 & \text{if } |u| < \beta H \text{ and } |v| < \beta W, \\ 0 & \text{else}, \end{cases}$$

where $\beta$ means the “cut off” ratio. It is noted that the complementary high- and low-pass filters are employed in our method, i.e., $\alpha = \beta$. Here, we use the source data $x_s$ for augmentation as an example. The first augmentation is called “Mix”, which is implemented as:

$$|\mathcal{F}(x^s)|_A = B|\mathcal{F}(x^s)| + \gamma D|\mathcal{F}(x^s)| + (1 - \gamma)D|\mathcal{F}(x^s)|,$$

where the $(u, v)$ is omitted for simplicity, $\gamma \in (0, 1)$ is random, $x^t$ is the target image, and $|\mathcal{F}(x^s)|_A$ is the frequency spectrum of source image after augmentation. And the second augmentation “Replace” is performed as:

$$|\mathcal{F}(x^s)|_A = B|\mathcal{F}(x^s)| + D|\mathcal{F}(x^t)|,$$

which is considered as replacing the low-frequency spectrum of the source image with target image. This operation is similar to [56]. However, we perform the random choice of target image instead of swapping in pairs [56]. To further enrich the frequency spectrum information, we propose the third augmentation “Extend” as:

$$|\mathcal{F}(x^s)|_A = B|\mathcal{F}(x^s)| + \gamma D|\mathcal{F}(x^t_1)| + (1 - \gamma)D|\mathcal{F}(x^t_2)|,$$

where $\gamma \in (0, 1)$ is random, $x^t_1$ and $x^t_2$ are two images respectively from target domain. In this way, we can extend the frequency spectrum information in the target domain. In the above three operations, we perform the same frequency spectrum augmentation on the corresponding RGB channels. Then, the fourth operation “Disorder” is channel-wise disorder “Replace”, which is calculated as:

$$|\mathcal{F}(x^s_i)|_A = B|\mathcal{F}(x^s_i)| + D|\mathcal{F}(x^t_j)|,$$

where $i$ and $j$ are different channels of source and target images, i.e. $i \neq j$. It is considered as channel-wise frequency spectrum augmentation.

After the frequency spectrum augmentation, we obtain the intermediate domain data. The size of the target images will be adjusted to be the same as the source by interpolation when we perform the augmentation on source data, and vice versa. In the following chapters, we also use similar operations for the target data.

4. Method

For DAOD, we have the access to an image $x^s$ of the source domain with labels $y^s$ and bounding boxes $b^s$. Meanwhile, we could also access an unlabeled image $x^t$ of target domain. Our goal is to improve the generalization ability in target domain. And our method contains two stages of training. In the first stage, we train a model with original and augmented source data. And in the second stage, we propose the prediction consistency based on the pre-trained model for original and augmented data from source and target domain.

4.1. Augmented Training

The Fig. 5 illustrates the details of augmented training, which is based on the Faster R-CNN [34] as the previous works. In view of the domain-specific information in augmented data, we train the model with both original and augmented source data as the first stage, which are respectively denoted as $x^s$ and $A(x^s)$:

$$L_1 = L_{AT} = L_{det}(x^s, y^s, b^s) + L_{det}(A(x^s), y^s, b^s),$$

where the augmentation $A(\cdot)$ here means the random choice from the four operations, $y^s$ is the corresponding set of object classes in $x^s$, and $b^s$ is the set of bounding boxes. Faster R-CNN is a detector mainly containing a region proposal network (RPN) and a region-of-interest (RoI) based classifier. Both of the RPN and RoI classifiers consist of two modules, i.e., classification and bounding box regression. The loss of the first stage training is composed of them, which is formulated as:

$$L_{det} = L_{RPN} + L_{RoI}. \tag{12}$$
When the model converges through the above training, thanks to the frequency spectrum augmentation, we obtain a pre-trained model with good generalization for the target domain. Obviously, the number of augmented source data may have an impact on performance. And we will experiment on it in Section 5.1 and 5.2.

4.2. Prediction Consistency

Self-Training. Given the pre-trained model with parameters \( \theta \), we perform the inference to generate the pseudo labels. And we adopt a typical method of the pseudo-labeling in [24,41] as follows:

\[
q(x) = \left[ \text{argmax}(p(x; \theta)) \right]_{\text{ONE-HOT}},
\]

\[
w(x) = 1 \text{ if } \max(p(x; \theta)) \geq \tau, \tag{13}
\]

where \( p \) is the prediction score of the model, \( q \) represents the prediction label, and \( w \) means that the sample \( x \) has a contribution to the loss when its confidence is greater than the threshold \( \tau \). And we need to choose the appropriate pseudo labels for further training. Therefore, we set a high threshold \( \tau \) for the confidence of pseudo labels. And in the forward pass of the pre-trained model, non-maximum suppression (NMS) is also used to remove redundant information as overlapping bounding boxes.

To make full use of the information in the data, we infer the original source and target data on the pre-trained model. Finally, we get a series of pseudo labels for source data \( x^s \) and target data \( x^t \), which are denoted as \((y^{*s}, b^{*s})\) and \((y^{*t}, b^{*t})\), respectively. Then, the self-training is formulated as:

\[
\mathcal{L}_{ST} = \mathcal{L}_{det}(A(x^s), y^{*s}, b^{*s}) + \mathcal{L}_{det}(A(x^t), y^{*t}, b^{*t}) \tag{14}
\]

This is the prediction consistency loss for original and augmented data. In this way, the model will improve generalization ability and tend to extract domain-independent information. Since it is mainly to ensure the semantics are unchanged after the domain-related augmentation, the pseudo labels of source data are utilized in this loss instead of ground truth. Due to the randomness of the augmentation, we pseudo-label the original data instead of augmented data.

Teacher Guiding. Considering that our goal is to improve performance on the target domain, we introduce a teacher model for target pseudo labels. From [43], averaging model parameters over training steps is more conducive to producing an accurate model than training directly. The teacher model is optimized by using the exponential moving average weights of the student model [23]. And the above pre-trained model \( \theta \) can be seen as the student model. Therefore, the parameters \( \theta'_t \) of the teacher model at training step \( t \) are defined as:

\[
\theta'_t = \delta \theta'_{t-1} + (1 - \delta) \theta_t \tag{15}
\]

where \( \delta \) denotes a smoothing coefficient hyperparameter and \( \theta_t \) is the student model at training step \( t \). Then, we obtain the more accurate pseudo labels of target data as \((y^{t+}, b^{t+})\). And the teacher model will guide the student model, which is formulated as:

\[
\mathcal{L}_T = \mathcal{L}_{det}(x^t, y^{t+}, b^{t+}) \tag{16}
\]

Different from the previous methods [1,51], our main purpose is to improve the performance on the target domain. And we use the prediction consistency for self-training instead of iterative training for the student and teacher model. To make a stable training, we still include the first-stage loss in the second-stage loss as:

\[
\mathcal{L}_2 = \mathcal{L}_{AT} + \lambda_1 \mathcal{L}_{ST} + \lambda_2 \mathcal{L}_T \tag{17}
\]

5. Experiment

In the experiment, our method is evaluated on single- and compound-target DAOD, respectively. For the single-target case, we evaluate our approach on four domain-shift scenes, i.e., PASCAL [9] \( \rightarrow \) Watercolor [21], Cityscapes [7] \( \rightarrow \) Foggy Cityscapes [37], Daytime-sunny \( \rightarrow \) Dusk-rainy, and Daytime-sunny \( \rightarrow \) Night-rainy [50,59]. For the compound case [30,50], we follow the previous work using Daytime-sunny as the source domain, and the domain-agnostic combination of Dusk-rainy and Night-rainy as the compound-target domain.

Datasets. For PASCAL \( \rightarrow \) Watercolor, we use the PASCAL VOC dataset as the source domain. It is a real-world dataset with 20 categories of objects and bounding box annotations. Following [36], we utilize PASCAL VOC 2007 and 2012 training and validation splits for experiments. Watercolor contains artistic images with 6 categories. For Cityscapes \( \rightarrow \) Foggy Cityscapes, there are 2,975 images for training and 500 images for validation in both of them. For a fair comparison, our settings and the splits of the training/test set are the same as the work [36]. For Daytime-sunny \( \rightarrow \) Dusk-rainy, Daytime-sunny \( \rightarrow \) Night-rainy, and the compound-target case, there are 27,708 daytime-sunny

| Method         | bike  | bird  | car   | cat   | dog   | person | mAP  |
|----------------|-------|-------|-------|-------|-------|--------|------|
| Source Only    | 68.8  | 46.8  | 37.2  | 32.7  | 21.3  | 60.7   | 44.6 |
| DAF [6]        | 75.2  | 40.6  | 48.0  | 31.5  | 20.6  | 60.0   | 46.0 |
| SW [56]        | 82.3  | 55.9  | 46.5  | 32.7  | 35.5  | 66.7   | 53.3 |
| MAF [16]       | 73.4  | 55.7  | 46.4  | 36.8  | 28.9  | 60.8   | 50.3 |
| ATF [17]       | 78.8  | 59.9  | 47.9  | 41.0  | 34.8  | 66.9   | 54.9 |
| SCL [38]       | 82.2  | 55.1  | 51.8  | 39.6  | 38.4  | 64.0   | 55.2 |
| DBGL [2]       | 83.1  | 49.3  | 50.6  | 39.8  | 38.7  | 61.3   | 53.8 |
| VDD [50]       | 90.0  | 56.6  | 49.2  | 39.5  | 38.8  | 65.3   | 56.6 |
| UMT [8]        | 88.2  | 55.3  | 51.7  | 39.8  | 43.6  | 69.9   | 58.1 |
| AT (first stage)| 84.2  | 53.1  | 45.6  | 33.8  | 34.9  | 64.3   | 52.6 |
| FSAC (ours)    | 89.8  | 58.7  | 53.3  | 42.4  | 39.8  | 68.7   | 58.8 |

Table 1. Results (%) on adaptation from PASCAL to Watercolor.
Implementation Details. We employ the two-stage training in our method. In the first stage, we perform the augmented training with a learning rate of 0.001 for 100K iterations, then with the learning rate of 0.0001 for 80K more iterations. The “cut off” ratios of ideal high- and low-pass filter is set as $\alpha = \beta = 0.09$, which are selected by the experiments. Then, we perform the prediction consistency and Mean Teacher of the second stage. The pre-trained model of the first stage can be regarded as the student model. We train the student model with the learning rate 0.0001 for 20K more iterations. The confidence threshold $\tau$ is set as 0.8 and the teacher EMA coefficient is set as 0.98. And we set the weights of Eq. (17) as $\lambda_1 = 0.3$ and $\lambda_2 = 0.1$. For all datasets, we report the average precision (AP, %) and mean average precision (mAP, %) with a threshold of 0.5.

5.1. Result Analysis

Results on Watercolor. The results are presented in Table 1 and the third row in Fig. 7. Here, we use ResNet101 [15] as the backbone. Our method obtains 58.8% mAP detection accuracy that surpasses the previous works. Compared with the adversarial training method VDD [50], our method has a higher accuracy with 2.2% mAP. UMT [8] has achieved the best performance recently. Particularly, our method improves the average precision of “car” and “cat” by 1.6% and 2.6% against UMT. And it finds the pre-trained model of the first stage has a comparable performance with SW [36]. It demonstrates the effectiveness of our frequency spectrum augmentation, which contributes to the accuracy of the generated pseudo labels.

Results on Foggy Cityscapes. Table 2 and the first row of Fig. 7 show the results. Here, VGG16 [39] is used as the backbone. The current best state-of-art result on this dataset is from the recent work MeGA [45]. And our method has a 1.1% performance gain on it. Compared with MeGA, we improve the average precision of “car” by about 8%, which is the most category of objects. It is noted that the model from our first-stage training achieve a better performance than some well-designed work, e.g., SW [36], MAF [16], CT [60] and PDA [18]. Considering that PDA adopts the CycleGAN based intermediate domain, it further illustrates the effectiveness of frequency spectrum augmentation. Besides, UMT proposes a cross-domain distillation method for an unbiased Mean Teacher model. We can see our method outperforms UMT by 1.2%, which also demonstrates the effectiveness of prediction consistency.

Results on Dusk-rainy. Table 3 and the second row of Fig. 7 shows the results. Here, VGG16 [39] is used as the backbone. The current best state-of-art result on this dataset is from the recent work MeGA [45]. And our method has a 1.1% performance gain on it. Compared with MeGA, we improve the average precision of “car” by about 8%, which is the most category of objects. It is noted that the model from our first-stage training achieve a better performance than some well-designed work, e.g., SW [36], MAF [16], CT [60] and PDA [18]. Considering that PDA adopts the CycleGAN based intermediate domain, it further illustrates the effectiveness of frequency spectrum augmentation. Besides, UMT proposes a cross-domain distillation method for an unbiased Mean Teacher model. We can see our method outperforms UMT by 1.2%, which also demonstrates the effectiveness of prediction consistency.

Results on Night-rainy. Table 4 and the third row of Fig. 7 shows the results. Here, VGG16 [39] is used as the backbone. The current best state-of-art result on this dataset is from the recent work MeGA [45]. And our method has a 1.1% performance gain on it. Compared with MeGA, we improve the average precision of “car” by about 8%, which is the most category of objects. It is noted that the model from our first-stage training achieve a better performance than some well-designed work, e.g., SW [36], MAF [16], CT [60] and PDA [18]. Considering that PDA adopts the CycleGAN based intermediate domain, it further illustrates the effectiveness of frequency spectrum augmentation. Besides, UMT proposes a cross-domain distillation method for an unbiased Mean Teacher model. We can see our method outperforms UMT by 1.2%, which also demonstrates the effectiveness of prediction consistency.

| Method   | prsn rider car truck bus train mcycl bcycl | mAP   |
|----------|-----------------------------------------|-------|
| Source Only | 24.7 31.9 33.1 11.0 26.4 9.2 18.0 27.9 22.8 |
| DAF [6]  | 23.0 31.0 40.5 22.1 35.3 20.2 20.0 27.1 27.6 |
| TVT [17] | 25.8 39.3 42.4 24.9 40.4 23.1 25.9 30.4 31.5 |
| SC-DA [64] | 33.5 38.0 48.5 26.5 39.0 23.3 28.0 33.6 33.8 |
| SW [36]  | 32.9 42.3 43.5 24.8 36.2 32.6 30.0 35.3 34.3 |
| MAF [18] | 28.2 39.5 43.9 23.8 39.9 33.3 29.2 33.9 34.0 |
| CT [69]  | 32.7 44.4 50.1 21.7 45.6 25.4 30.1 36.8 35.9 |
| SCL [35] | 31.6 44.0 44.8 30.4 41.8 40.7 33.6 36.2 37.9 |
| AIF [17] | 34.6 47.0 50.0 23.7 43.3 38.7 33.4 38.8 38.7 |
| MCAR [51] | 32.5 42.2 43.4 31.9 44.1 43.4 37.4 36.6 38.5 |
| PDA [18] | 36.0 45.5 54.4 24.3 44.1 25.8 29.1 35.9 36.9 |
| HTCN [3]  | 33.2 47.5 47.9 31.6 47.4 40.9 32.3 37.1 39.8 |
| ICR [1]   | 32.9 43.8 49.2 27.2 45.1 36.4 30.3 34.6 37.4 |
| VDD [50]  | 33.4 44.0 51.7 33.9 52.0 34.7 34.2 36.8 40.0 |
| DS5 [47]  | 42.9 51.2 53.6 33.6 49.2 18.9 36.2 41.8 40.9 |
| KINet [16] | 46.4 43.2 60.6 25.8 41.2 40.4 30.7 38.8 40.9 |
| UMT [4]   | 33.0 46.7 48.6 34.1 56.5 46.8 30.4 37.3 41.7 |
| MeGa [15] | 31.7 49.0 52.4 25.4 49.2 46.9 34.5 39.0 41.8 |
| AT (first stage) | 32.6 43.7 51.2 27.0 36.7 32.8 33.2 35.9 36.6 |
| FSAC (ours) | 39.3 46.2 60.8 34.5 46.3 40.9 37.7 37.5 42.9 |

Table 2. Results (%) on adaptation from Cityscapes to Foggy Cityscapes. ‘prsn’, ‘mcycl’, and ‘bcycl’ separately denote ‘person’, ‘motorcycle’, and ‘bicycle’ category.

images, 3,501 dusk-rainy images, and 2,494 night-rainy images with 7 common categories, which are selected from the BDD100K. In these three cases, we follow the settings as the work [50].
Results on Daytime-sunny → Night-rainy. We report the results of Daytime-sunny → Night-rainy in Table 4. From the previous works [50], we can find Night-rainy is a difficult domain adaption case. And the performance of the first-stage model is comparable to some adversarial training methods, which show the effectiveness of augmentation training. And the precision of the two main categories “car” and “person” is improved. Our method obtains a performance gain of 0.8%. This further shows our method is effective.

Results on the compound case. The compared results are shown in Table 5. The ResNet101 is used as the backbone. Here, we perform the augmentation training by randomly selecting compound target domain data. It seems to be a challenge for our method. However, our method achieves the state-of-the-art performance with 1.1% mAP gain. This shows our method can solve the compound case. And the prediction consistency can extract the domain-independent features. Besides, the performance of the first-stage model by frequency spectrum augmentation is second only to VDD [50], which further demonstrates the effectiveness. And we also show the experimental results in Table 6 and 7, which are evaluated on dusk-rainy and night-rainy, respectively.

5.2. Ablation Analysis

We perform an ablation analysis based on the single-target case. Considering the pipeline of our method, we decide to conduct the ablation study for two training stages separately. And the corresponding backbone networks are used for different datasets.

Different amounts of augmented data. In Fig. 6, we show the results of different amounts of augmented data after first-stage training. The abscissa indicates the amount of the augmented data by the multiple of the original data, i.e., “0” means “Source Only”, and “2” means using the augmented twice the amount of original data.

Figure 6. Results (mAP) for different amounts of augmented data.
Figure 7. Detection results of Foggy, Dusk-rainy, and Watercolor. The third row shows the results of augmented data by “Disorder”.

| Method | AT | ST | T | C → F | P → W |
|--------|----|----|----|-------|-------|
| FSAC   | ✓  | ✓  | ✓  | 36.3% | 53.1% |
| FSAC   | ✓  | ✓  | ✓  | 37.5% | 53.9% |
| FSAC   | ✓  | ✓  | ✓  | 38.9% | 55.8% |
| FSAC   | ✓  | ✓  | ✓  | 40.7% | 57.2% |
| FSAC   | ✓  | ✓  | ✓  | 42.9% | 58.8% |

Table 8. Ablation analysis of second stage training in our method. Here, we use mAP as the metric. “AT” and “ST” denote we use augmented training and self-training to optimize our model, respectively. “T” indicates the mean teacher guiding. “C → F” denotes the adaptation from Cityscapes to Foggy Cityscapes and adopts VGG16 as the backbone. “P → W” denotes the adaptation from VOC to Watercolor and uses ResNet101 as the backbone.

The weight of second-stage loss. In Table 9, we report the results of different weights for second-stage loss. We can find the weight of teacher guiding loss should be lower than self-training. Choosing the appropriate combination of weights can achieve balance on DAOD.

6. Conclusion

In this paper, we propose a novel method named Frequency Spectrum Augmentation Consistency (FSAC). We analyze the frequency spectrum of different domains and compare it with CycleGAN based intermediate domain data. Then, we explore four frequency spectrum operations for augmentation, i.e., “Mix”, “Replace”, “Extend” and “Disorder”. After the first-stage augmented training, we use the prediction consistency to extract domain-independent information for self-training. Meanwhile, Mean Teacher guiding is adopted for further improvement. In the experiment, our method is evaluated on the single- and compound-target case, respectively. The performance gain over baselines shows the effectiveness of our method.
References

[1] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11457–11466. Computer Vision Foundation / IEEE, 2019. 5, 6, 7

[2] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 2703–2712, October 2021. 5

[3] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8866–8875. Computer Vision Foundation / IEEE, 2020. 1, 3, 6

[4] Liangyu Chen, Tong Yang, Xiangyu Zhang, Wei Zhang, and Jian Sun. Points as queries: Weakly semi-supervised object detection by points. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8823–8832. Computer Vision Foundation / IEEE, 2021. 3

[5] Yuanqi Chen, Ge Li, Cece Jin, Shan Liu, and Thomas H. Li. SSD-GAN: measuring the realness in the spatial and spectral domains. In *AAAI Conference on Artificial Intelligence*, pages 1105–1112. AAAI Press, 2021. 1

[6] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster R-CNN for object detection in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3339–3348. IEEE Computer Society, 2018. 1, 2, 3, 5, 6, 7

[7] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223. IEEE Computer Society, 2016. 1, 5

[8] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4091–4101. Computer Vision Foundation / IEEE, 2021. 1, 3, 5, 6

[9] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 1, 5

[10] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, volume 37, pages 1180–1189. JMLR.org, 2015. 1, 2

[11] Ross B. Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision*, pages 1440–1448. IEEE Computer Society, 2015. 1

[12] R.C. Gonzalez and R.E. Woods. *Digital Image Processing, EBook, Global Edition*. Pearson Education, 2018. 1, 3

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):386–397, 2020. 1

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE Computer Society, 2016. 6

[16] Zhenwei He and Lei Zhang. Multi-adversarial faster-rnn for unrestricted object detection. In *IEEE International Conference on Computer Vision*, pages 6667–6676. IEEE, 2019. 1, 5, 6

[17] Zhenwei He and Lei Zhang. Domain adaptive object detection via asymmetric tri-way faster-rnn. In *European Conference on Computer Vision*, volume 12369, pages 309–324. Springer, 2020. 1, 2, 5, 6

[18] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Manesh Kumar Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *IEEE Winter Conference on Applications of Computer Vision*, pages 738–746. IEEE, 2020. 1, 3, 6

[19] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. FSDR: frequency space domain randomization for domain generalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6891–6902. Computer Vision Foundation / IEEE, 2021. 2, 3

[20] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Rda: Robust domain adaptation via fourier adversarial attacking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8988–8999, October 2021. 3

[21] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5001–5009. IEEE Computer Society, 2018. 1, 5, 6

[22] Jisoo Jeong, Vikas Verma, Minsung Hyun, Juho Kannala, and Nojun Kwak. Interpolation-based semi-supervised learning for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11602–11611. Computer Vision Foundation / IEEE, 2021. 3

[23] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, OpenReview.net, 2017. 5

[24] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013. 5

[25] Chunghun Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhui Wang, and Jianfei Cai. Domain-invariant disentangling (WREPL) ICML 2013 Workshop : Challenges in Representation Learning (WREPL), 07 2013. 5

[26] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 936–944. IEEE Computer Society, 2017. 1
[51] Qizhe Xie, Minh-Thang Luong, Eduard H. Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10684–10695. Computer Vision Foundation / IEEE, 2020. 5

[52] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11721–11730. Computer Vision Foundation / IEEE, 2020. 1, 2

[53] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12352–12361. Computer Vision Foundation / IEEE, 2020. 1, 3

[54] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14383–14392. Computer Vision Foundation / IEEE, 2021. 3

[55] Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9008–9017. Computer Vision Foundation / IEEE, 2020. 2, 3

[56] Yanchao Yang and Stefano Soatto. FDA: fourier domain adaptation for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4084–4094. Computer Vision Foundation / IEEE, 2020. 2, 3, 4

[57] Xingxu Yao, Sicheng Zhao, Pengfei Xu, and Jufeng Yang. Multi-source domain adaptation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3273–3282, October 2021. 2

[58] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems*, pages 13255–13265, 2019. 2

[59] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving video database with scalable annotation tooling. *CoRR*, abs/1805.04687, 2018. 1, 5

[60] Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin. Collaborative training between region proposal localization and classification for domain adaptive object detection. In *European Conference on Computer Vision*, volume 12363, pages 86–102. Springer, 2020. 2, 6, 7

[61] Zhen Zhao, Yuhong Guo, Haifeng Shen, and Jieping Ye. Adaptive object detection with dual multi-label prediction. In *European Conference on Computer Vision*, volume 12373, pages 54–69. Springer, 2020. 1, 2, 6

[62] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4081–4090. Computer Vision Foundation / IEEE, 2021. 3

[63] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, pages 2242–2251. IEEE Computer Society, 2017. 1

[64] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 687–696. Computer Vision Foundation / IEEE, 2019. 1, 6