FreeKD: Knowledge Distillation via Semantic Frequency Prompt

Yuan Zhang[1], Tao Huang[2], Jiaming Liu[1], Tao Jiang[3], Kuan Cheng[1], Shanghang Zhang[1†]
[1] National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University [2]The University of Sydney [3]Zhejiang University

Abstract

*Knowledge distillation (KD) has been applied to various tasks successfully, and mainstream methods typically boost the student model via spatial imitation losses. However, the consecutive downsamplings induced in the spatial domain of teacher model is a type of corruption, hindering the student from analyzing what specific information needs to be imitated, which results in accuracy degradation. To better understand the underlying pattern of corrupted feature maps, we shift our attention to the frequency domain. During frequency distillation, we encounter a new challenge: the low-frequency bands convey general but minimal context, while the high are more informative but also introduce noise. Not each pixel within the frequency bands contributes equally to the performance. To address the above problem: (1) We propose the Frequency Prompt plugged into the teacher model, absorbing the semantic frequency context during finetuning. (2) During the distillation period, a pixel-wise frequency mask is generated via Frequency Prompt, to localize those pixel of interests (PoIs) in various frequency bands. Additionally, we employ a position-aware relational frequency loss for dense prediction tasks, delivering a high-order spatial enhancement to the student model. We dub our **Fre**quency **K**nowledge **D**istillation method as **FreeKD**, which determines the optimal localization and extent for the frequency distillation. Extensive experiments demonstrate that FreeKD not only outperforms spatial-based distillation methods consistently on dense prediction tasks (e.g., FreeKD brings **3.8** AP gains for RepPoints-R50 on COCO2017 and **4.55** mIoU gains for PSPNet-R18 on Cityscapes), but also conveys more robustness to the student. Notably, we also validate the generalization of our approach on large-scale vision models (e.g., DINO and SAM).*

1. Introduction

In the quest for significant advancements, recent deep learning models have witnessed a substantial increase in both depth and width, as exemplified by notable works such as [16, 23, 28]. However, pursuing larger and more powerful models results in unwieldy and inefficient deployments on resource-limited edge devices. To address this dilemma, knowledge distillation (KD) [8, 11, 17, 20, 50] has emerged as a promising solution to transfer the knowledge encapsulated within a heavy model (teacher) to a more compact, pocket-size model (student).

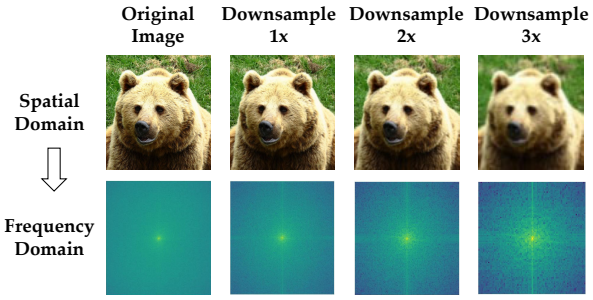

Figure 1. **Comparison of the presentation of the bear at different downsampling ratios on spatial and frequency domain.**

Among diverse computer vision tasks, the transfer of dark knowledge for dense prediction tasks poses unique challenges, particularly requiring fine-grained distillation at the feature level. Recent distillation methods have aimed to enhance performance through spatial-level distillation losses, refining valuable information within the features. However, the sequential downsampling applied in the spatial domain of the teacher model introduces a form of corruption. This corruption hampers the student's ability to discern specific information that should be mimicked, resulting in a decline in accuracy.

As illustrated in Figure 1, downsampling operations prominently remove high-frequency image details in the frequency domain, revealing underlying patterns not easily discernible from raw pixel values [3, 45, 47]. This observation prompts us to explore the potential of leveraging frequency signals for knowledge distillation. However, directly employing this approach raises two significant challenges: **(a)** The low-frequency bands from the teacher model convey general yet minimal contextual information,

[†]Corresponding author. Shanghang Zhang is supported by the National Key Research and Development Project of China (No. 2022ZD0117801).

characterized by smooth variations [44, 58]. If the student is forced to imitate all pixels of low-frequency bands directly, it tends to focus on easy but less informative samples, aiming to reduce loss. **(b)** The high-frequency range provides more fine-grained and distinctive signals, with salient transitions enhancing the student's robustness and generalizability [54]. However, when the student mimics high-frequency pixels, it also captures noise, leading to undesired degradation. Therefore, the challenge lies in localizing worthy pixels of interest (PoIs) in both frequency bands.

To address these challenges, we introduce the semantic Frequency Prompt as depicted in Figure 2 (c). Initially, a set of Points of Interest (PoIs) masks is generated by encoding similarities between prompts and frequency bands. Subsequently, the masked frequency bands, rather than the vanilla ones, are supervised by task loss. This approach provides precise guidance for the student in reconstructing the teacher's frequency bands — a crucial aspect of knowledge distillation. Importantly, the Frequency Prompt differs from previous spatial prompts in both insertion method and the transferred substance. In Figure 2, Prompt Tokens (VPTs) [21, 57] are inserted as tokens for transformer series tasks, while Contrastive Texture Attention Prompts (CTAP) [13] are summed point by point on the input image, avoiding occlusion. In contrast, the localization of our Frequency Prompts is flexible, depending on where the student intends to imitate. This involves incorporating a position-aware relational frequency loss, where positional channel-wise weights are derived from cross-layer information. These weights act as an adaptive gating operation, selectively choosing relevant channels from frequency bands.

With the above key designs, we propose a **Fre**quency **K**nowledge **D**istillation pipeline called **FreeKD**, where the student is under fine-grained frequency imitation principle. Extensive experimental results show that our method surpasses current state-of-the-art spatial-based methods consistently in standard settings of object detection and semantic segmentation tasks. For instance, FreeKD obtains 42.4 AP with RepPoints-R50 student on the COCO dataset, surpassing DiffKD [20] by 0.7 AP; while on semantic segmentation, FreeKD outperforms MGD [51] by 0.8% with PSPNet-R18 student on Cityscapes test set. Moreover, we implement FreeKD on large-scale vision model settings, and our method significantly outperforms the baseline method. Finally, we are surprised that the student distilled by FreeKD exhibits better domain generalization capabilities (e.g., FreeKD outperforms DiffKD by 1.0% rPC [39]).

In a nutshell, the contributions of this paper are threefold:

1. We introduce a novel knowledge distillation manner (FreeKD) from the frequency domain, and make the first attempt to explore its potential for distillation on dense prediction tasks, which breaks the bottleneck of spatial-based methods.

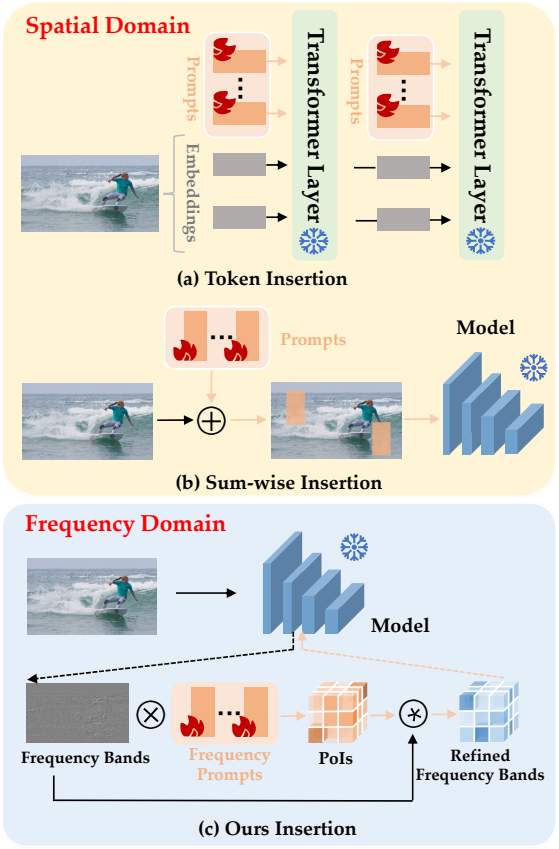

Figure 2. **Comparisons with other insertion methods of spatial prompts.** (a) Prompts are inserted into the encoder layer as tokens. (b) Sum-wise on RGB channels of input image. (c) Ours interact with intermediate features. Best view in color.

2. To the best of our knowledge, we are the first to propose Frequency Prompt especially for frequency knowledge distillation, eliminating unfavorable information from frequency bands, and a position-aware relational frequency loss for dense prediction enhancement.
3. We validate the effectiveness of our method through extensive experiments on various benchmarks, including large-scale vision model settings. Our approach consistently outperforms existing spatial-based methods, demonstrating significant improvements and enhanced robustness in students distilled by FreeKD.

2. Related Work

2.1. KD on Dense Prediction Tasks

In recent years, knowledge distillation for dense prediction tasks such as object detection and semantic segmentation has garnered significant attention, owing to its practical applications and the inherent challenges of distilling fine-grained pixel-level recognition and localization features. Early approaches [2, 24] primarily concentrated on

distilling classification and regression outputs or intermediate features using traditional loss functions such as Kullback–Leibler divergence and mean square error. However, recent research has shifted its focus towards mimicking valuable information while filtering out noisy features in the intermediate dense representations. This shift is driven by the observation that dense features often contain redundant information, which can burden the student model. To address this, contemporary works employ techniques like generating pixel-level masks based on ground-truth boxes [14, 40], leveraging feature attentions [37, 49], and introducing learnable mask tokens [19] for feature refinement. Besides, some approaches propose to reducing the representation gap between teacher and student via normalizing the features with Pearson correlation [1] or denoising the features with diffusion models [20]. However, the consecutive downsamplings induced in the spatial domain of the teacher model is a type of corruption, hindering the student from analyzing what specific information needs to be imitated, which results in accuracy degradation. To better understand the underlying pattern of corrupted feature maps, we shift our attention to the frequency domain.

2.2. Frequency Analysis Methods

Frequency domain analysis has found extensive application in various computer vision tasks, including image classification [42, 47], image generation [22], and image super-resolution [32]. Early studies [15, 31, 33] indicate that in the frequency domain, the phase component predominantly captures high-level semantics of the original signals, while the amplitude component retains low-level statistics. Consequently, underlying image patterns are more conveniently observed in the frequency representation compared to raw pixel values in the spatial domain. In this context, wavelet analysis stands out as a particularly effective method in image processing [12, 29, 54], as it can capture multiscale frequency domain information in a compact representation. Unlike other frequency analysis methods like Fourier analysis, wavelet analysis offers a more comprehensive perspective. Leveraging wavelet analysis, our method is tailored for dense prediction tasks, demonstrating superior distillation on image patterns when compared to distilling raw pixel values in the spatial domain.

3. Proposed Approach: FreeKD

In this section, we first demonstrate vanilla knowledge distillation via frequency loss. To further provide more precise PoIs, we design a novel Frequency Prompt to generate pixel imitation principles. Finally, we propose a position-aware relational loss to enhance the sensitivity to dense prediction. The architecture of FreeKD is illustrated in Figure 3.

3.1. Distillation with Frequency

Dilations and translations of the Mother function $\Phi(t)$, define an orthogonal wavelet basis:

$$\Phi_{(s,d)}(t) = 2^{\frac{s}{2}}\Phi(2^s t - d), \quad s, d \in \boldsymbol{Z} \tag{1}$$

where $\boldsymbol{Z}$ is the set of all integers and the factor $\frac{s}{2}$ maintains a constant norm independent of scale s. The variables s and d, scales and dilates the mother function Φ to generate wavelets in $\mathcal{L}_2$ spaces. To create frequency representations, Discrete Wavelet Transformation (DWT) ξ is applied for frequency bands decomposition via Φ to each channel as follows:

$$\mathcal{B}_l = \xi(x), \tag{2}$$

where l is the decomposition level. When the level is set to 1, the feature map $\boldsymbol{F} \in \mathbb{R}^{C\times H\times W}$ can be split into four bands, and $\mathcal{B}_1 = \{\texttt{LL}, \texttt{HL}, \texttt{LH}, \texttt{HH}\}$, where $\texttt{LL}$ indicates the low-frequency band ($\boldsymbol{R}_{\texttt{LL}} \in \mathbb{R}^{C\times H_{\texttt{LL}}\times W_{\texttt{LL}}}$ represents its corresponding tensor), and the others are high-frequency bands. When l is 2, the $\texttt{LL}$ band can be further decomposed into $\texttt{LL2}$, $\texttt{HL2}$, $\texttt{LH2}$ and $\texttt{HH2}$. In this paper, we set $l = 3$ for all distillation experiments.

In order to learn dark knowledge of the teacher, one typical manner is to mimic the tensor pixel-wisely. Regularly, $\boldsymbol{F}^{(t)} \in \mathbb{R}^{C\times H\times W}$ and $\boldsymbol{F}^{(s)} \in \mathbb{R}^{C_s\times H\times W}$ denote the feature maps of teacher and student networks respectively, and the frequency bands imitation can be fulfilled via:

$$\mathcal{L}_{\text{FKD}} = \sum_{k=1}^{L} \left\| a_k - b_k \right\|_1,$$

$$a_k \in \xi(\boldsymbol{F}^{(t)}), b_k \in \xi(\phi(\boldsymbol{F}^{(s)})), \tag{3}$$

where L is the number of frequency bands, and ϕ is a linear projection layer to adapt $\boldsymbol{F}^{(s)}$ to the same resolution as $\boldsymbol{F}^{(t)}$. The student model studies general laws via low-frequency imitation, and salient pattern (including fine textures, edges, and *noise*) from the high-frequency.

3.2. Semantic Frequency Prompt

Therefore, we introduce a learnable frequency prompt $\mathcal{P} \in \mathbb{R}^{B\times T\times C}$ to deliver T pixel imitation principles in C channels of B frequency bands, and it will finetune the teacher model first. For simplicity, we choose the frequency band $\texttt{HH}$ from B bands and its corresponding prompt $\boldsymbol{P} \in \mathbb{R}^{T\times C}$ as an example, and the rest are the same.

Unlike previous insertion methods of spatial-based prompts, our approach requests the frequency prompt to interact with the band, a better way to know the manifolds embedded in frequency spaces. In this paper, we adopt the matrix multiplication manner to calculate the mutual information $\boldsymbol{M} \in \mathbb{R}^{C\times H_{\texttt{HH}}W_{\texttt{HH}}}$ between prompt $\boldsymbol{P}$ and frequency pixels $\boldsymbol{R}^{(t)}$ in the teacher band:

$$\boldsymbol{M} = \boldsymbol{P} \times \boldsymbol{R}^{(t)}, \tag{4}$$

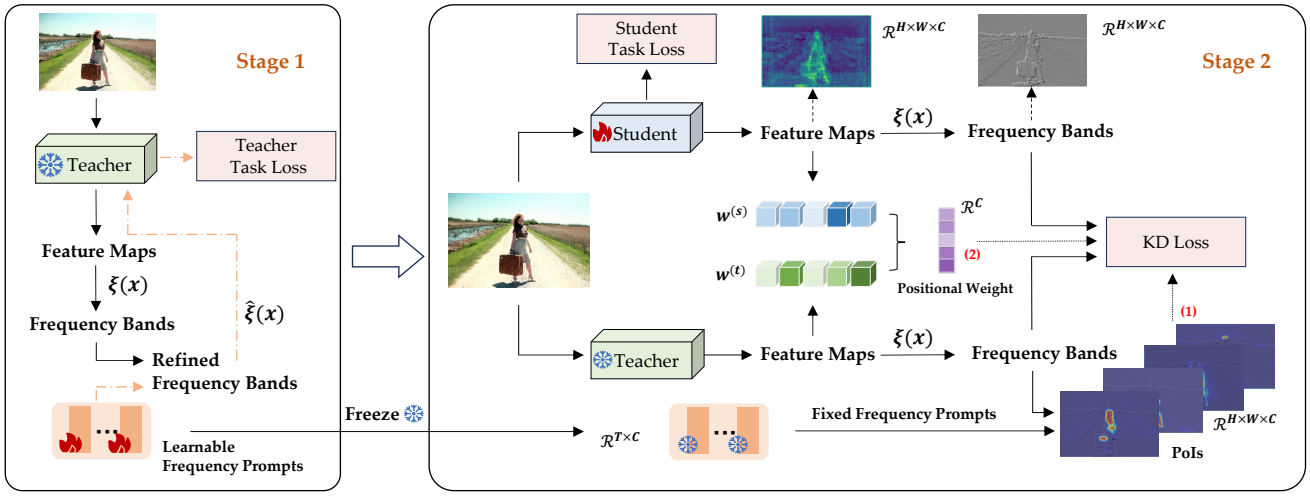

Figure 3. **Overview of our FreeKD pipeline.** The pipeline includes two stages. **Stage 1**: Frequency prompts make interaction with intermediate frequency bands, and are supervised by the teacher task loss. **Stage 2**: First, the distillation feature maps of student and teacher transform into the frequency domain, respectively. Then, receiving frequency prompts from stage 1, we request the frozen ones multiply with teacher frequency bands, and generate the PoIs of bands. Finally, a channel-wise positional-aware weight is determined by the teacher spatial gate and student gate together. The flow (1) in the figure decides where to distill and flow (2) indicates the extent of the distillation.

where we flatten the band HH into shape $(C, H_{\texttt{HH}} \times W_{\texttt{HH}})$ to fit matrix multiplication.

Then, to connect with the task loss $\mathcal{L}_{\text{finetune}}$ supervision and support stochastic gradient descent, a masked frequency band is utilized to substitute the original band HH:

$$\hat{\boldsymbol{R}^{(t)}} = \sum_{i=1}^{T} \sigma(\boldsymbol{M}_i) \circledast \boldsymbol{R}^{(t)}, \tag{5}$$

where we turn the mutual information $\boldsymbol{M}$ into a probability space to function as the masks. The symbol σ denotes the sigmoid function and $\circledast$ means element-wise multiplication.

After collecting all B masked frequency bands, we perform an Inverse Discrete Wavelet Transformation (IDWT) $\tilde{\xi}$ on them to the spatial domain:

$$\hat{\boldsymbol{F}^{(t)}} = \tilde{\xi}(\hat{\mathcal{B}_l}), \tag{6}$$

and we send the new feature map $\hat{\boldsymbol{F}^{(t)}}$ back to the teacher model. The finetune loss can be treated as an observation of mask quality, and minimize to force the frequency prompts to focus on the substantial pixels of the band.

However, simply minimizing $\mathcal{L}_{\text{finetune}}$ would lead to an undesired collapse of the T sets of masks generated by the frequency prompt. Specifically, some masks will be learned to directly recover all the bands, filled with 1 everywhere. To make the prompt represent T sets PoIs of the band, we propose a Prompt-Dissimilarity loss based on the Jaccard coefficient:

$$\mathcal{L}_{\text{dis}} = \frac{1}{T^2} \sum_{i=1}^{T} \sum_{j=1}^{T} \Theta_{\text{Jaccard}}(\boldsymbol{M}_i, \boldsymbol{M}_j) \tag{7}$$

with

$$\Theta_{\text{Jaccard}}(\boldsymbol{m}, \boldsymbol{n}) = \frac{|\boldsymbol{m} \cap \boldsymbol{n}|}{|\boldsymbol{m} \cup \boldsymbol{n}|}, \tag{8}$$

where $\boldsymbol{m} \in \mathbb{R}^N$ and $\boldsymbol{n} \in \mathbb{R}^N$ are two vectors. Jaccard loss is widely used to measure the degree of overlap between two masks in segmentation tasks. By minimizing the coefficients of each mask pair, we can make masks associated with different PoIs. As a result, the training loss of prompt is composed of finetune loss and dissimilarity loss:

$$\mathcal{L}_{\text{prompt}} = \mathcal{L}_{\text{finetune}} + \lambda \mathcal{L}_{\text{dis}}, \tag{9}$$

where λ is a factor for balancing the loss. In this paper, we set $\lambda = 1$ for all distillation experiments and allocate $T = 2$ imitation principles for each frequency band, as the frequency prompt is easier to converge (e.g., the teacher *FCOS ResNet101* has 40.8 mAP on COCO val set, and the finetuned one is 39.9). Notably, we still utilize the original teacher instead of the finetuned one to distill for the students for the fairness.

3.3. Position-aware Relational Loss

With the help of frequency prompt, we can already localize the PoIs of bands to improve the performance of frequency distillation. As frequency responses come from a local region, encoding original features with positional importance is thus necessary to distinguish the objects for dense prediction. Hence we introduce the Position-aware Relational Loss to provide high-order spatial enhancement for the student model. First, the relational attention from multi-

receptive fields can be represented as:

$$\boldsymbol{A} = Softmax(\psi(\boldsymbol{F})\boldsymbol{F}^T), \tag{10}$$

where $\psi(\boldsymbol{F})$ denotes the spatial feature of the latter layer than $\boldsymbol{F}$. Thus $\boldsymbol{A} \in \mathbb{R}^{C\times C}$ serves as a bridge to find the position-aware correlations across different layers. Then, the gating operation is generated based on the spatial perceptions to form the position-aware loss relation weight:

$$\omega = \mathcal{G}(\boldsymbol{A}) \in \mathbb{R}^{1\times C}, \tag{11}$$

where $\mathcal{G}$ denotes the gating weight generated by a Multi-layer Perceptron (MLP). Therefore, Eq. 3 can be reformulated as:

$$\mathcal{L}_{\text{FKD}} = \sum_{k=1}^{L} \omega^{(r)} \left\| a_k - b_k \right\|_1, \tag{12}$$

with $\omega^{(r)} = \omega^{(t)} \circledast \omega^{(s)}$ generated by the teacher and student position-aware relation weight. The reason is that the channels in distillation should consist of the ones both meaningful to the teacher and student. Our eventual frequency distillation loss can be formulated as:

$$\mathcal{L}_{\text{FreeKD}} = \sum_{k=1}^{L} \omega^{(r)} \left\| \boldsymbol{M} \circledast a_k - \boldsymbol{M} \circledast b_k \right\|_1. \tag{13}$$

3.4. Overall loss

To sum up, we train the student detector with the total loss formulated as:

$$\mathcal{L}_{\text{student}} = \mathcal{L}_{\text{task}} + \mu \mathcal{L}_{\text{FreeKD}}, \tag{14}$$

where μ is a factor for balancing the losses. The distillation loss is applied to intermediate feature maps (e.g., the feature pyramid network [26] (FPN) in object detection tasks), so it can be easily applied to different architectures.

4. Experiments

In this paper, to validate the superiority of our method, we conduct extensive experiments on object detection and semantic segmentation tasks, with various model architectures (including CNN-based and Transformer-based). Furthermore, we evaluate the robustness of detectors trained with FreeKD on the COCO-C benchmark, to exhibit its better domain generalization capabilities.

4.1. Object Detection

4.1.1 Datasets.

We experiment on MS COCO detection dataset [25], which contains 80 object classes. We train the student models on COCO `train2017` set and evaluate them with average precision (AP) on `val2017` set.

4.1.2 Network Architectures.

Our evaluation includes two-stage models [35], anchor-based one-stage models [27], as well as anchor-free one-stage models [38, 48], to validate the efficacy of FreeKD across diverse detection architectures.

4.1.3 Implementation Details.

For the object detection task, we conduct feature distillation on the predicted feature maps sourced from teacher's neck. We adopt ImageNet pretrained backbones and inheriting strategy following previous KD works [20, 51, 55] during training. All the models are trained with the official strategies (SGD, weight decay of 1e-4) of 2X schedule in MMDetection [4]. We train the student with our FreeKD loss $\mathcal{L}_{\text{FreeKD}}$, regression KD loss, and task loss for the object detection task. Concretely, the loss weights μ of $\mathcal{L}_{\text{FreeKD}}$ in Eq.14 on *Faster RCNN*, *RetinaNet*, *FCOS*, and *RepPoints* are 1, 5, 10, and 10.

4.1.4 Experimental Results.

Results on baseline settings. Our results compared with previous methods are summarized in Table 1, where we take ResNet-101 (R101) [16] backbone as the teacher network, and ResNet-50 (R50) as the student. Our FreeKD can significantly improve the performance of student models over their teachers on various network architectures. For instance, FreeKD improves FCOS-R50 by 4.4 AP and surpasses DiffKD [20] by 0.5 AP. Besides, FreeKD benefits more to detecting large-size objects (AP_L), as larger objects would involve more frequency bands and cross-domain information.

Results on stronger settings. We further investigate our efficacy on stronger teachers whose backbones are replaced by stronger ResNeXt (X101) [46]. The results in Table 2 demonstrate that student detectors achieve more enhancements with our FreeKD, especially when with a RepPoints-X101 teacher, FreeKD gains a substantial improvement of 3.8 AP over the RepPoints-R50. Additionally, our method outperforms existing KD methods by a large margin, and the improvement of FreeKD compared to DiffKD [20] is greater for all cases than the improvement of DiffKD [20] compared to FGD [50].

4.2. Semantic segmentation

4.2.1 Datasets.

We conduct experiments on Cityscapes dataset [7] to valid the effects of our method, which contains 5000 high-quality images (2975, 500, and 1525 images for the training, validation, and testing). We evaluate all the student networks with mean Intersection-over-Union (mIoU).

Table 1. **Object detection performance via FreeKD in baseline settings on COCO val set.**

Method		AP	AP_S	AP_M	AP_L
One-stage detectors					
T: RetinaNet-R101		38.9	21.0	42.8	52.4
S: RetinaNet-R50		37.4	20.0	40.7	49.7
FRS [10]	NeurIPS21	39.3 (1.9↑)	21.5	43.3	52.6
FGD [49]	CVPR22	39.6 (2.2↑)	**22.9**	43.7	53.6
DiffKD [20]	NeurIPS23	39.7 (2.3↑)	21.6	43.8	53.3
FreeKD		**39.9 (2.5↑)**	21.2	**44.0**	**53.7**
Two-stage detectors					
T: Faster RCNN-R101		39.8	22.5	43.6	52.8
S: Faster RCNN-R50		38.4	21.5	42.1	50.3
FRS [10]	NeurIPS21	39.5 (1.1↑)	22.3	43.6	51.7
FGD [49]	CVPR22	40.4 (2.0↑)	22.8	44.5	53.5
DiffKD [20]	NeurIPS23	40.6 (2.2↑)	23.0	44.5	54.0
FreeKD		**40.8 (2.4↑)**	**23.1**	**44.7**	**54.0**
Anchor-free detectors					
T: FCOS-R101		40.8	24.2	44.3	52.4
S: FCOS-R50		38.5	21.9	42.8	48.6
FRS [10]	NeurIPS21	40.9 (2.4↑)	25.7	45.2	51.2
FGD [49]	CVPR22	42.1 (3.6↑)	**27.0**	46.0	54.6
DiffKD [20]	NeurIPS23	42.4 (3.9↑)	26.6	45.9	54.8
FreeKD		**42.9 (4.4↑)**	26.8	**46.8**	**55.4**

Table 2. **Object detection performance via FreeKD in stronger settings on COCO val set.** *CM RCNN*: Cascade Mask RCNN.

Method		AP	AP_S	AP_M	AP_L
One-stage detectors					
T: RetinaNet-X101		41.2	24.0	45.5	53.5
S: RetinaNet-R50		37.4	20.0	40.7	49.7
FRS [10]	NeurIPS21	40.1 (2.7↑)	21.9	43.7	54.3
FGD [49]	CVPR22	40.7 (3.3↑)	**22.9**	45.0	54.7
DiffKD [20]	NeurIPS23	40.7 (3.3↑)	22.2	45.0	55.2
FreeKD		**41.0 (3.6↑)**	22.3	**45.1**	**55.7**
Two-stage detectors					
T: CM RCNN-X101		45.6	26.2	49.6	60.0
S: Faster RCNN-R50		38.4	21.5	42.1	50.3
CWD [37]	ICCV21	41.7 (3.3↑)	23.3	45.5	55.5
FGD [49]	CVPR22	42.0 (3.6↑)	23.7	46.4	55.5
DiffKD [20]	NeurIPS23	42.2 (3.8↑)	**24.2**	46.6	55.3
FreeKD		**42.4 (4.0↑)**	24.1	**46.7**	**55.9**
Anchor-free detectors					
T: RepPoints-X101		44.2	26.2	48.4	58.5
S: RepPoints-R50		38.6	22.5	42.2	50.4
FKD [53]	ICLR20	40.6 (2.0↑)	23.4	44.6	53.0
FGD [49]	CVPR22	41.3 (2.7↑)	**24.5**	45.2	54.0
DiffKD [20]	NeurIPS23	41.7 (3.1↑)	23.6	45.4	55.9
FreeKD		**42.4 (3.8↑)**	24.3	**46.4**	**56.6**

Table 3. **Semantic segmentation performance via FreeKD on Cityscapes val set.** FLOPs is measured based on an input image size of 512×512.

Method		Params (M)	FLOPs (G)	mIoU (%)
T: PSPNet-R101		70.43	574.9	78.34
S: PSPNet-R18		13.1	125.8	69.85
CWD [37]	ICCV21			73.53
MGD [51]	ECCV22			73.63
FreeKD				**74.40**
S: DeepLabV3-R18		12.6	123.9	73.20
CWD [37]	ICCV21			75.93
MGD [51]	ECCV22			76.02
FreeKD				**76.45**

Table 4. **Performance of robust object detection via FreeKD on COCO-C dataset.** Each experiment is averaged over 6 trials.

Method	mAP_{clean}	mPC	rPC
Source (Retina-R50)	37.4	18.3	48.9
FGD [49]	39.6	20.3	51.3
DiffKD [20]	39.7	20.3	51.1
FreeKD (Ours)	**39.9**	**20.8**	**52.1**

4.2.2 Network architectures.

For all segmentation experiments, we take PSPNet-R101 [56] as the teacher network. While for the students, we use various frameworks (DeepLabV3 [5] and PSPNet) with ResNet-18 (R18) to demonstrate the efficacy of our method.

4.2.3 Implementation Details.

For the semantic segmentation task, we conduct feature distillation on the predicted segmentation maps. All the models are trained with the official strategies of 40K iterations schedule with 512×512 input size in MMSegmentation [6], where the optimizer is SGD and the weight decay is 5e-4. A polynomial annealing learning rate scheduler is adopted with an initial value of 0.02.

4.2.4 Experimental results.

The experimental results are summarized in 3. FreeKD further improves the performance of state-of-the-art MGD [51] on both homogeneous and heterogeneous settings. For instance, the ResNet-18-based PSPNet gets 0.77 mIoU gain and that based DeepLabV3 gets 0.43 mIoU.

4.3. Natural Corrupted Augmentation

We evaluate the robustness of student detector RetinaNet-R50, trained with FreeKD on the COCO-C dataset [30].

Table 5. **The performance of DETR-like models via FreeKD on COCO.** *De-DETR*: Deformable DETR, *MBv2*: MobileNetV2.

Teacher	Student	Backbone	AP	AP_S	AP_M	AP_L
De-DETR R101 47.1 (50e)	De-DETR	MBv2	33.5	16.9	36.4	46.6
	+ FreeKD		**36.2 (2.7↑)**	**19.3**	**38.9**	**49.0**
	De-DETR	R18	36.4	19.6	39.0	49.3
	+ FreeKD		**38.9 (2.5↑)**	**22.0**	**41.2**	**51.9**
DINO Swin-L 56.6 (12e)	DINO	R50	48.4	30.9	51.3	63.4
	+ FreeKD		**50.4 (2.0↑)**	**33.1**	**53.6**	**64.9**
	DINO	R18	45.1	28.7	48.0	59.1
	+ FreeKD		**47.3 (2.2↑)**	**30.0**	**50.4**	**61.3**

Table 6. **The performance of SAM via FreeKD on SA-1B.**

Teacher	Students	Steps	mIoU
SAM ViT-H	SAM ViT-Tiny	20K	40.12
	+ MSE	20K	42.42
	+ FreeKD	20K	**44.63**

COCO-C is derived from `val2017` set of COCO, enriched with four types[†] of image corruption, and each type further comprises several fine-grained corruptions. The results on corrupted images compared in Table 4, the mPC improvement of FreeKD compared to DiffKD [20] is greater than mAP_{clean}, and FreeKD outperforms DiffKD [20] by 1.0% rPC[†]. Our method is beneficial to enhancing the extra robustness and domain generalization abilities of the student.

4.4. Large-Scale Vision Models Distillation

To fully investigate the efficacy of FreeKD, we further conduct experiments on much stronger large-scale teachers.

DETR-like Model. For the object detection task, we apply FreeKD for two popular DETR-based models (Deformable DETR [59] and DINO [52]) with various student backbones (R18, R50, and MobileNetV2 [36]). For De-DETR, FreeKD brings 2.5+ AP improvement for both De-DETR-R18 and De-DETR-MBv2 students. While for DINO model, it still has a 2.0+ AP gain for stronger students, e.g., DINO-R50 breaks the limit of 50 AP with the help of FreeKD. Notably, we only distill the output of the final encoder layer and train the students in 12 epochs (1X).

Segment Anything Model (SAM). For the semantic segmentation task, SAM [23] is our first choice to validate the generality of FreeKD. We take the original SAM as the teacher, and its default image encoder is based on the heavyweight ViT-H [9]. Therefore, we replace the ViT-H with ViT-Tiny as the student and transfer the dark knowledge from image embeddings, which are generated by the image encoder. The student is trained with the SA-1B dataset [23] for 20K iterations (The image encoder is distilled for 10, 000 steps with 1024 × 1024 input size, and then the mask decoder is fine-tuned for 10, 000 steps) and evaluated with mIoU between the original SAM and itself. We run all the experiments on 8 A100 GPUs. For comparisons with our baseline one spatial-level feature distillation, we also report the mean square error (MSE) results with the same distillation location as FreeKD. As summarized in Table 6, FreeKD obviously outperforms the MSE results by 2.21% on SAM ViT-Tiny and improves the student by 4.51%.

The above cases indicate that our precise frequency information in FreeKD is generic to large-scale vision models. Besides, sourced from Parameter-Efficient Fine-Tuning, the Prompt-guided distillation method thus is more fit for foundation vision teacher models, and effectively polishes up the performance of the students.

[†]including noise, blurring, weather, and digital corruption.
[†]rPC = mPC / mAP_{clean}

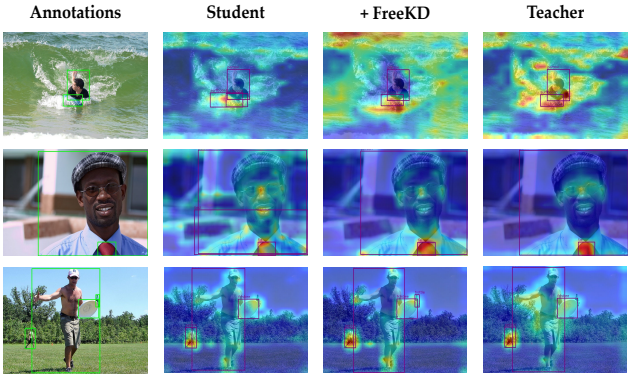

Figure 4. **Visualization of student features, student distilled with FreeKD features and teacher features on COCO dataset.** The cases are randomly selected from val set and the heatmaps are generated with AblationCAM [34].

Table 7. **Ablation study on Frequency Prompts (FP).** We use RepPoints-R50 student and RepPoints-X101 teacher on COCO with various frequency bands.

Frequency Bands		AP
Distill w/o FP.		
Low	High	
✓	✗	40.7
✗	✓	**41.8**
✓	✓	41.3

Frequency Bands		AP
Distill w/ FP.		
Low	High	
✓	✗	41.0
✗	✓	42.3
✓	✓	**42.4**

5. Analysis

5.1. Effects of Frequency Prompts

We propose a semantic Frequency Prompt (FP) to localize the PoIs of both high and low-frequency bands to com-

Table 8. **The comparison of attention weights on COCO (AP) via FreeKD.** Teacher: RepPoints-X101. Student: RepPoints-R50.

Student	SE	Non-local	CBAM	Ours
37.4	42.2	41.9	42.1	**42.4**

Table 9. **Various Frequency Transformation Manners for FreeKD.** We use RepPoints-R50 student and RepPoints-X101 teacher on COCO.

Method	Mother Function	AP
DCT	Cosine	41.9
DFT	Sine and Cosine	42.0
DWT	Wavelet	**42.4**

pensate for their own limitations in the distillation. Here we conduct experiments to compare the effects of FP on different frequency bands in Table 7. We can see that: **(a)** Only low-frequency distillation cannot help polish up the student, and even impair the performance (-0.5 AP) when combined with high-frequency bands. **(b)** When Frequency Prompt provides accurate PoIs, the low-frequency band eliminates harmful samples with 0.3 AP gain, and the high filters extra noise by 0.5 AP improvement. **(c)** In general, FP has improved frequency distillation by 0.6 AP and unified the distillation framework of frequency bands.

5.2. Effects of Position-aware Weight

To validate our Position-aware weight effectiveness, we choose several spatial attention (Squeeze and Excitation (SE) [18], Non-local Module [41], and Convolutional Block Attention Module (CBAM) [43]) to watch Frequency distillation. The results are reported on Table 8. We find that enhancing frequency distillation from channel dimension is a more effective method (SE and ours), compared with the other two. Besides, our position-aware weight includes distinguished object information with multi-scale receptive fields, which is more urgent to the frequency domain.

5.3. Effects of Frequency Transformation Manner

In terms of which frequency transformation is more suitable for distillation, we conduct detailed experiments on three methods (Discrete Cosine Transform (DCT), Discrete Fourier Transform (DFT), and Discrete Wavelet Transform (DWT)). As shown in Table 9, DWT based on Wavelet is significantly superior to DCT and DFT, whose mother functions are trigonometric functions. The reason is that wavelet provides frequency domain information at different scales, facilitating the analysis of local signal features, while the trigonometric function only provides global frequency domain information.

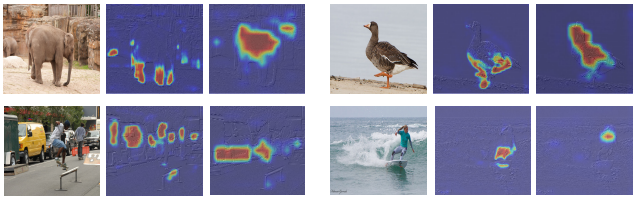

Figure 5. **Visualization of high-frequency pixels of interests on COCO dataset via RepPoints-X101.**

5.4. Visualization

We visualize the prediction results and heatmaps of the detector in Figure 4 to further investigate the efficacy of FreeKD. We utilize RepPoints-R50 student and RepPoints-X101 teacher as an example. In general, FreeKD yields more clear contrast between low-frequency pixels and high-frequency pixels in heat maps, and it provides more distinctive observation. For instance, in the third case, the student trained by FreeKD performs better than the teacher (e.g., it detects the bottle successfully). The reason is that: Firstly, the high-frequency imitation principle in FreeKD, with shorter wavelengths, enables the student to effectively capture details, edges, and richer textures, thereby focusing on small but crucial regions. Secondly, position-aware relational loss enhances the student's sensitivity to positional information, particularly for small objects. Lastly, the combination of ground truth and soft labels provides the student with robust supervision signals.

Meanwhile, FreeKD effectively avoid generating redundant bounding boxes in the first two cases, due to its spatial perception of objects. Besides, we visualize the two PoIs (masks) generated by frequency prompt in the high-frequency band `HH` in Figure 5. We find that the distinctive details in the band are marked out, while the noise is avoided to prevent performance degradation. This verifies our frequency prompt is effective in practice.

6. Conclusion

This research shifts the attention to frequency domain, and highlights its potential for knowledge distillation on dense prediction tasks. Meanwhile, to tackle the natural shortcomings of high and low frequency during mimicking, we introduce a novel pipeline named FreeKD, which determines both the optimal localization and extent for the frequency distillation. Specifically, we first propose Frequency Prompt to generate pixel-wise imitation principles. Besides, we design a channel-wise position-aware relational loss to enhance the sensitivity to objects for dense prediction. Extensive experiments demonstrated that FreeKD outperforms spatial-based distillation methods and provides more robustness to the student model.

References

[1] Weihan Cao, Yifan Zhang, Jianfei Gao, Anda Cheng, Ke Cheng, and Jian Cheng. Pkd: General distillation framework for object detectors via pearson correlation coefficient. *Advances in Neural Information Processing Systems*, 35: 15394–15406, 2022. 3, 1

[2] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *NeurIPS*, 30, 2017. 2

[3] Hanting Chen, Yunhe Wang, Han Shu, Yehui Tang, Chunjing Xu, Boxin Shi, Chao Xu, Qi Tian, and Chang Xu. Frequency domain compact 3d convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1641–1650, 2020. 1

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018. 6

[6] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 6

[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 *CVPR*, 2016. 5

[8] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *CVPR*, pages 7842–7851, 2021. 1

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7

[10] Z. Du, R. Zhang, M. Chang, X. Zhang, S. Liu, T. Chen, and Y. Chen. Distilling object detectors with feature richness. In *NeurIPS*, 2021. 6

[11] Zhixing Du, Rui Zhang, Ming-Fang Chang, Xishan Zhang, Shaoli Liu, Tianshi Chen, and Yunji Chen. Distilling object detectors with feature richness. In *NeurIPS*, 2021. 1

[12] Shin Fujieda, Kohei Takayama, and Toshiya Hachisuka. Wavelet convolutional neural networks for texture classification. *arXiv preprint arXiv:1707.07394*, 2017. 3

[13] Yulu Gan, Yan Bai, Yihang Lou, Xianzheng Ma, Renrui Zhang, Nian Shi, and Lin Luo. Decorate the newcomers: Visual domain prompt for continual test time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7595–7603, 2023. 2

[14] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *CVPR*, pages 2154–2164, 2021. 3

[15] Bruce C Hansen and Robert F Hess. Structural sparseness and spatial phase alignment in natural scenes. *JOSA A*, 24 (7):1873–1885, 2007. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 5

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *NeurIPS Workshop*, 2014. 1

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 8

[19] Tao Huang, Yuan Zhang, Shan You, Fei Wang, Chen Qian, Jian Cao, and Chang Xu. Masked distillation with receptive tokens. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[20] Tao Huang, Yuan Zhang, Mingkai Zheng, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge diffusion for distillation. *NeurIPS*, 30, 2023. 1, 2, 3, 5, 6, 7

[21] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 2

[22] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13919–13929, 2021. 3

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 7

[24] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *CVPR*, pages 6356–6364, 2017. 2

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 5

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 5

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 5

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[29] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999. 3

[30] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 6

[31] Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981. 3

[32] Yingxue Pang, Xin Li, Xin Jin, Yaojun Wu, Jianzhao Liu, Sen Liu, and Zhibo Chen. Fan: frequency aggregation network for real image super-resolution. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 468–483. Springer, 2020. 3

[33] Leon N Piotrowski and Fergus W Campbell. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, 11(3):337–346, 1982. 3

[34] Harish Guruprasad Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *WACV*, pages 983–991, 2020. 7

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. 5

[36] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 7

[37] Changyong Shu, Yifan Liu, Jianfei Gao, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *ICCV*, 2021. 3, 6, 1

[38] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, pages 9627–9636, 2019. 5

[39] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 2

[40] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *CVPR*, pages 4933–4942, 2019. 3

[41] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 8

[42] Travis Williams and Robert Li. Wavelet pooling for convolutional neural networks. In *International conference on learning representations*, 2018. 3

[43] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 8

[44] Shijun Xiang, Hyoung Joong Kim, and Jiwu Huang. Invariant image watermarking based on statistical features in the low-frequency domain. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(6):777–790, 2008. 2

[45] Jiahao Xie, Wei Li, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Masked frequency modeling for self-supervised visual pre-training. *arXiv preprint arXiv:2206.07706*, 2022. 1

[46] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500, 2017. 5

[47] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1740–1749, 2020. 1, 3

[48] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *ICCV*, pages 9657–9666, 2019. 5

[49] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *CVPR*, pages 4643–4652, 2022. 3, 6

[50] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *CVPR*, pages 4643–4652, 2022. 1, 5

[51] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. In *ECCV*, pages 53–69. Springer, 2022. 2, 5, 6

[52] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. 7

[53] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*, 2020. 6

[54] Linfeng Zhang, Xin Chen, Xiaobing Tu, Pengfei Wan, Ning Xu, and Kaisheng Ma. Wavelet knowledge distillation: Towards efficient image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12464–12474, 2022. 2, 3

[55] Yuan Zhang, Weihua Chen, Yichen Lu, Tao Huang, Xiuyu Sun, and Jian Cao. Avatar knowledge distillation: self-ensemble teacher paradigm with uncertainty. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5272–5280, 2023. 5

[56] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 6

[57] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9516–9526, 2023. 2

[58] Shengyang Zhu, Jizhong Yang, Chengbiao Cai, Zili Pan, and Wanming Zhai. Application of dynamic vibration absorbers in designing a vibration isolation track at low-frequency domain. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 231(5): 546–557, 2017. 2

[59] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 7