# FIT: Frequency-Based Image Translation for Domain Adaptive Object Detection

Siqi Zhang[1,2], Lu Zhang[1], Zhiyong Liu[1,2(✉)], and Hangtao Feng[1,2]

[1] State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{zhangsiqi2020,lu.zhang,zhiyong.liu,fenghangtao2018}@ia.ac.cn
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences,
Beijing 100190, China

**Abstract.** Domain adaptive object detection (DAOD) aims to adapt the detector from a labelled source domain to an unlabelled target domain. In recent years, DAOD has attracted massive attention since it can alleviate performance degradation due to the large shift of data distributions in the wild. To align distributions between domains, adversarial learning is widely used in existing DAOD methods. However, the decision boundary for the adversarial domain discriminator may be inaccurate, causing the model biased towards the source domain. To alleviate this bias, we propose a novel Frequency-based Image Translation (FIT) framework for DAOD. First, by keeping domain-invariant frequency components and swapping domain-specific ones, we conduct image translation to reduce domain shift at the input level. Second, hierarchical adversarial feature learning is utilized to further mitigate the domain gap at the feature level. Finally, we design a joint loss to train the entire network in an end-to-end manner without extra training to obtain translated images. Extensive experiments on three challenging DAOD benchmarks demonstrate the effectiveness of our method.

**Keywords:** Unsupervised Domain Adaptation · Object Detection · Frequency Domain · Image Translation · Adversarial Learning

## 1 Introduction

In recent years, object detectors [1–3] based on deep convolutional neural networks have demonstrated outstanding performance on a variety of datasets. However, existing object detection models still face serious challenges when deployed in practice such as autonomous driving and robotic manipulation, due to various changes in weather, illumination, object appearance, *etc.* These changes may lead to domain gaps between the training and testing data, which has been

This work was supported in part by the National Key Research and Development Plan of China under Grant 2020AAA0108902 and the Strategic Priority Research Program of Chinese Academy of Science under Grant XDB32050100.

M. Tanveer et al. (Eds.): ICONIP 2022, LNCS 13625, pp. 240–252, 2023.
https://doi.org/10.1007/978-3-031-30111-7_21

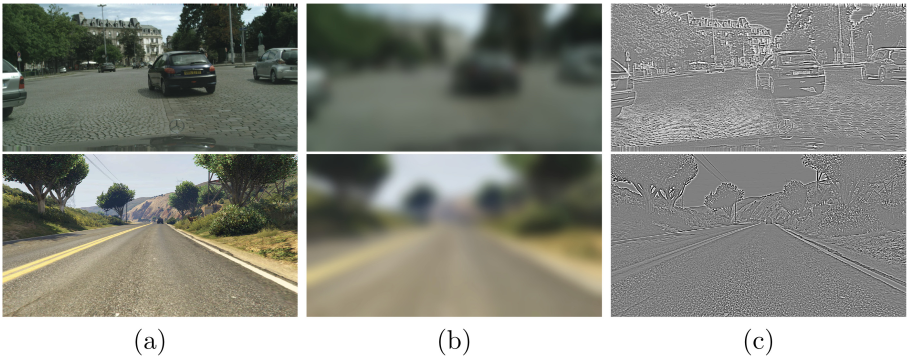


**Fig. 1.** Visualization of frequency decomposition of source image: (a), (b) and (c) show original image, low-pass and high-pass filtered image.

observed to cause dramatic drops in the performance of the trained detector [6]. Although we can annotate for each new data to mitigate the problem, it is costly and even infeasible because of the countless situations in the real world. Therefore, adaptive object detectors that can bridge the domain gap from the source to the target domain are highly desirable.

Domain adaptive object detection (DAOD), which trains with labelled source datasets and unlabelled target datasets, aims to tackle domain shift to get better performance on the visually distinct target domain. Many previous works [4,6,7,20,22–24] attempt to utilize adversarial feature learning [8] to align feature distributions to extract domain-invariant features. But the adversarial training process could be unstable [22,24], which makes the decision boundary for the adversarial domain discriminator inaccurate, causing the model biased towards the source domain. To alleviate this problem, some methods [5,9,16,21] utilize the image translation model GANs, like CycleGAN [10] to translate source images to target-like images or vice versa to further mitigate the domain gap and make the detector perform better on the target datasets. However, GANs for domain adaption object detection have two following limitations. First, GANs could fail to keep semantic consistency and tend to lose important structural characteristics [25]. Second, GANs-based methods need extra training to prepare translated images before training the adaptive detector, which is time-consuming.

To address the above limitations, we propose a novel Frequency-based Image Translation method to mitigate the input-level domain gap without extra time-consuming training. Inspired by digital signal processing theories [17], we exploit the frequency information to translate the image style and maintain semantic consistency. Intuitively, the low-frequency component largely captures domain-specific information, such as colours and illuminations [26], while the high-frequency component mainly obtains domain-invariant information, such as edges and shapes, which are important details of objects [19], as shown in Fig. 1. Motivated by this, we present the Frequency-based Image Translation (FIT)

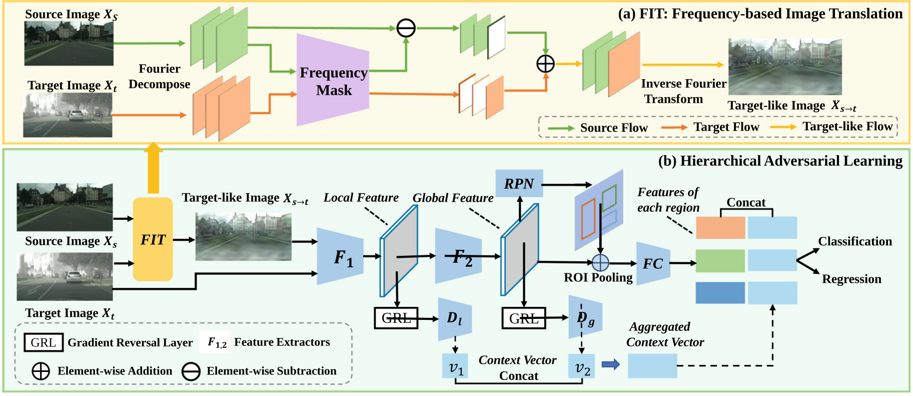


**Fig. 2.** Overview of the proposed framework. (a) illustrates Frequency-based Image Translation (FIT) module, where green arrows represent the flow of source data and orange arrows represent the flow of target data. The overall pipeline is illustrated in (b). The source $X_s$ and target $X_t$ images are fed into FIT to obtain target-like images $X_{s\to t}$, and $X_{s\to t}$ and $X_t$ are as the inputs for the object detector. We align the local and global feature by a local domain classifier $D_l$ and a global domain classifier $D_g$. $F_1$ and $F_2$ denote the different level feature extractors. The context vectors $v_{1,2}$ are extracted by the domain classifiers and concatenated with features of regions before the final fully connected layer. (Color figure online)

module, which decomposes the image into multiple frequency components, keeps domain-invariant frequency components unchanged and swaps domain-specific ones. Moreover, a novel module called Frequency Mask is designed to identify whether the frequency component is domain-specific in FIT. Then, hierarchical adversarial feature learning is utilized to further boost the performance. The entire network can be optimized in an end-to-end manner under the supervision of a joint loss function. The contributions of this work can be summarized as follows:

- A novel Frequency-based Image Translation (FIT) method is presented for DAOD, which leverages frequency information to mitigate the domain shift at the input level. To further boost the adaptation performance, we introduce hierarchical adversarial learning to align distributions at the feature level.
- Different from traditional GANs-based methods, the entire network can be trained in an end-to-end manner without extra time-consuming training, since the proposed frequency-based image translation is embedded as a module in the detection network.
- We conduct extensive experiments on three challenging DAOD benchmarks and our FIT achieves favorable performance under various domain-shift scenarios, demonstrating the effectiveness of the proposed method.

## 2 Proposed Method

### 2.1 Overview

**Problem Definition.** The domain adaptation [8] task typically considers two domains: the source domain $S$ and target domain $T$. Specifically, we have access to a labelled source dataset $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ and a target dataset $\mathcal{D}_t = \left\{x_j^t\right\}_{j=1}^{n_t}$ with no ground-truth annotations. Here, $x_i^s$ denotes $i^{th}$ source image, $y_i^s$ denotes the corresponding label and $n_s$ denotes the number of source images. Similarly, $x_j^t$ denotes $j^{th}$ target image and $n_t$ denotes the number of target images. The source and target domains with different data distributions share the same label space, and the goal of domain adaptive object detection is to train an detector with $\mathcal{D}_s$ and $\mathcal{D}_t$, which performs well on the target dataset $\mathcal{D}_t$. Following the mainstream domain adaptive object detection methods [4,6,7,20,22–24], the proposed method is based on the Faster RCNN [1] framework.

**Overall Framework.** The overall framework of the proposed method is shown in Fig. 2. We first transform source images $X_s$ to target-like images $X_{s\rightarrow t}$ via frequency-based image translation (FIT), as shown in Fig. 2(a). The key idea is to decompose the image into multiple frequency components and then feed them to the Frequency Mask to identify domain-specific frequency components. Then we replace the domain-specific components of the source image with the corresponding ones of the target image and get the target-like image $X_{s\rightarrow t}$ via the Inverse Fourier Transform. Afterwards, we put target-like images $X_{s\rightarrow t}$ and target images $X_t$ into object detector and align the local and global feature by hierarchical adversarial learning, as shown in Fig. 2(b). Through this framework, the domain gap at both input and feature level can be mitigated. The details of the proposed method are given in the following sections.

### 2.2 Frequency-Based Image Translation

In order to mitigate the domain gap at the input level, a novel frequency-based image translation is presented to obtain translated images without changing their semantic structures. The framework of frequency-based image translation is shown in Fig. 2(a).

First, Fourier transform $\mathcal{F}(\cdot)$ is performed on the image $x$ of size $H \times W$:

$$\mathcal{F}(x)(a,b) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x(h,w) e^{-i2\pi\cdot\left(\frac{ha}{H}+\frac{wb}{W}\right)}, \tag{1}$$

for $a = 0, \ldots, H-1$, $b = 0, \ldots, W-1$.

Then, we decompose the frequency space representation $\mathcal{F}(x)$ of the image into $N$ components $\left\{x^1, x^2, \ldots, x^N\right\}$ of equal bandwidth via band-pass filter $\mathcal{B}(\cdot;\cdot)$:

$$x^{fs} = \mathcal{B}(\mathcal{F}(x); N) = \left\{x^1, x^2, \ldots, x^{N-1}, x^N\right\}, \tag{2}$$

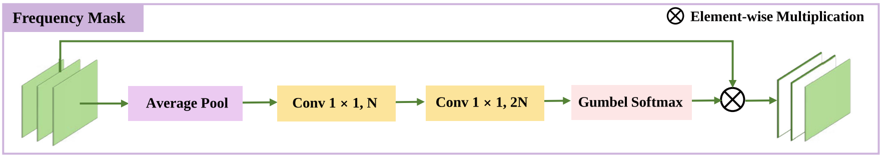


**Fig. 3.** Structure of Frequency Mask.

$$x^n = \begin{cases} \mathcal{F}(x)(i,j), & \text{if } \frac{n-1}{N} < d\left((i,j),(c_i,c_j)\right) < \frac{n}{N} \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

where $c_i$ and $c_j$ denote the image centroid, $d(\cdot,\cdot)$ denotes the Euclidean distance, and $N$ is the number of components. In our experiments, we set $N = 64$.

To identify which frequency component is domain-specific, we design a module called Frequency Mask and its structure is shown in Fig. 3. Motivated by the Squeeze-and-Excitation Networks [15], which model the interdependencies between the channels and recalibrate the channel-wise feature responses adaptively, we design similar structure and add Gumbel-Softmax [18] to make the value close to one-hot vector. '1' means the frequency component is domain-specific, while '0' denotes it is domain-invariant. By Frequency Mask, we find the domain-specific components $DS(x_s^{fs})$:

$$DS(x_s^{fs}) = M\left(x_s^{fs}\right) \cdot x_s^{fs}, \tag{4}$$

where $M(x_s^{fs})$ represents the output of Gumbel-Softmax in Fig. 3. Then, we replace the domain-specific components of the source image with the corresponding ones of the target image:

$$\hat{x}_{s\to t}^{fs} = x_s^{fs} - DS(x_s^{fs}) + DS(x_t^{fs}), \tag{5}$$

After replacing components, we combine all frequency components and perform Inverse Fourier transform $\mathcal{F}^{-1}(\cdot)$. Finally, we obtain the target-like image:

$$x_{s\to t} = \mathcal{F}^{-1}\left(\sum \hat{x}_{s\to t}^{fs}\right). \tag{6}$$

In order to keep the consistency of semantic information, we regulate the reconstruction loss:

$$\mathcal{L}_{rec}(X) = \left\| H(X) - H(\hat{X}) \right\|_1, \tag{7}$$

where $X$ and $\hat{X}$ represent the original and translated image. $H(\cdot)$ represents the band-pass filter that extracts the middle and high-frequency components, which largely capture the semantic information.

### 2.3 Hierarchical Adversarial Feature Learning

After the frequency-based image translation, we put target-like and target images into the object detector and further mitigate the feature-level domain gap by the domain classifier and gradient reversal layer (GRL) [8]. Since different domains could have distinct scene layouts, fully matching the entire distributions of source and target images at the global image-level may fail [4,20]. Therefore, we adopt different strategies on the local and global features.

The global feature alignment module consists of a global domain classifier $D_g$ and a GRL. The GRL connects the global domain classifier and the backbone, which reverses the gradients that flow through the backbone, as shown in Fig. 2(b). It means that the global domain classifier $D_g$ aims to distinguish which domain the global feature comes from, whereas the backbone attempts to confuse the classifier. Here, the source images are given the domain label $d = 0$ and the label is 1 for the target images. The loss of the global feature alignment module is calculated as follows,

$$\mathcal{L}_{glb_s} = -\frac{1}{n_s}\sum_{i=1}^{n_s} D_g\left(F_2(F_1\left(x_i^s\right))\right)^{\gamma} \cdot \log\left(1 - D_g\left(F_2(F_1\left(x_i^s\right))\right)\right), \tag{8}$$

$$\mathcal{L}_{glb_t} = -\frac{1}{n_t}\sum_{i=1}^{n_t} \left(1 - D_g\left(F_2\left(F_1\left(x_i^t\right)\right)\right)\right)^{\gamma} \cdot \log\left(D_g\left(F_2\left(F_1\left(x_i^t\right)\right)\right)\right), \tag{9}$$

$$\mathcal{L}_{glb} = \frac{1}{2}\left(L_{glb_s} + L_{glb_t}\right), \tag{10}$$

where $n_s$ and $n_t$ represent the number of source and target images, $x^s$ and $x^t$ are the target-like and target images, and $F_1$ and $F_2$ denotes the first seven convolutional layers of the backbone VGG16 and the rest convolutional layers. The detailed structure of global domain classifier $D_g$ is shown in Fig. 4(a).

Similar with the adversarial training in global alignment, the local domain classifier $D_l$ and shallow layers of the backbone are connected by the GRL. The loss function of local alignment can be written as:

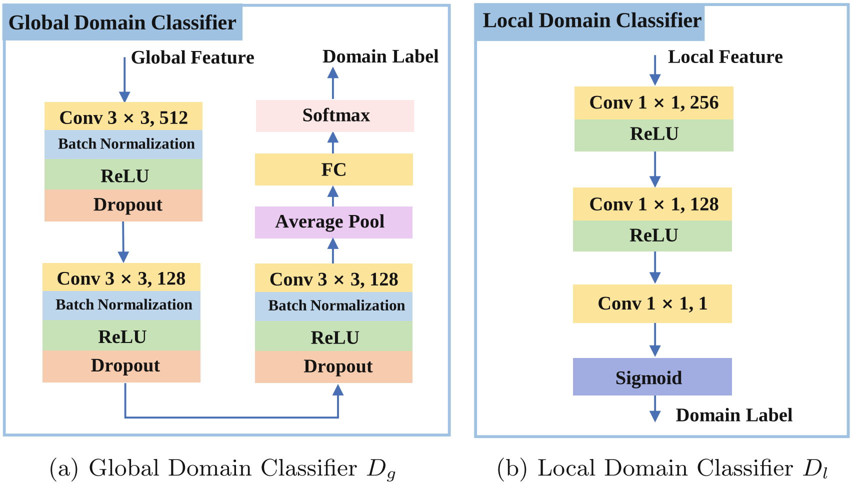
$$\mathcal{L}_{loc_s} = \frac{1}{n_s HW}\sum_{i=1}^{n_s}\sum_{w=1}^{W}\sum_{h=1}^{H} D_l\left(F_1\left(x_i^s\right)\right)_{wh}^2, \tag{11}$$

$$\mathcal{L}_{loc_t} = \frac{1}{n_t HW}\sum_{i=1}^{n_t}\sum_{w=1}^{W}\sum_{h=1}^{H} \left(1 - D_l\left(F_1\left(x_i^t\right)\right)_{wh}\right)^2, \tag{12}$$

$$\mathcal{L}_{loc} = \frac{1}{2}\left(\mathcal{L}_{loc_s} + \mathcal{L}_{loc_t}\right), \tag{13}$$

where $D_l\left(F_1\left(x_i\right)\right)_{wh}$ represents the output of the local domain classifier $D_l$ in each location. The detailed structure of local domain classifier $D_l$ is shown in Fig. 4(b).

To achieve better adaptation, we regularize the domain discriminator. Previous work has shown that it is effective for stabilizing the adversarial training by regularizing the domain classifier with the segmentation loss in domain adaptive segmentation [27]. Similar with this approach, we regularize the domain discriminator with the detection loss. Formally, we extract the different levels of



(a) Global Domain Classifier $D_g$ (b) Local Domain Classifier $D_l$

**Fig. 4.** Structure of Domain Classifiers.

context vectors $v_1$ and $v_2$ from the middle layers of the domain classifiers $D_l$ snd $D_g$ respectively. Each context vector has 128 dimensions. Then, we concatenate the vectors to obtain the aggregated context vector and all region-wise features are concatenated with the aggregated context to train the domain classifiers to minimize the detection loss and domain classification loss, as illustrated in Fig. 2(b).

### 2.4 Overall Objective

We denote the loss of Faster RCNN [1] as $\mathcal{L}_{det}$ and the overall loss function $\mathcal{L}_{total}$ can be summarized as:

$$\mathcal{L}_{total} = \mathcal{L}_{det} + \mathcal{L}_{rec} + \lambda(\mathcal{L}_{glb} + \mathcal{L}_{loc}), \tag{14}$$

where $\lambda$ is the hyper-parameter to balance the detection, reconstruction loss and hierarchical alignment losses.

## 3 Experiments

### 3.1 Datasets

We extensively evaluate our approach on three challenging domain adaptive object detection tasks with distinct domain shifts, including adaptation under different weather (Cityscapes [11] → Foggy Cityscapes [12]), adaptation from the synthetic to the real scene (Sim10K [13] → Cityscapes) and adaptation under different cameras (KITTI [14] → Cityscapes). Cityscapes [11] is a dataset of urban street scenes with 8 categories captured with on-board cameras, which

**Table 1.** Results (%) on the adaptation from Cityscapes to Foggy Cityscapes. 'No DA' indicates the model is only trained with the source images and directly tested on the target images without any domain adaptation. The best results are in **bold**, and the second best results are underlined.

| Methods | person | rider | car | truck | bus | train | moto | bicycle | mAP |
|---|---|---|---|---|---|---|---|---|---|
| No DA | 23.3 | 27.9 | 32.8 | 11.4 | 23.5 | 9.3 | 12.2 | 25.2 | 20.7 |
| DA [6]$_{CVPR'2018}$ | 25.0 | 31.0 | 40.5 | 22.1 | 35.3 | 20.2 | 20.0 | 27.1 | 27.6 |
| DivMatch [9]$_{CVPR'2019}$ | 30.8 | 40.5 | 44.3 | 27.2 | 38.4 | 34.5 | 28.4 | 32.2 | 34.6 |
| SWDA [4]$_{CVPR'2019}$ | 29.9 | 42.3 | 43.5 | 24.5 | 36.2 | 32.6 | 30.0 | 35.3 | 34.3 |
| HTCN [5]$_{CVPR'2020}$ | 33.2 | **47.5** | 47.9 | 31.6 | 47.4 | **40.9** | 32.3 | 37.1 | 39.8 |
| CDN [7]$_{ECCV'2020}$ | 35.8 | 45.7 | 50.9 | 30.1 | 42.5 | 29.8 | 30.8 | 36.5 | 36.6 |
| ATF [20]$_{ECCV'2020}$ | 34.6 | 47.0 | 50.0 | 23.7 | 43.3 | 38.7 | 33.4 | 38.8 | 38.7 |
| Progressive [16]$_{WACV'2020}$ | 36.0 | 45.5 | **54.4** | 24.3 | 44.1 | 25.8 | 29.1 | 35.9 | 36.9 |
| VDD [22]$_{ICCV'2021}$ | 33.4 | 44.0 | 51.7 | **33.9** | **52.0** | 34.7 | 34.2 | 36.8 | 40.0 |
| CDTD [21]$_{IJCV'2021}$ | 31.6 | 44.0 | 44.8 | 30.4 | 41.8 | 40.7 | 33.6 | 36.2 | 37.9 |
| RPA [23]$_{CVPR'2021}$ | 33.4 | 44.3 | 50.1 | 29.9 | 44.8 | 39.1 | 29.9 | 36.3 | 38.5 |
| DDF [24]$_{TMM'2022}$ | **37.2** | 46.3 | 51.9 | 24.7 | 43.9 | 34.2 | 33.5 | **40.8** | 39.1 |
| FIT-DA (Ours) | 36.6 | 45.8 | 52.2 | 32.2 | 48.1 | 34.6 | **34.7** | 37.2 | **40.2** |

has 2975 training images and 500 validating images. Foggy Cityscapes [12] is the synthetic foggy version of Cityscapes. Sim10K [13] is a virtual dataset including 10000 images generated by the Grand Theft Auto gaming engine. KITTI [14] is an autonomous driving dataset that has 7481 images, which is captured by a standard station wagon with two high-resolution video cameras. In the test, we use mean average precision (mAP) metrics for evaluation.

## 3.2 Implementation Details

Our detector is original Faster R-CNN [1] without extra modules. We adopt VGG-16 [28] pre-trained on ImageNet [29] as our backbone. In our experiments, the shorter side of the image is resized to 600. Each batch is composed of one source image and one target image. The networks are trained with a learning rate of 0.001 for 50K iterations, then with a learning rate of 0.0001 for 20K more iterations. We use a momentum of 0.9 and a weight decay of 0.0005. $N$ is 64 in Eq. (2). For Sim10K $\rightarrow$ Cityscapes, we set $\lambda = 0.1$ in Eq. (14). For the rest two tasks, we set $\lambda = 1$. Our method is implemented with PyTorch.

## 3.3 Comparison Experiments

**Adaptation Under Different Weather.** Table 1 shows the performance of our method on Cityscapes → Foggy Cityscapes. We can see that our method alleviates the domain gap across different weather conditions and outperforms all competitors in Table 1. Compared with GANs-based methods: DivMatch [9], Progressive [16] and CDTD [21], our method improves the result by +5.6%, +3.3% and +2.3% in mAP, which demonstrates the advantage of the proposed Frequency-based Image Translation for domain adaptive object detection.

**Table 2.** Sim10K to Cityscape.

| Methods | mAP |
|---|---|
| Source Only | 34.2 |
| DA [6]$_{CVPR'2018}$ | 39.0 |
| SWDA [4]$_{CVPR'2019}$ | 40.1 |
| HTCN [5]$_{CVPR'2020}$ | 42.5 |
| ATF [20]$_{ECCV'2020}$ | 42.8 |
| CDTD [21]$_{IJCV'2021}$ | 42.6 |
| RPA [23]$_{CVPR'2021}$ | 45.7 |
| DDF [24]$_{TMM'2022}$ | 44.3 |
| FIT-DA (Ours) | **48.6** |

**Table 3.** KITTI to Cityscapes.

| Methods | mAP |
|---|---|
| Source Only | 32.2 |
| DA [6]$_{CVPR'2018}$ | 38.5 |
| SWDA [4]$_{CVPR'2019}$ | 43.1 |
| CDN [7]$_{ECCV'2020}$ | 44.9 |
| ATF [20]$_{ECCV'2020}$ | 42.1 |
| Progressive [16]$_{WACV'2020}$ | 43.9 |
| DDF [24]$_{TMM'2022}$ | 46.0 |
| FIT-DA (Ours) | **46.3** |

**Table 4.** Ablation analysis of our method. LA is local feature alignment and GA is global feature alignment. CTV represents the context vector and FIT denotes the frequency-based image translation.

| Methods | LA | GA | CTV | FIT | C → F | S → C |
|---|---|---|---|---|---|---|
| Source only | | | | | 20.7 | 34.2 |
| FIT-DA | | | | ✓ | 31.8 | 40.5 |
| FIT-DA | | ✓ | ✓ | ✓ | 38.5 | 46.4 |
| FIT-DA | ✓ | | ✓ | ✓ | 35.7 | 42.2 |
| FIT-DA | ✓ | ✓ | | ✓ | 37.4 | 44.3 |
| FIT-DA | ✓ | ✓ | ✓ | | 34.5 | 40.3 |
| FIT-DA (Ours) | ✓ | ✓ | ✓ | ✓ | **40.2** | **48.6** |

**Adaptation from the Synthetic to Real Scene.** We evaluate the detection performance on car on Sim10K to Cityscapes benchmark. As we can see the results in Table 2, our method has a significant performance boost over other methods, further indicating the effectiveness of our method.

**Adaptation Under Different Cameras.** There exists a domain gap between datasets captured through different cameras due to the diversity of hardware devices. We conduct the cross-camera adaptation from KITTI to Cityscapes. The results are presented in Table 3, and our method has competitive performance among all the comparison methods.

### 3.4 Ablation Study

**Effectiveness of Each Component.** We conduct the ablation experiments on Cityscapes → Foggy Cityscapes (C → F) and Sim10K → Cityscapes (S → C) to validate the effectiveness of each module in our framework. The results in Table 4 show that all the modules contribute to the performance improvement

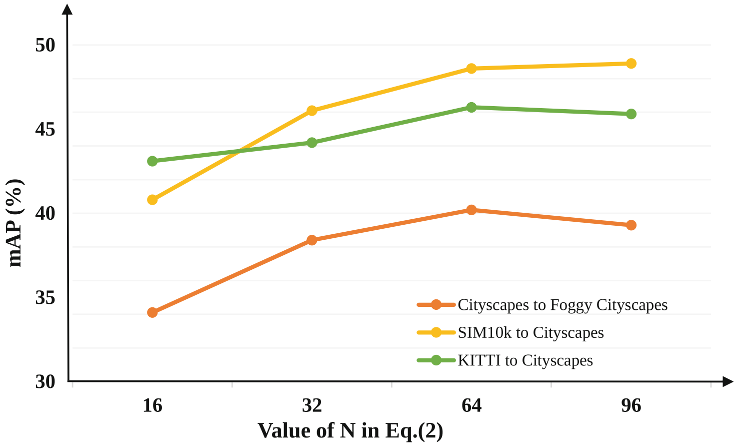


**Fig. 5.** Detection performances on the three benchmarks with different values of $N$.

**Table 5.** Performances with different choices of domain-specific frequency components.

| Settings | Choice | C → F | S → C |
|---|---|---|---|
| Non-learnable | FC[1] | 38.3 | 44.2 |
| | FC[1, 2] | 37.3 | 45.9 |
| | FC[1, 2, 3] | 36.1 | 41.8 |
| | FC[1, 2, 3, 4] | 33.9 | 40.0 |
| Learnable | Frequency mask | **40.2** | **48.6** |

(especially FIT module), which indicates the effectiveness of each component in our method.

**Method of Choosing Domain-Specific Frequency Components.** In FIT framework, Frequency Mask is the core module, which choose the domain-specific frequency component in a learnable way. We compare the adaptation performance of using Frequency Mask with using fixed low-frequency components to determine domain-specific components in Table 5. The results suggest the Frequency Mask captures the domain-specific information better. Although low-frequency components largely captures domain-specific information, the distributions of domain-specific components are not completely consistent for images from different domains, making it difficult to capture these components preciously just using fixed low-frequency components.

**Value of $N$ in Eq. 2.** $N$ is the number of frequency components after Fourier decomposing. Figure 5 shows the influence on adaptation performance with different $N$. As $N$ is related to the division of frequency bands, which is critical for finding domain-specific frequency components, it affects the quality of translated images. In our experiments, $N = 64$ is the best choice considering the performance on the three benchmarks.

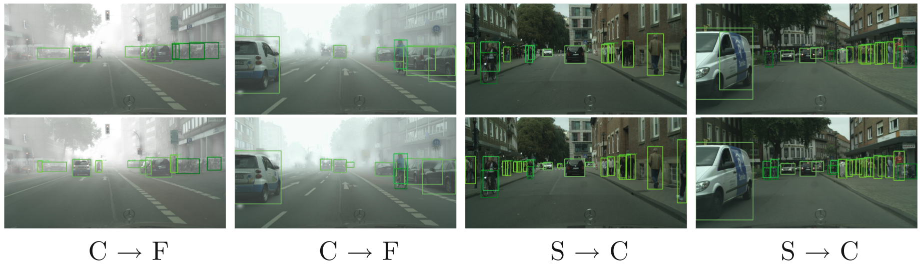


**Fig. 6.** Example results on Cityscapes to Foggy Cityscapes (C → F) and Sim10K to Cityscapes (S → C). The fist row is the results of SWDA and the second row is the results of our FIT-DA. The class and score predictions are at the top left corner of the bounding box. Zoom in to visualize the details.

### 3.5 Visualization

Figure 6 illustrates some examples of detection results on Cityscapes to Foggy Cityscapes and Sim10K to Cityscapes. Obviously, our method produces more accurate bounding box predictions and has a stronger ability to detect obscured instances.

## 4 Conclusion

In this paper, a novel Frequency-based Image Translation (FIT) method for DAOD is presented to reduce domain shift at the input level. Compared to other image translation methods for DAOD, it is embedded in the detection network and does not need extra time-consuming training. Additionally, we introduce hierarchical adversarial feature learning to further mitigate the domain gap at the feature level. Meanwhile, a joint loss function is designed to optimize the entire network in an end-to-end manner. Extensive experiments on three challenging DAOD benchmarks validate the effectiveness of our method. In the future, we will utilize the frequency information in the feature space to investigate the feature augmentation for DAOD.

**Acknowledgements.** This work was supported in part by the National Key Research and Development Plan of China under Grant 2020AAA0108902 and the Strategic Priority Research Program of Chinese Academy of Science under Grant XDB32050100.

## References

1. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. Adv. Neural. Inf. Process. Syst. **28**, 91–99 (2015)

2. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
3. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: fully convolutional one-stage object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 9627–9636 (2019)
4. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6956–6965 (2019)
5. Chen, C., Zheng, Z., Ding, X., Huang, Y., Dou, Q.: Harmonizing transferability and discriminability for adapting object detectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8869–8878 (2020)
6. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster R-CNN for object detection in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3339–3348 (2018)
7. Su, P., et al.: Adapting object detectors with conditional domain normalization. In: Vedaldi, Andrea, Bischof, Horst, Brox, Thomas, Frahm, Jan-Michael. (eds.) ECCV 2020. LNCS, vol. 12356, pp. 403–419. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58621-8_24
8. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning, pp. 1180–1189. PMLR (2015)
9. Kim, T., Jeong, M., Kim, S., Choi, S., Kim, C.: Diversify and match: a domain adaptive representation learning paradigm for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12456–12465 (2019)
10. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
11. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
12. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. Int. J. Comput. Vis. **126**(9), 973–992 (2018). https://doi.org/10.1007/s11263-018-1072-8
13. Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S.N., Rosaen, K., Vasudevan, R.: Driving in the matrix: can virtual worlds replace human-generated annotations for real world tasks? In: 2017 IEEE International Conference on Robotics and Automation, pp. 746–753. IEEE (2017)
14. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the KITTI dataset. Int. J. Robot. Res. **32**(11), 1231–1237 (2013)
15. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
16. Hsu, H.K., et al.: Progressive domain adaptation for object detection. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, pp. 749–757 (2020)
17. Oppenheim, A.V.: Discrete-Time Signal Processing. Pearson Education India (1999)
18. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with Gumbel-Softmax. arXiv preprint arXiv:1611.01144 (2016)

19. Li, J., Duan, L.Y., Chen, X., Huang, T., Tian, Y.: Finding the secret of image saliency in the frequency domain. IEEE Trans. Pattern Anal. Mach. Intell. **37**(12), 2428–2440 (2015)
20. He, Zhenwei, Zhang, Lei: Domain adaptive object detection via asymmetric tri-way faster-RCNN. In: Vedaldi, Andrea, Bischof, Horst, Brox, Thomas, Frahm, Jan-Michael. (eds.) ECCV 2020. LNCS, vol. 12369, pp. 309–324. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58586-0_19
21. Shen, Z., et al.: CDTD: a large-scale cross-domain benchmark for instance-level image-to-image translation and domain adaptive object detection. Int. J. Comput. Vis. **129**(3), 761–780 (2021). https://doi.org/10.1007/s11263-020-01394-z
22. Wu, A., Liu, R., Han, Y., Zhu, L., Yang, Y.: Vector-decomposed disentanglement for domain-invariant object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 9342–9351 (2021)
23. Zhang, Y., Wang, Z., Mao, Y.: RPN prototype alignment for domain adaptive object detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12425–12434 (2021)
24. Liu, D., et al.: Decompose to adapt: cross-domain object detection via feature disentanglement. IEEE Trans. Multimed. (2022)
25. Chen, Y., Li, G., Jin, C., Liu, S., Li, T.: SSD-GAN: measuring the realness in the spatial and spectral domains. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 1105–1112 (2021)
26. Piotrowski, L.N., Campbell, F.W.: A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. Perception **11**(3), 337–346 (1982)
27. Sankaranarayanan, S., Balaji, Y., Jain, A., Lim, S.N., Chellappa, R.: Learning from synthetic data: addressing domain shift for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3752–3761 (2018)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
29. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)