Journal of Intelligent Computing and Networking
https://www.ffspub.com/index.php/jicn/index
ISSN 3079-9228 (print)
E-mail: jicn.office@ffspub.com

Article

Exploring Test-time Adaptive Object Detection in the Frequency Domain

Kunyu Wang[1], Qi Qi[2], Wei Zhai[1,†]
[1]*Department of Automation, University of Science and Technology of China, Hefei 230026, China*
[2]*Department of Information Systems and Analytics, National University of Singapore, 117417, Singapore*
[†]*E-mail: wzhai056@ustc.edu.cn*

Received: October 18, 2025 / Revised: November 30, 2025 / Accepted: December 01, 2025 / Published online: December 08, 2025

Abstract: Continual test-time adaptive object detection (CTTA-OD) aims to online adapt a pre-trained detector to changing environments. Most CTTA-OD methods adapt the test domain shifts in the spatial domain, while overlooking the potential of the frequency cues. In this paper, we propose a novel frequency-specific adaptation framework that enables the detector rapidly adapt to the evolving target domain, enabling stable detection performance at test-time. Our motivation stems from the observation that distribution shifts caused by different perturbations primarily manifest in specific frequency bands. By selectively adapting only the affected bands, we reduce the optimization dimension, enabling rapid convergence to the new distribution. Specifically, we first decompose the spectrum into distinct bands by splitting it into patches, with each band handled by a specialized frequency adaptation expert. A learnable sparse gating network then selects the top-k affected frequency bands, assigning the corresponding experts to adapt each target domain. Moreover, we introduce a frequency alignment loss at both patch and instance levels, guiding the learning of the gating network and experts. Experiments on three benchmarks show that our method accelerates the CTTA-OD process within fewer batches, outperforming recent SOTA methods.

Keywords: Test-time adaptation; object detection; frequency learning
https://doi.org/10.64509/jicn.12.40

1 Introduction

Object detection [1, 2] is a fundamental task in computer vision, with diverse applications such as autonomous driving. However, real-world machine perception systems [3, 4] operate in unpredictable and constantly changing environments, where dynamic domain shifts can occur. Applying pre-trained detectors in these scenarios often results in substantial performance degradation. Therefore, it is crucial for source pre-trained detectors to rapidly adapt to continually changing target domains during inference, i.e., continual test-time adaptive object detection (CTTA-OD), ensuring stable performance in dynamic environments.

While promising, existing CTTA-OD methods [5–8] primarily adapt to test distribution shifts in the spatial domain. However, perturbation factors [9–11], such as weather variations, typically affect the entire scene, inducing alterations across all pixels in the spatial domain. This requires optimization across the high-dimensional space of the entire image to accommodate pixel-level distribution shifts, resulting in a slow convergence process to the new distribution [12, 13]. In contrast, the Fourier transform [14], with its ability to decompose and compress information, partitions the image into distinct frequency bands. Our experiments show that perturbations such as fog, night, and rain, while affecting the entire scene in the spatial domain, primarily manifest in specific frequency bands [15] after being transformed into the frequency domain, as shown in Figure 1. This observation motivates our approach of selectively adapting the frequency bands relevant to the perturbations, which reduces the dimensionality of the optimization space, accelerates the adaptation process, and enables the model to rapidly adapt to continually changing target domains.

In this paper, we propose a novel frequency-specific learning framework that achieves rapid online adaptation. Our core idea is to selectively adapt the most affected frequency bands by different target domains, thereby accelerating CTTA-OD process. Specifically, we first transform spatial features into the frequency domain, where frequency components are compactly organized from low to high frequencies. Based on this

[†] Corresponding author: Wei Zhai
[*] Academic Editor: Chunxiao Jiang

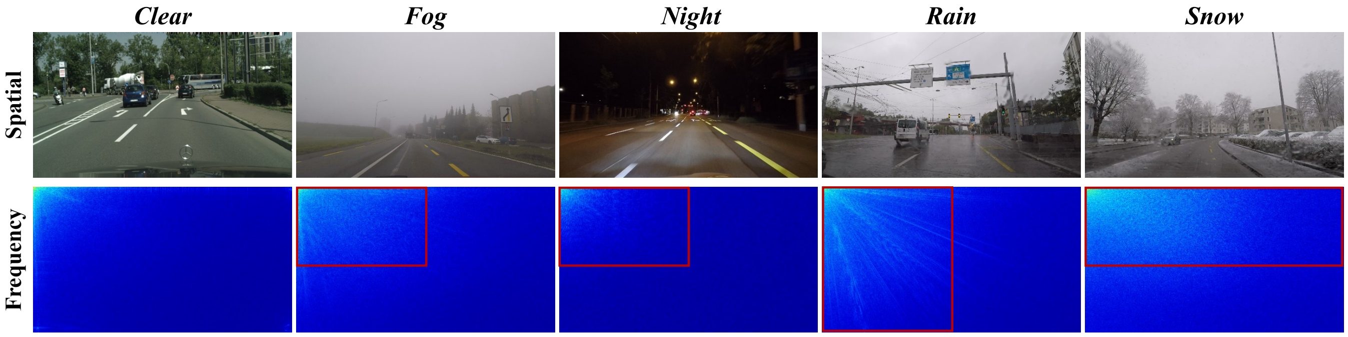


Figure 1: Visual comparison of spatial domain and frequency domain representations under perturbations. Red boxes highlight frequency bands with noticeable changes.

property, we decompose the frequency spectrum into distinct bands by splitting it into patches, with each band being handled by a specialized frequency adaptation expert. We then employ a learnable sparse gating network to select the top-k most affected frequency bands, and assign the corresponding frequency experts to adapt each target domain. Furthermore, we propose a frequency alignment loss that minimizes the MSE between source and target frequency distributions at both patch and instance levels, guiding the gating network and frequency experts toward effective selection and adaptation. By focusing learning on specific frequency bands for different target domains, our method accelerates the CTTA-OD process within a few adaptation batches, thereby improving overall performance. We conduct extensive experiments across three benchmarks, achieving an average mAP improvement of 1.7% over state-of-the-art methods, validating the superiority of our method.

In summary, our contributions are listed as follows:

1. We provide a novel frequency perspective for CTTA-OD, where selective frequency bands adaptation reduces optimization dimensional and enables rapid adaptation.
2. We propose a frequency-specific adaptation framework that uses a sparse gating network to select the most affected frequency bands for each target domain and assign specialized experts for focused adaptation.
3. We propose a frequency alignment loss that aligns the frequency distributions of the source and target domains at both patch and instance levels, guiding the effective selection and adaptation of gating network and experts.

2 Related Works

Continual Test-time Adaptive Object Detection. CTTA-OD [5–8, 16–19] aims to online adapts a pre-trained detector to handle a sequence of evolving target domains. Current methods employ learning paradigms such as statistics calibration, pseudo-labeling, and consistency regularization. For instance, MemCLR [5] proposes an online adaptation framework using memory-enhanced contrastive learning with the MemXformer module to store prototypes and generate contrastive pairs. STFAR [6] adopts a teacher-student network, generating pseudo-labeled objects and incorporating feature alignment regularization to improve robustness. ActMAD [7] focuses on fine-grained alignment of activation statistics between test and training data. MLFA [19] introduces multi-level feature alignment, aligning distributions globally and at the category cluster level to extract domain-invariant features. However, the coupling of perturbations with scene semantics in the spatial domain makes rapid online adaptation challenging, due to the high-dimensional optimization required for pixel-level shifts across the entire image. In contrast, we leverage the frequency cues, where perturbation-induced shifts concentrate in specific bands, and propose frequency-specific learning for rapid adaptation.

Fourier Transform in Vision. Fourier transform (FT) [14] has been widely used in signal processing for decades [20]. Recently, its integration into deep neural networks (DNNs) has gained attention. In vision, FT's ability to convert spatial domain data into the frequency domain makes it invaluable for extracting critical features from noisy or high-dimensional data [21]. Based on this, current research explores applications of FT in optimizing DNN architectures [22–26], enhancing data augmentation [27–30], analyzing DNN behaviors [31–33], etc. Leveraging FT's ability to decompose and compress high-dimensional information, it is a promising tool for online dynamic scenarios like CTTA-OD, which require rapid adaptation. When perturbations, coupled with scene semantics in the spatial domain, are transformed to the frequency domain, they primarily concentrate in specific frequency bands. This motivates our use of frequency-specific learning, instead of pixel-level spatial adaptation, enabling rapid online adaptation and improving overall performance.

Mixture of Experts. The concept of mixture of experts (MoE) [34], initially introduced in [35], is based on the idea that different parts of a network, known as experts, specialize in distinct tasks or aspects of the data. Each expert learns different discriminative features, while the gating network dynamically generates scores that weight the outputs of the expert networks. Recent advancements [36–39] have introduced sparse activation, where only a few experts are selected based on gating scores, significantly reducing the computational cost in large-scale model training [40–42]. Related to our work, we draw inspiration from the MoE framework. We adopt frequency-specific learning to selectively adapt to the most affected frequency bands, where the selection of these bands dynamically changes with the target domain. To handle this, we assign the specialized adaptation expert for each band and employ a dynamic sparse gating mechanism to control the assignment of frequency experts for each target domain, enabling fast and flexible adaptation.

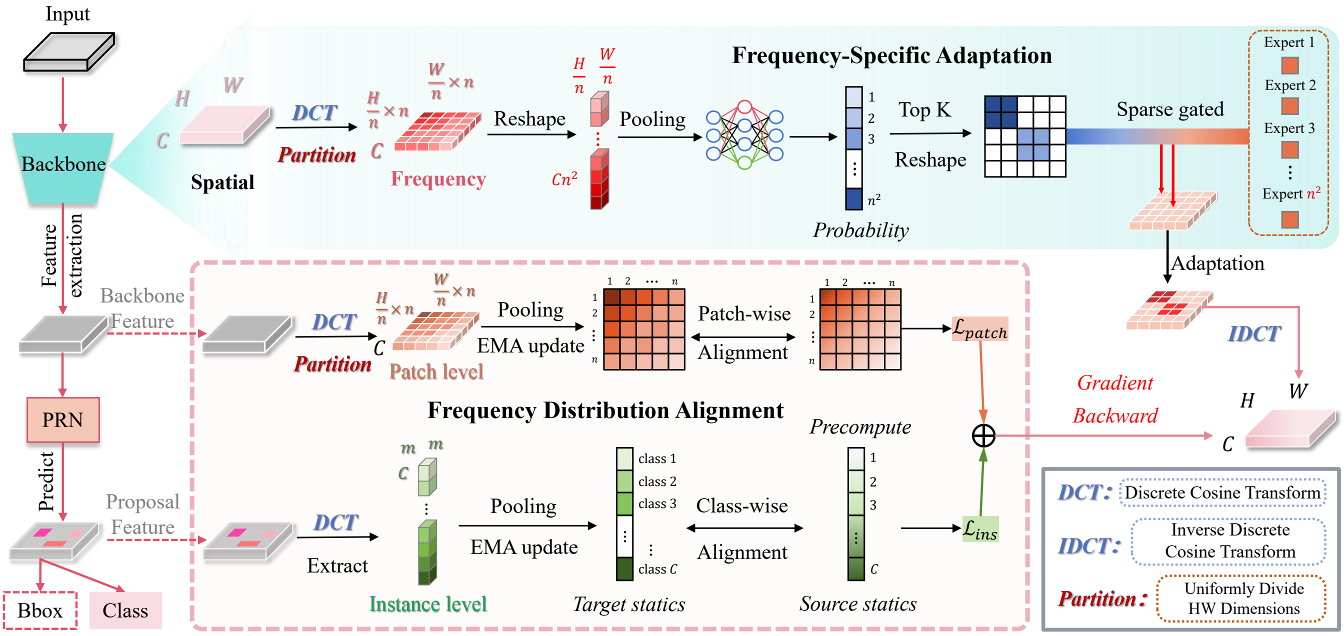


Figure 2: Overview of the proposed method, which incorporates two key components: frequency-specific adaptation and frequency distribution alignment.

3 Methodology

In this section, we first introduce the task setup in Section 3.1, followed by our frequency-specific adaptation framework in Section 3.2. We then present the frequency distribution alignment loss at both patch-level and instance-level in Section 3.3, before concluding with the overall optimization objective in Section 3.4. Figure 2 provides an overview of our method.

3.1 Problem Definition

Assume we have an object detector pre-trained on the source domain $D_s = \{x_s, y_s\}$, where x_s denotes the source images and y_s denotes the corresponding bounding boxes and category labels. Our goal is to adapt the detector to a sequence of continually changing target domains $\{D_t^1, D_t^2, ..., D_t^N\}$ using only the target data while making predictions. The target domain at time period n is denoted as $D_t^n = \{x_t^n\}$, where x_t^n denotes the target images at time period n and $P_{test}^n \neq P_{test}^{n-1}$. Following prior work [6–8], as the source domain is inaccessible during adaptation, pre-computed source feature statistics, such as the mean and variance, are available.

3.2 Frequency-Specific Adaptation

Frequency-specific adaptation aims to selectively adjust the most affected frequency bands of features across different target domains, enabling rapid CTTA-OD. Given a target feature $x \in \mathbb{R}^{C\times H\times W}$ in the spatial domain, we first transform it into the frequency domain using a channel-wise two-dimensional Discrete Cosine Transform (2D DCT) [43]. Specifically, 2D DCT represents a signal by projecting it onto a set of orthogonal cosine basis functions. The basis function for the 2D DCT is formulated as:

$$B_{h,w}^{i,j} = \cos\left(\frac{\pi}{H}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{\pi}{W}\left(j+\frac{1}{2}\right)\right). \quad (1)$$

Each basis function $B_{h,w}^{i,j}$ corresponds to a frequency component indexed by (h,w), while the indices (i,j) enumerate all spatial positions of the input feature. Based on this definition, the DCT coefficient $f_{h,w}^{2d}$ is obtained by taking the inner product between the spatial feature x^{2d} and the basis function $B_{h,w}^{i,j}$. This directly rewrites Equation 1 into:

$$f_{h,w}^{2d} = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j}^{2d} B_{h,w}^{i,j} \quad (2)$$
$$\text{s.t.} \quad h \in \{0,1,\ldots,H-1\}, w \in \{0,1,\ldots,W-1\},$$

where $f^{2d} \in \mathbb{R}^{H\times W}$ represents the 2D frequency spectrum and $x^{2d} \in \mathbb{R}^{H\times W}$ is the input spatial feature. This formulation explicitly shows how each frequency coefficient is computed from all spatial locations, thereby verifying how spatial features are transformed into frequency domain. Given that the frequency feature $f \in \mathbb{R}^{C\times H\times W}$ exhibits a characteristic distribution, where low frequencies concentrated in the top-left corner and high frequencies in the bottom-right corner, we decompose the spectrum into distinct frequency bands by partitioning the 2D plane. Specifically, given a partition factor n, we uniformly divide both the height H and width W dimensions into n segments. This transformation reshapes the original tensor $f \in \mathbb{R}^{H\times W\times C}$ into a new shape:

$$f' \in \mathbb{R}^{C\times n^2\times \frac{H}{n}\times\frac{W}{n}}. \quad (3)$$

For each frequency band $f'_{i,j} \in \mathbb{R}^{C\times\frac{H}{n}\times\frac{W}{n}}$, where $i,j \in \{1,\ldots,n\}$, we assign a specialized frequency expert $E_{i,j}$ to adapt the distribution shift. The expert structure consists of a simple yet effective point-wise convolution followed by batch normalization:

$$\hat{f}_{i,j} = \text{BN}(\text{Conv}_{1\times1}(f'_{i,j})). \quad (4)$$

To dynamically select the most affected frequency bands for each target domain, we employ a learnable sparse gating network with a top-k routing policy. The process for generating sparse weights operates as follows:

$$f_g' = \mathrm{GAP}(f') + \mathrm{GMP}(f'), \tag{5}$$

$$P_g = f_g' \cdot W_g + N(0,1) \cdot \mathrm{SoftPlus}(f_g' \cdot W_{noise}), \tag{6}$$

$$W_g = \begin{cases} 1 & \text{if } v_i \text{ is in the top } k \text{ elements of } P_g, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where we first process the features f' using average pooling $\mathrm{GAP}(\cdot)$ and maximum pooling $\mathrm{GMP}(\cdot)$ operations, followed by summing them to obtain $f_g' \in \mathbb{R}^{Cn^2}$. Next, f_g' is passed through the fully connected layers $W_g, W_{noise} \in \mathbb{R}^{Cn^2 \times n^2}$ to produce the probability vector $P_g \in \mathbb{R}^{n^2}$. The addition of a learnable noise term introduces randomness, facilitating the transition during expert selection [38]. By applying the Top-K operation, we select the k positions with the highest values in P_g, assign a value of 1 to the selected positions, and set the unselected expert weights to 0, resulting in the $W_g \in \mathbb{R}^{n^2}$. Based on the selected frequency bands from the sparse weight vector W_g, we use the corresponding frequency experts to adapt to the distribution changes, while the unselected bands remain unchanged:

$$\bar{f}_{i,j} = \begin{cases} \hat{f}_{i,j} & \text{if } W_g^{i,j} = 1, \\ f'_{i,j} & \text{if } W_g^{i,j} = 0, \end{cases} \quad \text{for} \quad i,j \in \{1,\ldots,n\}. \tag{8}$$

Finally, we apply the channel-wise inverse 2D DCT to $\bar{f}$ to return to the spatial domain, obtaining the adjusted spatial feature $\bar{x} \in \mathbb{R}^{C \times H \times W}$:

$$\bar{x}_{i,j}^{2d} = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} \bar{f}_{h,w}^{2d} B_{h,w}^{i,j}, \\ \text{s.t. } i \in \{0,1,\ldots,H-1\}, \quad j \in \{0,1,\ldots,W-1\}. \tag{9}$$

3.3 Frequency Distribution Alignment

To better adapt the distribution shift in the frequency domain, we introduce a frequency distribution alignment loss. This loss aligns the frequency feature distributions between the source and target domains, guiding both gating network selection and frequency expert adaptation. For the object detection task, we propose aligning two types of features: backbone and foreground features, considering the importance of instance information.

In a typical Faster RCNN framework [44], the backbone outputs global feature map $x_{\text{img}} \in \mathbb{R}^{C \times H \times W}$, along with proposal features after RPN and ROI alignment, denoted as $x_{\text{ins}} \in \mathbb{R}^{C_1 \times m \times m}$, where m represents the output size of the ROI alignment. We first collect the global feature map and proposal features for each source image, and apply 2D DCT to transform them into frequency domain, yielding f_{img} and f_{ins}. Then, we pre-compute the mean statistics of these source features. For global feature maps, instead of using global average pooling as in prior work [6–8], we build upon our earlier practice of dividing the frequency features into n^2 patches. We apply average pooling on each patch to compute patch-level mean statistics, resulting in n^2 statistics. This ensures consistent granularity with the gating and expert networks, facilitating frequency bands selection and adaptation:

$$\mu_s^{i,j} = \frac{1}{N_{\text{img}}} \sum_{f_{\text{img}} \in D_s} \frac{n^2}{HW} \sum_{\frac{H}{n}, \frac{W}{n}} f_{img}^{i,j}, \tag{10}$$

where $f_{img}^{i,j} \in \mathbb{R}^{C \times \frac{H}{n} \times \frac{W}{n}}$, N_{img} denotes the total number of source domain global feature maps. For proposal features, which are already at the instance level, we apply global average pooling to each proposal. We then compute class-wise instance-level mean statistics using the object category ground-truth labels, resulting in C statistics, which correspond to C object categories:

$$\mu_s^c = \frac{1}{N_{\text{ins}}^c} \sum_{f_{\text{ins}}^c \in D_s} \frac{1}{m^2} \sum_{m,m} f_{ins}^c, \tag{11}$$

where $c \in \{1,\ldots,C\}$, $f_{ins}^c \in \mathbb{R}^{C_1 \times m \times m}$, N_{ins}^c denotes the total number of proposals for a specific class c.

As batches of target data arrive, we collect the global feature maps and proposal features for each batch through the forward pass. Since the target domain lacks object category labels, we rely on the high-quality pseudo-labels generated by the network for each proposal. Proposals with background scores exceeding a specific threshold (e.g., 0.5) are discarded, while the remaining proposals are assigned to the foreground class with the highest probability, generating the corresponding class labels. The computation of patch-level and instance-level mean statistics for each target batch follows the same approach as in the source domain:

$$\mu_t^{i,j} = \frac{1}{B_{\text{img}}} \sum_{f_{\text{img}} \in B_t} \frac{n^2}{HW} \sum_{\frac{H}{n}, \frac{W}{n}} f_{img}^{i,j}, \tag{12}$$

$$\mu_t^c = \frac{1}{B_{\text{ins}}^c} \sum_{f_{\text{ins}}^c \in B_t} \frac{1}{m^2} \sum_{m,m} f_{ins}^c, \tag{13}$$

where B_t denotes a batch of target data, B_{img} and B_{ins}^c denotes the number of global features and proposals of class c in a batch. To estimate the statistics for the target domain, we update them using an exponentially moving average:

$$\mu_t^{i,j}(0) = \mu_s^{i,j}, \quad \mu_t^c(0) = \mu_s^c, \tag{14}$$

$$\mu_t^{i,j}(T) = (1-\alpha) \cdot \mu_t^{i,j}(T-1) + \alpha \cdot \mu_t^{i,j}, \tag{15}$$

$$\mu_t^c(T) = (1-\alpha) \cdot \mu_t^c(T-1) + \alpha \cdot \mu_t^c, \tag{16}$$

where $\mu(\cdot)$ denotes mean statistics at a given time, and α is the smoothing factor.

To align the frequency feature distributions, we first compute the MSE of per-patch mean statistics at the global feature map level and obtain the patch-level frequency distribution alignment loss through averaging:

$$\mathscr{L}_{\text{patch}} = \frac{1}{n^2} \sum_{i,j} \|\mu_t^{i,j}(T) - \mu_s^{i,j}\|_2, \tag{17}$$

At the instance level, we calculate the class-wise mean statistics MSE. Following prior work [8], we introduce a weighting scheme to better align features of infrequent classes, addressing the severe class imbalance where certain instances may appear multiple times in a single image:

$$\omega_c = \log\left(\frac{\max\{o_i\}_{i=1}^{c}}{o_c}\right) + 0.01 \tag{18}$$

$$\mathscr{L}_{\text{ins}} = \frac{1}{C}\sum_{c} \omega_c \cdot \|\mu_t^c(T) - \mu_s^c\|_2, \tag{19}$$

where o_i denotes the number of occurrences of class i in the target domain, and ω_i represents the loss weight for class i.

3.4 Optimization Objective

During the CTTA-OD process, we estimate the frequency statistics for the target domain through the forward pass. We then calculate the patch-level and instance-level frequency alignment losses by comparing these statistics with the pre-computed source frequency statistics. These losses are used to optimize the frequency-specific adapter in the detector, enabling fast and stable online adaptation:

$$\mathscr{L}_{\text{total}} = \mathscr{L}_{\text{patch}} + \lambda \mathscr{L}_{\text{ins}}, \tag{20}$$

where λ is the hyper-parameter that balances the losses.

3.5 Theoretical Explanation of Our Design

Our framework is motivated by the observation that domain shifts caused by real-world corruptions are concentrated in specific frequency bands. To justify this, we provide a theoretical analysis based on standard degradation models and frequency characteristics of driving scenes. We model a corrupted image as:

$$I = \mathscr{A}I_0 + \varepsilon, \tag{21}$$

where I_0 is the clean image, $\mathscr{A}$ is a corruption operator (e.g., blur, haze), and ε denotes additive noise. Applying the linear DCT transform $\mathscr{T}(\cdot)$ gives:

$$F = \mathscr{T}(I) = \widehat{\mathscr{A}}F_0 + F_\varepsilon, \tag{22}$$

where F_0 and F are the frequency representations of I_0 and I, and $\widehat{\mathscr{A}}$ is the frequency response of the corruption. Thus, the frequency-domain shift is:

$$\Delta F = F - F_0 = (\widehat{\mathscr{A}} - \mathbf{I})F_0 + F_\varepsilon. \tag{23}$$

This shows that the shift depends on the frequency-selective behavior of $\widehat{\mathscr{A}}$. For example, blur and haze act as low-pass filters suppressing high frequencies, while motion blur and structured noise (e.g., rain) impact specific mid- or high-frequency bands. Therefore, ΔF is typically non-uniform across the spectrum. Additionally, driving scenes contain large low-frequency areas (e.g., sky, roads) and semantically rich mid-/high-frequency regions (e.g., objects, edges). Perturbations that affect object visibility or contrast lead to spectral deviations that align with these regions.

In summary, the non-uniform frequency response of real-world corruptions, combined with the frequency structure of typical scenes, explains why domain shifts concentrate in specific frequency bands. This supports our design of frequency-specific adaptation.

4 Experiments

4.1 Experimental Settings

Datasets. We conduct extensive experiments on three benchmarks, including **Cityscapes-C**, **ACDC**, and **SHIFT**. The **Cityscapes** dataset [45] consists of 2,975 training images and 500 validation images with 8 categories of objects. We construct **Cityscapes-C** based on benchmark robustness tasks [9], selecting four common corruption types: fog, motion blur, contrast, defocus blur. These corruptions are applied to the validation set at the maximum severity level 5, with each corruption type forming an individual target domain consisting of 500 images. The **ACDC** dataset [10] shares the same class categories as Cityscapes but includes four different adverse visual conditions: fog, night, rain, and snow, with each condition containing 400 unlabeled images. The **SHIFT** dataset [11] is a synthetic dataset for autonomous driving, simulating real-world environmental changes. We select 3,000 clear daytime images from SHIFT-discrete as the source domain training set, and 500 images each of foggy, rainy, night, and overcast conditions as the target test domains.

Implementation Details. We use Faster R-CNN [44] with ResNet50 [46] backbone pre-trained on ImageNet [47] as the detector. During test-time adaptation, we update the frequency-specific adapter while freezing all other parameters pre-trained on the source domain. The batch size is 2, and the learning rate for the Adam optimizer is 0.005. Hyper-parameters are set as follows: partition factor $n = 2$, top-k percentage $p = 50\%$, smoothing factor $\alpha = 0.99$, and loss balancing coefficient $\lambda = 3$. We use mAP@50 (%) as the evaluation metric for detection performance. Following prior work [48], the source model is continually adapted to the target domains for 10 cycles. For computational efficiency, the baseline detector requires approximately 280 GFLOPs for a single forward pass, and contains about 31M parameters. Our method introduces 0.18M additional parameters (less than 0.6% of the total). During test-time adaptation, each step includes a forward and a backward pass. The backward pass involves two components: (1) gradient propagation across layers, which passes through the entire network and has a similar cost to the forward pass (280 GFLOPs), and (2) gradient computation for the learnable parameters, which only applies to the adapter and costs around 4 GFLOPs. As a result, each adaptation step requires approximately 564 GFLOPs in total. On an RTX 3090 GPU with batch size 1, the baseline detector runs at 10.2 FPS, while our method achieves 4.8 FPS with one adaptation step per image.

4.2 Comparisons on Benchmark Datasets

We compare our method with several recent representative CTTA-OD approaches, including DUA [16], CoTTA [48], STFAR [6], MemCLR [5], ActMAD [7], MLFA [19], ViDA [49] and WHW [8]. For open-source methods, we follow their implementation details, adopt the corresponding hyper-parameter settings, and adapt them to the Faster R-CNN detector. For non-open-source methods, we reproduce them based on the descriptions in their papers. Direct test refers to directly evaluating the source domain trained detector on the target domains. We conduct comparative experiments on three representative CTTA-OD benchmark datasets: Cityscapes → Cityscapes-C, Cityscapes → ACDC, and SHIFT → SHIFT-C. The results are shown in Tables 1, 2, and 3. Our method achieves the best average performance across all three datasets, outperforming the baseline by 9.0%, 2.0%, and 2.6%, and surpassing the second-best method by 1.7%, 1.0%, and 1.1%, respectively. It is important to note that in these three tables, we follow the common CTTA-OD evaluation protocol, which employs an online cumulative computation strategy: the model dynamically updates, predicting results only for the current batch at each time step, and all predictions across time steps are concatenated to determine the final performance on the target domain. To more fairly assess the model's adaptation speed to distribution shifts, we evaluate its overall performance on the current target domain after each optimization step and plot the performance variation curve, as shown in Figure 3. The curve demonstrates that our method achieves the fastest performance improvement when the target domain changes, indicated by the steepest upward slope, and maintains stable performance after the rapid increase, validating its fast learning capability. This rapid learning ability is also a key reason why our method achieves superior performance under the online cumulative evaluation metric, as it enables better early-stage adaptation in the current domain, thereby improving overall performance.

4.3 Ablation Study

Effectiveness of different components. As shown in Table 4, we conduct ablation studies to verify the effectiveness of the two core components in our framework: the frequency-specific adapter and the frequency distribution alignment loss. To demonstrate the advantage of performing adaptation in the frequency domain, we introduce a spatial adapter [8] as a comparison baseline. Unlike the frequency domain, where information is naturally decomposed and compactly represented, the spatial domain requiring optimization over all spatial dimensions. As a result, the spatial adapter suffers from high optimization dimensionality, leading to slower convergence and inferior performance. We further evaluate the contribution of the frequency distribution alignment loss. Both patch-level alignment loss $\mathscr{L}_{\text{patch}}$ and instance-level alignment loss $\mathscr{L}_{\text{ins}}$ bring performance improvements when applied individually or jointly. Moreover, we assess the effect of patch-wise computation by comparing with a global average pooling variant $\mathscr{L}_{\text{img}}$. Experimental results confirm that patch-wise computation is more effective, as it matches the patch-wise selection mechanism in the adapter, ensures consistent receptive fields.

Partition factor n and Top-k ablation. We analyze the influence of different partition factors and Top-k values on model performance, revealing a trade-off relationship, as shown in Table 5. Without partitioning the frequency spectrum (i.e., $n = 1$) and learning a single global frequency expert, the optimization dimensionality remains similar to spatial adaptation, failing to leverage the decomposition and compression benefits of frequency domain learning. Conversely, when n increases (e.g., $n = 4$), the spectrum is more finely decomposed,

Table 1: Comparisons on Cityscapes → Cityscapes-C. We report mAP@50 for detection performance. 'Mean' denotes the average CTTA-OD performance across domains, while 'Gain' indicates the improvement over the source detector.

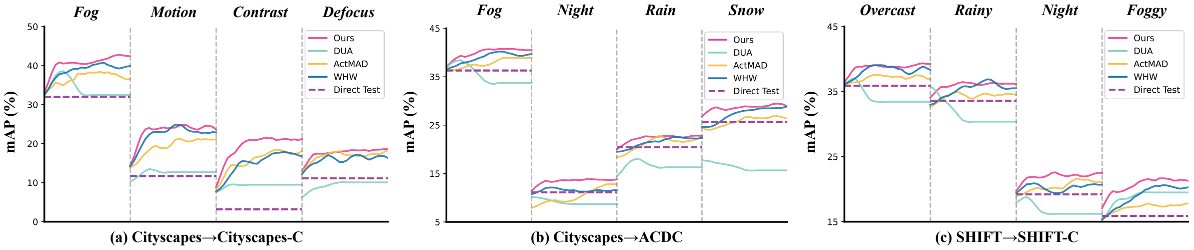
Round	1				5				10				Mean↑	Gain
Condition	Fog	Motion	Contrast	Defocus	Fog	Motion	Contrast	Defocus	Fog	Motion	Contrast	Defocus		
Direct test	32.0	11.7	3.2	11.1	32.0	11.7	3.2	11.1	32.0	11.7	3.2	11.1	14.5	/
DUA [16]	32.9	12.5	9.7	9.2	31.3	9.4	13.5	8.1	30.6	9.1	13.5	7.9	15.8	+1.3
CoTTA [48]	34.0	13.5	10.5	13.0	33.5	12.8	11.2	12.5	33.0	13.2	12.0	12.8	17.0	+2.5
STFAR [6]	35.7	15.9	12.4	14.2	34.7	14.1	14.4	15.2	34.0	15.0	16.4	15.1	19.9	+5.4
MemCLR [5]	36.3	14.8	16.5	16.0	35.3	15.8	17.1	15.1	36.0	15.2	16.5	15.5	20.9	+6.4
ActMAD [7]	36.0	18.0	16.4	17.5	35.1	19.4	15.2	15.5	32.1	21.1	17.1	14.5	21.4	+6.9
MLFA [19]	37.2	16.6	17.1	15.4	36.2	16.5	16.4	15.2	37.0	16.7	16.1	16.0	21.5	+7.0
ViDA [49]	36.5	17.2	15.2	16.2	35.8	16.5	15.0	15.0	36.2	17.0	15.8	15.5	20.4	+5.9
WHW [8]	37.1	18.5	15.8	16.8	36.7	18.4	15.3	15.8	37.1	17.9	16.7	15.2	21.8	+7.3
Ours	38.4	19.2	18.4	17.6	38.8	20.9	17.6	18.0	38.3	20.1	18.0	17.0	23.5	+9.0

Table 2: Comparisons on Cityscapes → ACDC. We report mAP@50 for CTTA-OD performance, where 'Mean' represents the average performance over across domains, and 'Gain' denotes the improvement over the source detector.

Round	1				5				10				Mean↑	Gain
Condition	Fog	Night	Rain	Snow	Fog	Night	Rain	Snow	Fog	Night	Rain	Snow		
Direct test	36.3	11.1	20.4	25.7	36.3	11.1	20.4	25.7	36.3	11.1	20.4	25.7	23.4	/
DUA [16]	34.4	9.6	16.8	15.7	32.7	8.8	15.0	15.6	31.4	9.8	15.9	14.6	18.4	-5.2
CoTTA [48]	35.2	10.0	19.0	25.0	34.8	9.6	18.5	24.5	34.5	10.2	18.8	25.2	21.2	-2.2
STFAR [6]	36.9	10.7	20.3	26.3	37.0	10.9	21.3	26.1	36.5	10.7	20.6	26.8	23.7	+0.4
MemCLR [5]	36.5	10.4	20.3	26.0	37.5	11.4	20.6	25.7	36.4	10.5	20.1	26.3	23.5	+0.2
ActMAD [7]	37.4	10.6	20.7	25.9	37.8	11.3	20.1	27.8	37.6	10.5	20.4	26.1	23.9	+0.6
MLFA [19]	37.9	11.8	20.0	27.6	38.0	12.0	20.1	27.9	37.5	11.8	20.5	26.6	24.3	+0.9
ViDA [49]	38.0	11.5	20.5	27.0	37.2	11.0	20.0	26.5	37.8	11.4	20.3	27.2	24.1	+0.7
WHW [8]	38.8	11.3	20.7	26.3	38.3	11.5	21.0	27.3	37.5	11.4	20.0	27.6	24.4	+1.0
Ours	40.1	12.8	21.1	27.9	39.0	12.5	21.8	29.0	38.6	12.0	20.7	29.0	25.4	+2.0

Table 3: Comparisons on SHIFT → SHIFT-C. We report mAP@50 for CTTA-OD performance, where 'Mean' represents the average performance over across domains, and 'Gain' denotes the improvement over the source detector.

Round	1				5				10				Mean↑	Gain
Condition	Overcast	Rainy	Night	Foggy	Overcast	Rainy	Night	Foggy	Overcast	Rainy	Night	Foggy		
Direct test	35.9	33.6	19.2	15.9	35.9	33.6	19.2	15.9	35.9	33.6	19.2	15.9	26.2	/
DUA [16]	33.6	30.4	16.4	19.0	32.1	27.8	16.1	17.5	31.5	28.6	16.1	17.5	24.0	-2.2
CoTTA [48]	36.2	33.0	18.5	17.8	35.8	32.6	18.2	17.5	35.6	32.8	18.7	17.6	25.2	-1.0
STFAR [6]	35.9	33.5	19.5	16.7	35.8	34.5	19.8	17.7	36.0	33.1	19.7	17.7	26.7	+0.5
MemCLR [5]	36.8	34.6	20.7	16.5	36.7	35.6	18.7	16.0	36.0	34.2	19.9	16.9	27.0	+0.8
ActMAD [7]	37.0	34.8	20.1	17.0	36.5	35.5	20.4	18.0	35.7	34.2	20.3	18.3	27.4	+1.2
MLFA [19]	37.2	34.6	20.2	19.2	36.9	34.0	21.5	18.9	36.0	33.9	20.5	18.0	27.7	+1.5
ViDA [49]	37.4	35.0	19.8	18.6	36.8	34.2	19.5	18.1	37.0	34.4	19.7	18.4	26.9	+0.7
WHW [8]	38.0	35.7	19.6	18.7	36.1	35.1	19.7	18.2	36.3	34.1	19.6	18.9	27.7	+1.5
Ours	38.6	35.8	21.8	20.5	37.5	36.1	22.6	19.8	36.7	34.9	21.6	19.4	28.8	+2.6



(a) Cityscapes→Cityscapes-C (b) Cityscapes→ACDC (c) SHIFT→SHIFT-C

Figure 3: To illustrate adaptation speed, we plot performance curves over adaptation steps on three benchmarks. The curves are smoothed for better visualization.

but more experts are required to cover sufficient bands. This leads to optimization instability and degraded performance under limited data. For a fixed partition factor (e.g., $n = 2$), choosing the Top-k value reflects a balance between two competing goals: reducing the optimization space by selecting fewer frequency bands and ensuring sufficient coverage for capturing domain shifts. From a signal analysis perspective, natural perturbations (e.g., fog, blur, noise) typically affect a localized subset of frequency bands rather than the entire spectrum. Selecting too few bands (e.g., 25%) may miss key affected regions, while selecting too many (e.g., 100%) introduces redundancy and slows down adaptation. Our empirical results suggest that selecting 50% of the bands captures most of the domain shift while preserving fast convergence, aligning with the underlying frequency-localized nature of visual corruptions. This observation provides a principled basis for choosing $n = 2$ and Top-k=50% as a balanced configuration.

Hyper-parameter λ ablation. We perform an ablation study on the loss balancing hyper-parameter λ, as shown in Figure 4. While the patch-level alignment loss $\mathscr{L}_{\text{patch}}$ operates on global frequency statistics, it lacks semantic granularity. To complement this, the instance-level alignment loss $\mathscr{L}_{\text{ins}}$ enforces category-specific frequency consistency, which is particularly important for object detection tasks that depend heavily on instance-wise feature quality. Therefore, incorporating $\mathscr{L}_{\text{ins}}$ is essential for capturing fine-grained shifts across classes. However, a key difference is that $\mathscr{L}_{\text{ins}}$ requires pseudo-labels to compute class-wise statistics in the target domain, which could be noisy, especially during early adaptation stages. In contrast, $\mathscr{L}_{\text{patch}}$ is inherently more reliable. Assigning an excessively large weight to $\mathscr{L}_{\text{ins}}$ (*i.e.*, a high λ) risks overfitting to incorrect pseudo-labels and destabilizing the optimization. Our empirical results show that $\lambda = 3$ strikes a good balance: it allows the model to benefit from class-level alignment while preserving training stability.

Table 4: Ablation analysis on the framework components.

Base Detector	Adapter Type	$\mathscr{L}_{\text{patch}}$	$\mathscr{L}_{\text{ins}}$	Mean↑
✓	/	/	/	14.5
✓	Spatial [8]	✓	✓	22.5
✓	Frequency (Ours)	$\mathscr{L}_{\text{img}}$		22.1
✓	Frequency (Ours)	✓		22.4
✓	Frequency (Ours)		✓	23.0
✓	Frequency (Ours)	✓	✓	23.5

Table 5: Ablation on the partition factor n and top-k (%).

Partition Factor n	1	2	2	2	2	4	4	4
Top-k selection	100%	25%	50%	75%	100%	12.5%	25%	50%
Mean↑	22.0	22.6	23.5	21.9	20.9	23.0	22.5	19.1

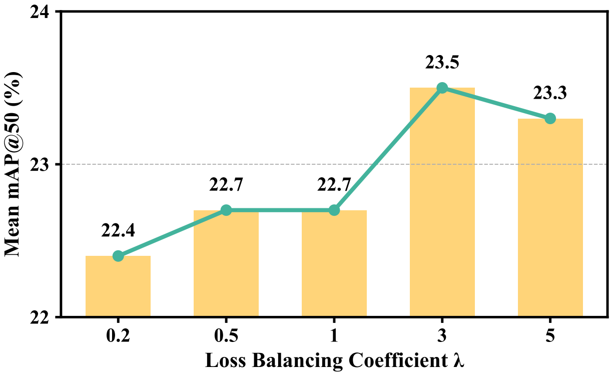
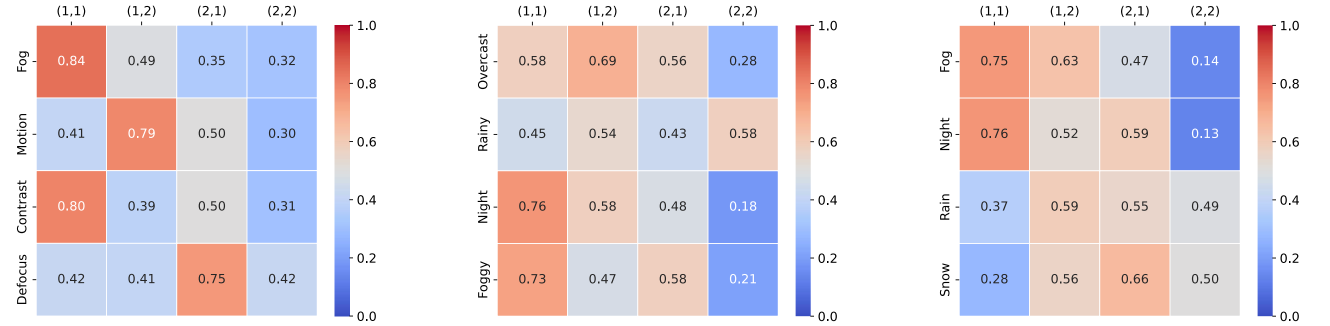


Figure 4: Ablation analysis on the hyper-parameter λ.



(a) Cityscapes→Cityscapes-C **(b) Cityscapes→ACDC** **(c) SHIFT→SHIFT-C**

Figure 5: We visualize the selection probabilities of different frequency bands during adaptation across three benchmarks. In the heatmap, the coordinates represent (row, column) indices of patches.

Design choice of framework and loss. We conduct ablation studies on the expert structure and loss function, as shown in Table 6. For expert structures, we compare $\text{Conv}_{3\times3}$, $\text{Conv}_{1\times1}$, and element-wise learnable tensor. Due to the global nature of the Fourier transform, positions in the frequency domain independently correspond to specific frequency components, unlike spatial pixels that exhibit local correlations. Thus, $\text{Conv}_{1\times1}$, which enables cross-channel interactions without introducing redundant spatial operations, proves to be the more suitable choice. In contrast, element-wise learnable tensors only perform channel-wise scaling without capturing cross-channel dependencies, leading to inferior performance. For the loss function, we evaluate L1, MSE, and KL divergence (following [8], which computes the KL divergence between two Gaussian distributions) as alignment losses. Since the magnitude of frequency-domain signals directly determines the contribution of different frequency components, the loss function should prioritize magnitude alignment. KL divergence, which primarily focuses on probability matching, is less effective for frequency-domain adaptation. In contrast, MSE directly minimizes magnitude discrepancies [50], resulting in better adaptation performance.

Table 6: Ablation analysis on the design choice.

Expert Structure			Loss Design			Mean↑
$\text{Conv}_{3\times3}$	$\text{Conv}_{1\times1}$	Learnable Tensor	L1	MSE	KL Div	
✓				✓		22.7
	✓			✓		23.5
		✓		✓		23.1
	✓		✓			23.0
	✓			✓		23.5
	✓				✓	20.1

Frequency bands selection visualization. We analyze the selection probabilities of different bands during adaptation by tracking the top-k selections across target domains, as shown in Figure 5. The results reveal distinct patterns: for low-frequency perturbations like fog and night, the model prioritizes low-frequency bands, while for high-frequency perturbations such as motion blur and rain, the selection shifts toward mid-to-high frequencies. However, as most signal energy is concentrated in the low-frequency range, adapting to these frequencies yields larger gradients. Thus, regardless of the perturbation type, the model consistently maintains a certain probability of selecting low-frequency bands.

5 Conclusion and Future Work

Conclusion. In this paper, we propose a novel frequency-specific adaptation framework for CTTA-OD, enabling fast online adaptation in dynamic environments. By selectively adapting the most affected frequency bands instead of optimizing in the high-dimensional spatial domain, our method accelerates adaptation process and maintain stable detection performance. Extensive experiments validate our efficacy.
Future work. One promising direction is to develop a more adaptive frequency spectrum partitioning strategy, as the current fixed scheme may not generalize well across varying target domains. Additionally, integrating spatial-frequency hybrid adaptation could further enhance performance by leveraging the global properties of frequency information while preserving spatial details for finer-grained adaptation.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions

Conceptualization, Kunyu Wang and Zhai Wei; methodology, Kunyu Wang; software, Kunyu Wang; validation, Kunyu Wang and Qi Qi; formal analysis, Kunyu Wang; investigation, Kunyu Wang; resources, Zhai Wei; data curation, Kunyu Wang; writing—original draft preparation, Kunyu Wang; writing—review and editing, Qi Qi and Zhai Wei; visualization, Kunyu Wang and Qi Qi; supervision, Zhai Wei; project administration, Zhai Wei. All authors have read and agreed to the published version of the manuscript.

Conflict of Interest

All the authors declare that they have no conflict of interest.

References

[1] Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J: Object detection in 20 years: A survey. Proceedings of the IEEE, **111**(3), 257–276 (2023). https://doi.org/10.1109/JPROC.2023.3238524

[2] Zhao, Z., Zheng, P., Xu, S.T., Wu, X: Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems, **30**(11), 3212–3232 (2019). https://doi.org/10.1109/TNNLS.2018.2876865

[3] Cui, Y., Huang, S., Zhong, J., Liu, Z., Wang, Y., Sun, C., Li, B., Wang, X., Khajepour, A.: Drivellm: Charting the path toward full autonomous driving with large language models. IEEE Transactions on Intelligent Vehicles, **9**(1), 1450–1464 (2023). https://doi.org/10.1109/TIV.2023.3327715

[4] Zhang, J., Wang, K., Xu, R., Zhou, G., Hong, Y., Fang, X., Wu, Q., Zhang, Z., Wang, H.: Navid: Video-based vlm plans the next step for vision-and-language navigation. arXiv preprint arXiv:2402.15852 (2024). https://doi.org/10.48550/arXiv.2402.15852

[5] Geiger, A., Lenz, P., Urtasun, R.: Towards online domain adaptive object detection. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 478–488 (2023). https://doi.org/10.1109/WACV56688.2023.00055

[6] Chen, Y., Xu, X., Su, Y., Jia, K.: Stfar: Improving object detection robustness at test-time by self-training with feature alignment regularization. arXiv preprint arXiv:2303.17937 (2023). https://doi.org/10.48550/arXiv.2303.17937

[7] Mirza, M.J., Soneira, P.J., Lin, W., Kozinski, M., Possegger, H., Bischof, H.: ActMAD: Activation Matching to Align Distributions for Test-Time-Training. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 24152–24161 (2023). https://doi.org/10.1109/CVPR52729.2023.02313

[8] Yoo, J., Lee, D., Chung, I., Kim, D., Kwak, N.: What how and when should object detectors update in continually changing test domains? In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 23354–23363 (2024). https://doi.org/10.1109/CVPR52733.2024.02204

[9] Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019). https://doi.org/10.48550/arXiv.1903.12261

[10] Sakaridis, C., Dai, D., Van Gool, L.: ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 10765–10775 (2021). https://doi.org/10.1109/ICCV48922.2021.01059

[11] Sun, T., Segu, M., Postels, J., Wang, Y., Van Gool, L., Schiele, B., Tombari, F., Yu, F.: SHIFT: A synthetic driving dataset for continuous multi-task domain adaptation. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 21371–21382 (2022). https://doi.org/10.1109/CVPR52688.2022.02068

[12] Bottou, L., Bousquet, O.: The tradeoffs of large scale learning. In Proceedings of the 21st International Conference on Neural Information Processing Systems, pp. 161–168 (2007)

[13] Ge, R., Huang, F., Jin, C., Yuan, Y.: Escaping from saddle points—online stochastic gradient for tensor decomposition. In Conference on learning theory, pp. 797–842 (2015).

[14] Nussbaumer, H.J.: The fast Fourier transform in Fast Fourier Transform and Convolution Algorithms, pp. 80-111, Springer, Berlin (1981). https://doi.org/10.1007/978-3-642-81897-4_4

[15] Jiang, X., Zhang, X., Gao, N., Deng, Y.: When fast fourier transform meets transformer for image restoration. In European conference on computer vision, pp. 381–402 (2024). https://doi.org/10.1007/978-3-031-72995-9_22

[16] Mirza, M.J., Micorek, J., Possegger, H., Bischof, H.: The norm must go on: Dynamic unsupervised domain adaptation by normalization. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14765–14775 (2022). https://doi.org/10.1109/CVPR52688.2022.01435

[17] Cao, S., Zheng, J., Liu, Y., Zhao, B., Yuan, Z., Li, W., Dong, R., Fu, H.: Exploring test-time adaptation for object detection in continually changing environments. arXiv preprint arXiv:2406.16439 (2024). https://doi.org/10.48550/arXiv.2406.16439

[18] Sinha, S., Gehler, P., Locatello, F., Schiele, B.: Test: Test-time self-training under distribution shift. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 2759–2769 (2023). https://doi.org/10.1109/WACV56688.2023.00278

[19] Liu, Y., Wang, J., Huang, C., Wu, Y., Xu, Y., Cao, X.: MLFA: Towards realistic test time adaptive object detection by multi-level feature alignment. IEEE Transactions on Image Processing, **33**, 5837–5848 (2024). https://doi.org/10.1109/TIP.2024.3473532

[20] Pitas, I.: Digital image processing algorithms and applications. John Wiley & Sons, New York (2000).

[21] Mevenkamp, N., Berkels, B.: Variational multi-phase segmentation using high-dimensional local features. In 2016 IEEE winter conference on applications of computer vision, pp. 1–9 (2016). https://doi.org/10.1109/WACV.2016.7477729

[22] Chi, L., Jiang, B., Mu, Y.: Fast fourier convolution. In Advances in Neural Information Processing Systems 33

(NeurIPS 2020), pp. 1–10 (2020).

[23] Buchholz, T.O., Jug, F.: Fourier image transformer. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1846–1854 (2022). https://doi.org/10.1109/CVPRW56347.2022.00201

[24] Nguyen, T., Pham, M., Nguyen, T., Nguyen, K., Osher, S., Ho, N.: Fourierformer: Transformer meets generalized fourier integral theorem. In Proceedings of the 36th International Conference on Neural Information Processing Systems, pp. 29319–29335 (2020)

[25] Pratt, H., Williams, B., Coenen, F., Zheng, Y.: FCNN: Fourier convolutional neural networks. In Joint European conference on machine learning and knowledge discovery in databases, pp. 786–798 (2017). https://doi.org/10.1007/978-3-319-71249-9_47

[26] Guibas, J., Mardani, M., Li, Z., Tao, A., Anandkumar, A., Catanzaro, B.: Adaptive fourier neural operators: Efficient token mixers for transformers. arXiv preprint arXiv:2111.13587 (2021). https://doi.org/10.48550/arXiv.2111.13587

[27] Huang, J., Guan, D., Xiao, A., Lu, S.: Fsdr: Frequency space domain randomization for domain generalization. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6891–6902 (2021). https://doi.org/10.1109/CVPR46437.2021.00682

[28] Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P. A.: FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1013–1023 (2021). https://doi.org/10.1109/CVPR46437.2021.00107

[29] Xu, Q., Zhang, R., Zhang, Y., Wang, Y., Tian, Q.: A fourier-based framework for domain generalization. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14383–14392 (2021). https://doi.org/10.1109/CVPR46437.2021.01415

[30] Yang, Y., Soatto, S.: FDA: Fourier domain adaptation for semantic segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4085–4095 (2020). https://doi.org/10.1109/CVPR42600.2020.00414

[31] Wang, H., Wu, X., Huang, Z., Xing, E.P.: High-frequency component helps explain the generalization of convolutional neural networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8684–8694 (2020). https://doi.org/10.1109/CVPR42600.2020.00871

[32] Xu, Z.J.: Understanding training and generalization in deep learning by fourier analysis. arXiv preprint arXiv:1808.04295 (2018). https://doi.org/10.48550/arXiv.1808.04295

[33] Yin, D., Gontijo Lopes, R., Shlens, J., Cubuk, E.D., Gilmer, J.: A fourier perspective on model robustness in computer vision. In Proceedings of the 33rd International Conference on Neural Information Processing Systems , pp. 13276–13286 (2019).

[34] Fedus, W., Dean, J., Zoph, B.: A review of sparse expert models in deep learning. arXiv preprint arXiv:2209.01667 (2022). https://doi.org/10.48550/arXiv.2209.01667

[35] Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural Computation, **3**(1), 79–87 (1991). https://doi.org/10.1162/neco.1991.3.1.79

[36] Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.: Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668 (2020). https://doi.org/10.48550/arXiv.2006.16668

[37] Aljundi, R., Chakravarty, P., Tuytelaars, T.: Expert gate: Lifelong learning with a network of experts. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3366–3375 (2017). https://doi.org/10.1109/CVPR.2017.753

[38] Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017). https://doi.org/10.48550/arXiv.1701.06538

[39] Eigen, D., Ranzato, M. A., Sutskever, I.: Learning factored representations in a deep mixture of experts. arXiv preprint arXiv:1312.4314 (2013). https://doi.org/10.48550/arXiv.1312.4314

[40] DeepSeek-AI, Liu, A., Feng, B., Wang, B., Wang, B., Liu, B., Zhao, C., Dengr, C., Ruan, C., Dai, D., et al.: Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. arXiv preprint arXiv:2405.04434 (2024). https://doi.org/10.48550/arXiv.2405.04434

[41] Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., de las Casas, D., Hanna, E.B., Bressand, F., et al.: Mixtral of experts. arXiv preprint arXiv:2401.04088 (2024). https://doi.org/10.48550/arXiv.2401.04088

[42] Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research, **23**(1), 5232–5270, (2022). https://dl.acm.org/

doi/abs/10.5555/3586589.3586709

[43] Ahmed, N., Natarajan, T., Rao, K.R.: Discrete cosine transform. IEEE Transactions on Computers, **100**(1), 90–93 (2006). https://doi.org/10.1109/T-C.1974.223784

[44] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, **39**(6), 1137–1149 (2016). https://doi.org/10.1109/TPAMI.2016.2577031

[45] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3213–3223 (2016). https://doi.org/10.1109/CVPR.2016.350

[46] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

[47] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848

[48] Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 7201–7211 (2022). https://doi.org/10.1109/CVPR52688.2022.00706

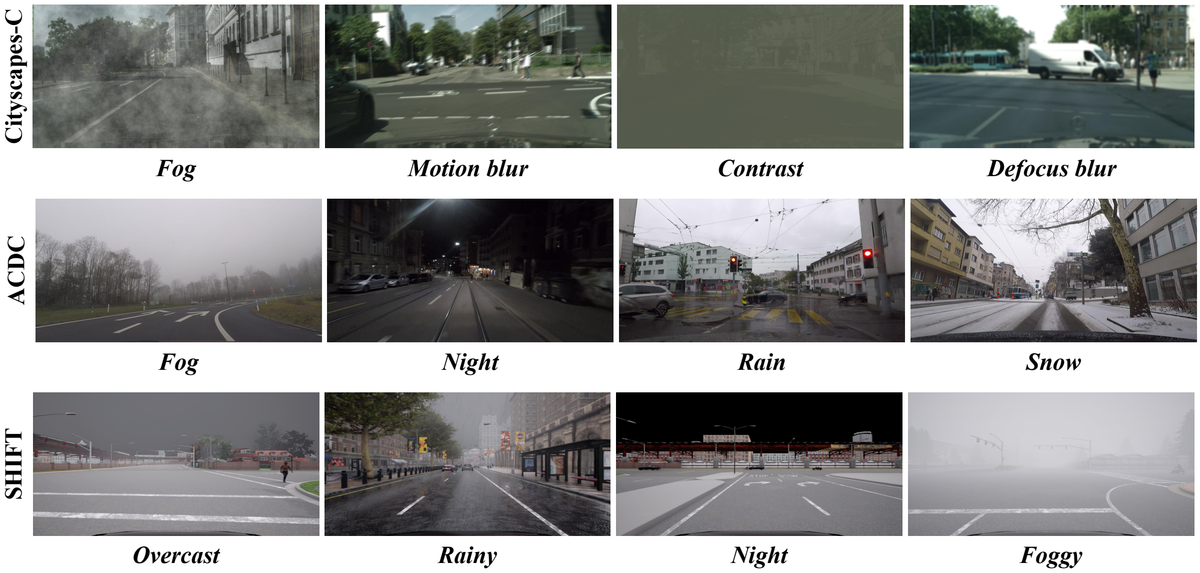
[49] Liu, J., Yang, S., Jia, P., Zhang, R., Lu, M., Guo, Y., Xue, W., Zhang, S: Vida: Homeostatic visual domain adapter for continual test time adaptation. arXiv preprint arXiv:2306.04344 (2023). https://doi.org/10.48550/arXiv.2306.04344

[50] Yan, C.W., Foo, S.Q., Trinh, V.H., Yeung, D.Y., Wong, K.H., Wong, W.K.: Fourier amplitude and correlation loss: Beyond using L2 loss for skillful precipitation nowcasting. In Proceedings of the 38th International Conference on Neural Information Processing Systems, pp. 100007–100041 (2024).

Appendix

A Target Domain Visualization

We conduct extensive experiments on Cityscapes-C, ACDC, and SHIFT. For Cityscapes-C, we use the maximum severity level 5 across four perturbations to evaluate our method under extreme conditions. ACDC and SHIFT are designed to simulate real-world autonomous driving scenarios, with ACDC collected from real-world environments and SHIFT generated from a simulator. Both datasets are specifically tailored for dense prediction tasks, providing a challenging benchmark for evaluating adaptation performance. The diverse evaluations validate our method's effectiveness under both extreme conditions and real-world scenarios. The visualizations are presented in Figure A1.



Fog *Motion blur* *Contrast* *Defocus blur*

Fog *Night* *Rain* *Snow*

Overcast *Rainy* *Night* *Foggy*

Figure A1: Examples of different target domains across three benchmarks.