FOAM: A General Frequency-Optimized Anti-Overlapping Framework for Overlapping Object Perception

Mingyuan Li, Tong Jia*, Han Gu, Hui Lu, Hao Wang, Bowen Ma, Shuyang Lin, Shiyi Guo, Shizhuo Deng, and Dongyue Chen

***Abstract*—Overlapping object perception aims to decouple the randomly overlapping foreground-background features, extracting foreground features while suppressing background features, which holds significant application value in fields such as security screening and medical auxiliary diagnosis. Despite some research efforts to tackle the challenge of overlapping object perception, most solutions are confined to the spatial domain. Through frequency domain analysis, we observe that the degradation of contours and textures due to the overlapping phenomenon can be intuitively reflected in the magnitude spectrum. Based on this observation, we propose a general Frequency-Optimized Anti-Overlapping Framework (FOAM) to assist the model in extracting more texture and contour information, thereby enhancing the ability for anti-overlapping object perception. Specifically, we design the Frequency Spatial Transformer Block (FSTB), which can simultaneously extract features from both the frequency and spatial domains, helping the network capture more texture features from the foreground. In addition, we introduce the Hierarchical De-Corrupting (HDC) mechanism, which aligns adjacent features in the separately constructed base branch and corruption branch using a specially designed consistent loss during the training phase. This mechanism suppresses the response to irrelevant background features of FSTBs, thereby improving the perception of foreground contour. We conduct extensive experiments to validate the effectiveness and generalization of the proposed FOAM, which further improves the accuracy of state-of-the-art models on four datasets, specifically for the three overlapping object perception tasks: Prohibited Item Detection, Prohibited Item Segmentation, and Pneumonia Detection. The code will be open source once the paper is accepted.**

***Index Terms*—Overlapping object perception, frequency domain learning, object detection, transformer detection.**

I. Introduction

Overlapping object perception, a fundamental task within the realm of computer vision, involves tasks like prohibited item detection, prohibited item segmentation, and pneumonia detection. These tasks aim to decouple the randomly overlapping foreground-background features and extract foreground features while suppressing background features, holding significant application value in fields such as security inspection and medical auxiliary diagnosis.

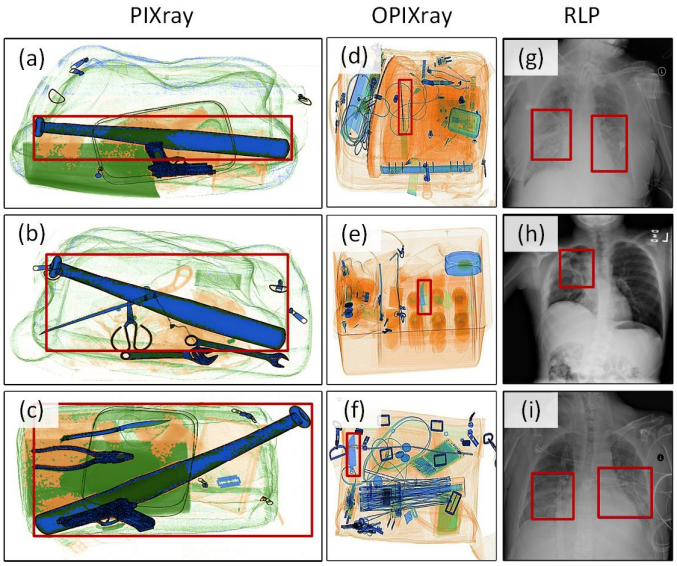

Fig. 1. The overlapping phenomenon causes the contours and textures of foreground objects to be randomly degraded in different background scenarios. (a)-(c) show the baseball "Bat" in the PIXray [1] dataset, (d)-(f) display the "Straight Knife" in the OPIXray [2] dataset, and (g)-(i) depict "Lung Opacity" in the RLP subset of the RSNA [3] dataset.

Fig. 1 illustrates typical overlapping phenomena in the PIXray, OPIXray, and RLP datasets. The contours and textures of foreground objects, such as the metal baseball "Bat" and "Straight Knife", in subfigures (a)-(c) and (e)-(f), are degraded to varying degrees by randomly occurring backgrounds. Therefore, in real-world security inspection scenarios, even the most advanced general vision models struggle to perform accurate object detection or instance segmentation of meticulously concealed prohibited items in X-ray images. Similarly, in subfigures (g)-(i), the imaging texture and contour of "Lung Opacity" are unclear, making them highly susceptible to interference from background elements [4], such as potential EKG leads, external tubes, artifacts, overlapping devices, bones, and healthy tissues. Therefore, in real-world medical auxiliary

Mingyuan Li, Han Gu, Hui Lu, Hao Wang, Bowen Ma, Shuyang Lin, Shiyi Guo, Shizhuo Deng and Dongyue Chen are with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang, 110819, Liaoning, China, and also with the College of Information Science and Engineering, Northeastern University, Shenyang, 110819, Liaoning, China (e-mail: 542027743@qq.com; 2400800@stu.neu.edu.cn; 2603813543@qq.com; ddsywh@yeah.net; 2010285@stu.neu.edu.cn; 2210329@stu.neu.edu.cn; guoshiyi@ise.neu.edu.cn; dengshizhuo@mail.neu.edu.cn; chendongyue@ise.neu.edu.cn).

Tong Jia is with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang, 110819, Liaoning, China, the College of Information Science and Engineering, Northeastern University, Shenyang, 110819, Liaoning, China, and the Key Laboratory of Data Analytics and Optimization for Smart Industry, Ministry of Education, Northeastern University, Shenyang, 110819, Liaoning, China (e-mail: jiatong@ise.neu.edu.cn). *(Corresponding author: Tong Jia.)*

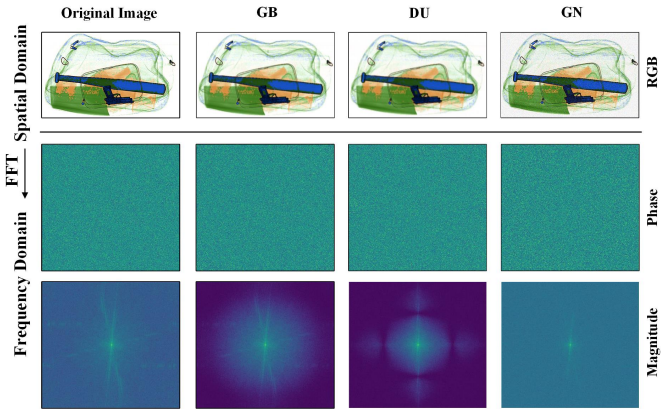

Fig. 2. X-ray images analysis in spatial and frequency domains. GB, DU, and GN represent Gaussian Blurring, Downsampling and Upsampling, and Gaussian Noise corruption strategies, respectively.

diagnosis scenarios, general vision models face significant challenges in performing accurate object detection on diverse pathological tissues.

Recently, an increasing number of studies have proposed advanced deep learning techniques to specifically address this task, especially after the introduction of several large-scale pseudo-colored X-ray datasets. Specifically, methods such as GADet [5], AO-DETR [6], Xdet [7], and CLCXray [8] introduce label assignment strategies to ensure that models consistently focus on high-quality foreground objects during training. OPIXray [2], SIXray [9], and PID-YOLOX [10] leverage attention mechanisms to assist models in decoupling and extracting foreground features from overlapping foreground and background. Additionally, PIXray [1] and AO-DETR [6] adopt multistage regression approaches to perceive blurred contours. However, the aforementioned methods rely on spatial domain information to perceive contraband, making the models susceptible to interference and deception from unknown texture and contour in the background [11].

A promising novel approach is to leverage frequency domain learning to enhance the contour and texture details of the foreground, complement spatial domain information, and thereby improve the model's ability to perceive overlapping objects. Specifically, FAPID [12] truncates the frequency domain information obtained from the Fast Fourier Transform (FFT) [13] using a fixed high-pass filter, which serves as contour and texture cues to correct spatial domain features. FDTNet [14] uses a CBAM-like attention mechanism to refine the local frequency domain information obtained from the SRM [15] filter, aiming to adaptively extract informative frequency information to complement the spatial domain information. However, the implementation of a fixed high-pass filter in FAPID results in the exclusion of valuable low-frequency information, whereas the use of the SRM filter in FDTNet [14] is limited in its global perception capabilities when compared to FFT. Therefore, although these two approaches validate and demonstrate the effectiveness and research value of frequency domain learning for overlapping object perception tasks, their understanding and development are limited, leaving significant room for improvement.

We conduct an in-depth analysis of the characteristic representations of the magnitude spectrum and phase spectrum derived from the frequency domain information obtained through FFT. As shown in Fig. 2, after applying classic corruption methods such as Gaussian Blurring (GB), Downsampling and Upsampling (DU), and Gaussian Noise (GN), the contours and textures information of the images are compromised. Obviously, neither the spatial domain image nor the phase spectrum in the frequency domain, which is adept at capturing object shapes and structures, shows significant changes. In contrast, the magnitude spectrum is more sensitive to changes in texture and contour information, revealing underlying patterns in the image that are not easily observed from the raw pixel values. For example, in the magnitude spectrum, it can be observed that GB and DU primarily remove mid- to high-frequency fine details from the image, while GN injects mid- to high-frequency noise. Therefore, we believe that the magnitude spectrum provides informative information for frequency domain learning for overlapping object perception tasks, facilitating the decoupling of foreground and background information, and the experimental results in Sec. IV-E5 corroborate our theory.

Based on this observation, we propose a general Frequency-Optimized Anti-Overlapping Framework (FOAM) for accurate overlapping object perception, which combines global frequency features with local spatial features to capture texture and contour information, thereby enhancing feature discriminability.

To implement FOAM, we first design a fundamental building block named Frequency Spatial Transformer Block (FSTB), which simultaneously extracts features from both the spatial and frequency domains and helps the network extract more texture features from the foreground. It consists of three components: the Frequency Domain Bands Self-Attention (FDBA) mechanism, the Spatial Domain Channel Self-Attention (SDCA) module, and the Frequency Spatial Feed-forward Network (FSFN). FDBA mechanism leverages the global dependency relationship among different frequency bands to reconstruct their proportions while keeping the phase unchanged, thereby correcting the texture and contour information perceived in the spatial domain. The lightweight SDCA module optimizes the local spatial details representation of objects. FSFN is responsible for the integration and optimization of both spatial and frequency domain information. Then, as a fundamental unit, FSTB is utilized to iteratively optimize the features of the backbone for N iterations, and obtains the basic feature set. This branch is referred to as the base branch and is enabled during both training and inference.

Furthermore, we propose a Hierarchical De-Corrupting (HDC) mechanism, which establishes a corruption branch that is enabled only during the training phase. We apply corruption strategies to the original image to simulate the blurring of textures and contours of the current foreground object under more severe overlapping phenomena, as well as the introduction of background noise, resulting in a low-quality corrupted image. Subsequently, we optimize the features through a shared-weight backbone and FSTBs to obtain the corrupted feature set. Finally, we employ a consistent loss

to align the features from the corruption branch with those from the base branch, guiding the FSTBs to suppress the response to irrelevant background features and adapting to the anti-overlapping perception task.

We utilize FOAM across four datasets to further improve the accuracy of state-of-the-art models on three overlapping object perception tasks: Prohibited Item Detection, Prohibited Item Segmentation, and Pneumonia Detection.

Our main contributions are summarized as follows:

1) We propose the Frequency-Optimized Anti-Overlapping Framework (FOAM), which leverages both frequency domain and spatial domain cues to help models capture more texture and contour under the negative impact of overlapping scenes for object perception. This architecture is designed to be compatible with most CNN-based and Transformer-based object detection and instance segmentation models.
2) To improve the comprehensive understanding capability of texture features, we design the Frequency Spatial Transformer Block (FSTB) to simultaneously extract foreground cues from both the frequency domain and the spatial domain.
3) To improve the perception ability for the foreground contour of networks, we propose the Hierarchical De-Corrupting (HDC) mechanism, which utilizes features from the base branch to supervise the features from the corruption branch during the training phase, suppressing the response to irrelevant background features of FSTBs.

II. Related Work

A. Frequency Domain Learning

Frequency domain information, distinct from spatial domain information, represents a unique form of high-order information with global feature representation, offering a distinctive perspective for image processing and understanding. Therefore, frequency domain learning has often been utilized for analysis and applications in the fields of image compression and super-resolution [16]–[19]. Recently, some works [20]–[26] on frequency domain learning have made progress in remote sensing and camouflaged object detection, sparking a wave of exploration among visual perception researchers. Specifically, Xu et al. [20] builds upon the SE-block [27] and proposes a learning-based dynamic channel selection method to identify trivial frequency components for static removal during inference, which is the first work to explore frequency domain learning in object detection and instance segmentation. FcaNet [28] proposes to leverage frequency domain learning to address the information loss problem in channel attention mechanisms. SPANet [29] handles the balancing problem of high- and low-frequency components in visual features. However, the aforementioned work did not explore the interaction between RGB images and frequency domain cues. Zhong et al. [30] applies the Discrete Cosine Transform (DCT) [31] to every 8×8 patch to extract frequency domain clues and uses a multi-head attention mechanism to combine frequency domain information with RGB domain information. The frequency domain information obtained through DCT consists of real-valued data, lacking the representation of the phase spectrum feature that is critical for capturing object position and structural details. Moreover, the approach of dividing features into patches before transformation causes the frequency domain representation to lose the advantage of global perception. FSEL [32] further employs the Fast Fourier Transform (FFT) to extract global frequency domain clues and integrates frequency domain and spatial domain information using a variant of the self-attention mechanism. However, this work lacks the independent design based on the distinct characteristics of magnitude and phase. In the domain of prohibited item detection, to the best of my knowledge, there is only two relevant works. FAPID [12] truncates the frequency domain information obtained from the Fast Fourier Transform (FFT) [13] using a fixed high-pass filter, which serves as contour and texture cues to correct spatial domain features, whereas completely ignoring the low-frequency information. FDTNet [14] uses an SRM filter to provide frequency domain information, but its filtering approach is essentially a set of fixed large kernel convolutions, which lack the global perspective compared to the frequency domain information obtained from Fourier transforms.

In this paper, we delve into the characteristics of the frequency domain signals obtained through FFT transformation, decoupling the learning of magnitude and phase, and combine frequency domain clues with RGB domain information to extract the contour and texture details of foreground objects.

B. Attention Mechanism in Computer Vision

The main goal of the attention mechanism is to help the model mimic the human visual system's ability to focus on foreground information in images rather than irrelevant background. This mechanism can typically be divided into CNN-based attention mechanisms and Transformer-based self-attention mechanisms.

For CNN-based attention mechanisms, SENet [27] proposes the most well-known channel attention mechanism, which compresses features into a vector using global average pooling, and then applies a fully connected layer to weight the features of each layer. GE [33] employs spatial attention to better exploit the feature context. Building upon these works, models like CBAM [34], CA [35], BAM [36], DAN [37], and PID-YOLOX [10] explore the integration of spatial and channel attention. CBAM [34] argues that global average pooling leads to information loss, prompting the introduction of global max pooling, which achieves superior performance. Inspired by this, GSoP [38] and SRM [15] further incorporate second-order pooling and global standard deviation pooling. SkNet [39] and ResNeSt [40] propose selective channel aggregation and attention mechanisms.

For the Transformer-based self-attention mechanism, the Transformer [41] was originally designed by Vaswani et al. for NLP tasks, relying solely on attention mechanisms and dispensing with recurrence and convolutions entirely. The Transformer has the ability to model global semantics and long-range dependencies, and its ideas have inspired many works in computer vision, including classification models like ViT [42] and PVT [43], object detection models

like DETR [44] and Deformable-DETR [45], and instance segmentation models like CondInst [46], Cascade-Mask-R-CNN [47], MaskFormer [48], and Mask2Former [49]. For example, ViT [42] divides images into independent patches to reduce the cost of capturing long-range relationships. The Swin Transformer [50] further enhances model efficiency through a shift operation. Additionally, other works, such as EfficientViT [51] and PVTv2 [52], have also achieved good performance.

However, these methods focus on global modeling of spatial domain features, neglecting the powerful representational capability of frequency domain information for textures and contours. Therefore, we aim to propose a transformer-based attention mechanism for integrating and refining both spatial and frequency domain features to enhance the model's ability to perceive informative foreground elements in X-ray images.

C. Prohibited Item Perception

Following the introduction of the first pseudo color large-scale X-ray image dataset, SIXray [9], a multitude of modern pseudo color X-ray prohibited perception datasets has been developed. These datasets are specifically designed for various tasks, including classification with SIXray, object detection using OPIXray, PIXray-det [1], HIXray [53], PIDray-det [54], DvXray [55], and CLCXray [8], as well as segmentation with PIXray-seg [1] and PIDray-seg [54]. Given the significant practical application value of object detection tasks, the majority of contemporary prohibited item perception methodologies predominantly focus on the development and enhancement of prohibited item detectors. Most of them are optimized based on traditional object detectors to cater to the unique imaging characteristics of X-ray images. Specifically, SIXray introduces a feature pyramid network FPN-like approach named class-balanced hierarchical refinement, which supervises lower-level features with higher-level features, thereby enhancing the focus on foreground information. OPIXray [2] and OVXD [56] propose the DOAM module and bottleneck-like adapter, respectively, which emphasize foreground materials and contour information. GADet [5] and Xdet [7] present the IAA and HSS labeling strategies, respectively, to improve foreground perception accuracy by alleviating the issue of class imbalance between foreground and background categories. AO-DETR [6] is the first to introduce a DETR-like architecture, DINO [57], in the field of prohibited detection, proposing the CSA strategy to train category-specific content queries that are specifically responsible for perceiving particular categories of contraband. Furthermore, MMCL [58] and CSPCL [59] propose plug-and-play contrastive learning strategies to address the issue of distribution imbalance of category-specific content queries in Deformable-DETR-like models. However, the aforementioned methods are all limited to spatial domain feature perception and fail to utilize the representational power of frequency domain information regarding contours and textures. FDTNet and FAPID employ convolutional or self-attention mechanisms to extract frequency domain information, thereby acquiring contour and texture information from foreground objects. However, their frequency domain information is derived from the SRM filter or high-pass filter, which incurs more feature loss compared to FFT transformation, resulting in them being only locally optimal solutions. After decoupling it into magnitude and phase spectra with distinct characteristics, we design a frequency domain attention mechanism in an attempt to identify the optimal strategy for frequency domain information extraction.

III. Methodology

In this section, we first introduce the principles and properties of image Fourier transformation. We then present the overall structure of the proposed Frequency-Optimized Anti-Overlapping Framework (FOAM), as depicted in Fig. 3. Subsequently, we describe the core module, the Frequency Spatial Transformer Block (FSTB), which internally includes the Frequency Domain Bands Self-Attention (FDBA) mechanism, the Spatial Domain Channel Self-Attention (SDCA) module, and the Frequency Spatial Feed-forward Network (FSFN). Finally, we explain how the Hierarchical De-Corrupting (HDC) mechanism utilizes the proposed consistent loss to deeply activate the anti-overlapping perception capability of the FSTB.

A. Image Fourier Transformation

2D Discrete Fourier Transform (2D DFT) is a widely used method for analyzing the frequency content of images. Compared to the 2D Discrete Cosine Transform (2D DCT), it additionally provides phase information, offering a comprehensive representation and description about location and structure of the objects [60]. In contrast to the 2D Discrete Wavelet Transform (2D DWT) [60], it dynamically transforms low and high-frequency elements and does not need predefined kernels to separate frequency components [61], thereby providing more informative clues about the detection of foreground objects for the model. For multi-channel image signals, the Fourier transform is typically applied to each channel individually. To simplify, we omit the channel notation in the subsequent equations. Let $X \in \mathbb{R}^{C\times H\times W}$ represent the image, and the 2D DFT maps it to the complex components in the Fourier space $F(u,v)$, which can be expressed as:

$$F(u,v)=\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}X(h,w)e^{-j2\pi(\frac{h}{H}u+\frac{w}{W}v)}, \tag{1}$$

where $F(u,v)\in\mathbb{C}^{H\times W}$. $u\in\{0,1,...,H-1\}$ and $v\in\{0,1,...,W-1\}$ represent the vertical and horizontal frequency indices, respectively. The 2D Inverse Discrete Fourier Transform (2D IDFT) is denoted as:

$$X(h,w)=\frac{1}{HW}\sum_{u=0}^{H-1}\sum_{v=0}^{W-1}F(u,v)e^{j2\pi(\frac{h}{H}u+\frac{w}{W}v)}. \tag{2}$$

In our work, we employ efficient and equivalent Fast Fourier Transform (FFT) and Invert Fast Fourier Transform (IFFT) [13] to replace the 2D DFT and its inverse transform, processing each image channel individually, as followed in [32], [61].

The magnitude component $M(u,v)$ and phase component $P(u,v)$ are defined as:

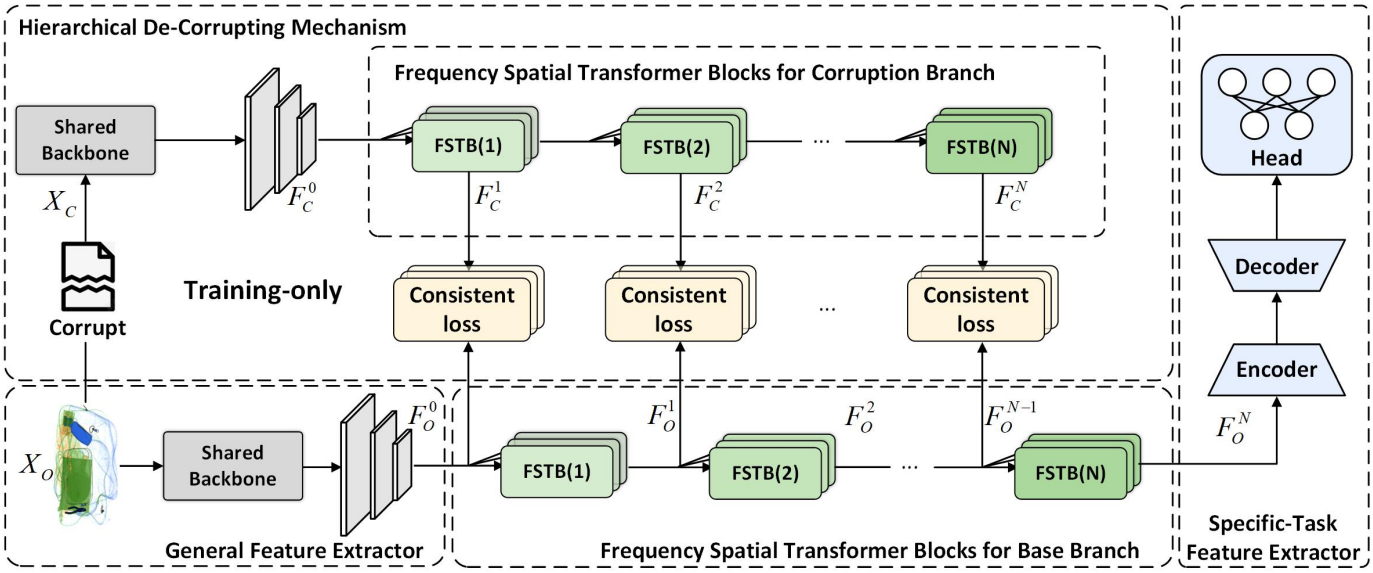

Fig. 3. Overall architecture of our proposed FOAM. The core module is FSTB, which is designed for joint learning in both the frequency and spatial domains, enabling the backbone network to effectively perceive and extract texture features, including both foreground and background. The HDC mechanism first corrupts the original image X_O to obtain the corrupted image X_C. After feature extraction through a shared-parameter extractor, cascaded FSTBs are then applied to generate the base branch and the corruption branch. Finally, a consistent loss is employed to align adjacent features from the two branches, suppressing the FSTB's response to irrelevant background features and clarifying the edges of foreground objects.

$$M(u,v) = \sqrt{R^2(u,v) + I^2(u,v)}, \tag{3}$$

$$P(u,v) = \arctan(\frac{I(u,v)}{R(u,v)}), \tag{4}$$

$$R(u,v) = (F(u,v) + conj(F(u,v)))/2, \tag{5}$$

$$I(u,v) = (F(u,v) - conj(F(u,v)))/2j. \tag{6}$$

Here, $R(u,v)$ and $I(u,v)$ represent the real and imaginary parts of $F(u,v)$ respectively, and $conj(\cdot)$ denotes the conjugate complex number operator.

As shown in Fig. 2, the frequency spectrum and the phase spectrum emphasize different aspects of image representation. The former is adept at capturing texture and contour information, while the latter excels in capturing shape and structural information [60]. In Sec. III-C, we design a joint learning method that combines the frequency and spatial domains, leveraging their complementary information based on the aforementioned characteristics.

B. Frequency-Optimized Anti-Overlapping Framework

As shown in Fig. 3, the architecture of FOAM differs between the training and inference phases. In the inference phase, the network structure only requires the insertion of the base branch, constructed using the FSTB proposed in Sec. III-C, into the general object perception model. This allows for fine-tuning the features extracted by the backbone network, thereby aiding the Specific-Task Feature Extractor in better handling the overlapping object perception task. In the training phase, the HDC mechanism is additionally introduced, where the corruption branch is constructed to supervise the features of the base branch. This process guides the FSTB to reduce its response to background features, further enhancing the perception ability for the foreground contour.

1) Base Branch: The base branch is enabled during both training and inference, realized by a set of FSTB modules inserted into the original model, whose details will be presented in Sec. III-C. Specifically, given the original image X_O, the network first applies a standard backbone, *i.e.*, ResNet, ResNeXt, and Swin Transformer, to obtain the initial multi-scale features $F_O^0 = \{F_{O,l}^0\}_{l=0}^L$ for the base branch, where L are the stage number of the backbone. Then, N independent FSTBs are cascaded together to adaptively perform entanglement learning on F_O^0 from both the spatial and frequency domains, resulting in a discriminative feature set $F_O = \{F_O^1, F_O^2, ..., F_O^N\}$. Among them, the feature enhanced by n FSTB modules is expressed as follows:

$$F_O^n = (FSTB^n \circ FSTB^{n-1} \circ ... \circ FSTB^1)(F_O^0), \tag{7}$$

where $\circ$ is the composition operator, and $FSTB^n$ represents the n-th FSTB operator. Finally, the prediction results are obtained through a specific-task feature extractor, consisting of the encoder, decoder, and head for specific-task.

2) Hierarchical De-Corrupting Mechanism: HDC mechanism is enabled only during the training phase. Corruption operation, such as Gaussian Blurring (GB), Downsampling and Upsampling (DU), and Gaussian Noise (GN), is applied to the original image to obtain a corrupted image X_C, which further disrupts the textures and contours of overlapped target objects to simulate more severe overlapping scenes. Similarly to the base branch, the initial multi-scale features F_C^0 for the corruption branch is obtained through a backbone with shared weights, followed by N FSTB operations, whose parameters are consistent with those of the N FSTBs in the base flow, resulting in the corruption version of the discriminative feature set $F_C = \{F_C^1, F_C^2, ..., F_C^N\}$. Theoretically, since the corruption operations disrupt the texture and outline details,

the effective features in the corrupted feature F_C^n at the n-th stage are fewer than those in the base branch F_O^n. Ideally, the FSTB is designed to extract and enhance relevant information. Therefore, the quality of discriminative information in F_C^{n+1}, which undergoes an additional FSTB operation, is superior to that in F_C^n. We propose a Type I consistent loss based on KL divergence, which employs the strategy of supervising F_C^{n+1} with F_O^n to achieve fine-grained alignment of the two multi-scale feature sets F_C and F_O, thereby directing FSTBs to enhance the de-corruption capability in extracting and reinforcing contours and textures. The process is as follows:

$$L_C^I(F_O, F_C) = -\sum_{n=1}^{N}\sum_{l\in\mathcal{L}}\sum_{i=1}^{HW}\hat{F}_{O,l}^{n-1}(i)\log(\frac{\hat{F}_{O,l}^{n-1}(i)}{\hat{F}_{C,l}^{n}(i)}), \quad (8)$$

$$\hat{F}_{O,l}^{n-1}(i) = \frac{\exp(F_{O,l}^{n-1}(i))}{\sum_{i=1}^{HW}\exp(F_{O,l}^{n}(i))}, \quad (9)$$

$$\hat{F}_{C}^{n}(i) = \frac{\exp(F_{C,l}^{n}(i))}{\sum_{i=1}^{HW}\exp(F_{C,l}^{n}(i))}, \quad (10)$$

where $\mathcal{L}$ is the target layer set involved in fine-grained alignment. Note that F_O^n is multi-scale features with L scales, where the higher-level features typically extract abstract global information, excelling in representing semantic characteristics and contextual relationships. In contrast, the lower-level features often capture local information but may contain redundant information and noise. Therefore, it is necessary to select the appropriate layer set $\mathcal{L}$ for feature alignment in order to achieve an optimal balance. Related ablation experiment results and analysis are shown in Sec. IV-E4.

The exact form of the consistent loss is not crucial. We also propose an alternative variant of the consistent loss that possesses similar properties and yields comparable results. This variant is designed based on the MSE loss and is referred to as the Type II consistent loss, as expressed in the following formula:

$$L_C^{II}(F_O, F_C) = \frac{1}{HW}\sum_{n=1}^{N}\sum_{l\in L}\sum_{i=1}^{HW}(F_{O,l}^{n-1}(i) - F_{C,l}^{n}(i))^2. \quad (11)$$

The experimental results comparing the consistent loss of Type II and Type I are presented in Sec. IV-E4. Both of them achieve the global optimal value when $\{\{F_{O,l}^n\}_{n=0}^{N-1}\}_{l\in\mathcal{L}}$ and $\{\{F_{C,l}^n\}_{n=1}^{N}\}_{l\in\mathcal{L}}$ are perfectly aligned, indicating that FSTBs has obtained ideal de-corruption capabilities by reducing the model's response to background noise and blur caused by overlapping phenomena. According to Theorem I, this suppression ability of background features ultimately manifests as an enhancement in contour perception, which is also corroborated by the feature map visualizations in Sec. IV-F1.

Theorem I: *Assume that in an overlapping scene, a homogeneous foreground overlaps with a homogeneous background. Let positive numbers f and b represent the response values of the neural network to the foreground and background, respectively. In the ideal linear case, the response value of the foreground region is $f+b$. The contrast at the foreground contour is defined as: $\frac{f+b}{b}$. When the model's response to the background decreases by c, where $c < b$, the following inequality holds:*

$$\frac{f+b-c}{b-c} > \frac{f+b}{b}. \quad (12)$$

C. Frequency Spatial Transformer Block

Unlike previous methods [12] and [14], which use a constant high-pass filter and a fixed SRM filter, respectively, to attempt to filter low-frequency features while preserving texture and contour information, or methods [50], [62] only model long-range dependencies based on local features in the spatial domain. Our FSTB integrates information from both the frequency and spatial domains simultaneously. It utilizes different adaptive learning approaches for the two domain features, allowing for a dynamic and targeted understanding and integration of information such as color, texture, edges, spectral characteristics, magnitude, and energy. This entanglement learning approach facilitates the learning of discriminative foreground features from coupled features. As depicted in Fig. 4, the proposed FSTB consists of three key components: Spatial Domain Channel Self-Attention (SDCA), Frequency Domain Bands Self-Attention (FDBA), and Frequency Spatial Feed-forward Network (FSFN). First, SDCA employs a variant of the self-attention mechanism to capture local information in the spatial domain. Second, FDBA combines the attention mechanism with Fourier transformations to extract global representations from the frequency domain. Third, FSFN enhances the information flow between the frequency and spatial domains, facilitating the learning of complementary representations

1) SDCA: In the spatial domain, we design a channel-oriented self-attention variant that can capture long-range interdependencies among channels. This mechanism adaptively recalibrates the responses of each channel's features, thereby enhancing the latent foreground information perceived in the spatial domain. Specifically, as illustrated in Fig. 4, given the input feature $P \in \mathbb{R}^{C\times H\times W}$, we obtain the positional embedding through a 1×1 convolution, following [41]. Then, two depthwise separable dilated convolutions are used to generate the query the $query$ Q_s, key K_s, and $value$ V_s. For example, the calculation process of Q_s is as follows:

$$Q_s = \text{Cat}(\text{DD}_3^{\frac{C}{2}}(\text{Conv}_1^C(P)), \text{DD}_5^{\frac{C}{2}}(\text{Conv}_1^C(P))) \quad (13)$$

Here, $Q_s = \in \mathbb{R}^{C\times H\times W}$ with $C = 256$. DD_k^C and Conv_k^C respectively denote a depthwise separable dilated convolution and a standard convolution, both with $k\times k$ kernel and C output channels. The computation process for K_s and V_s is similar, while their corresponding DD operators and standard convolutions update parameters independently. Subsequently, we obtain the channel attention map A_s by flattened $query$ $\overline{Q}_s \in \mathbb{R}^{C\times N}$ and key $\overline{K}_s \in \mathbb{R}^{C\times N}$, where $N = H\times W$, as follows:

$$A_s = Sof(\overline{Q}_s \odot \overline{K}_s^T) \in \mathbb{R}^{C\times C}, \quad (14)$$

where $Sof(\cdot)$ is the SoftMax function, and $\odot$ is matrix multiplication. Then, the activated attention map $A_s \in \mathbb{R}^{C\times C}$ is used to recalibrate the parameters of flattened $value$ $\overline{V}_s$.

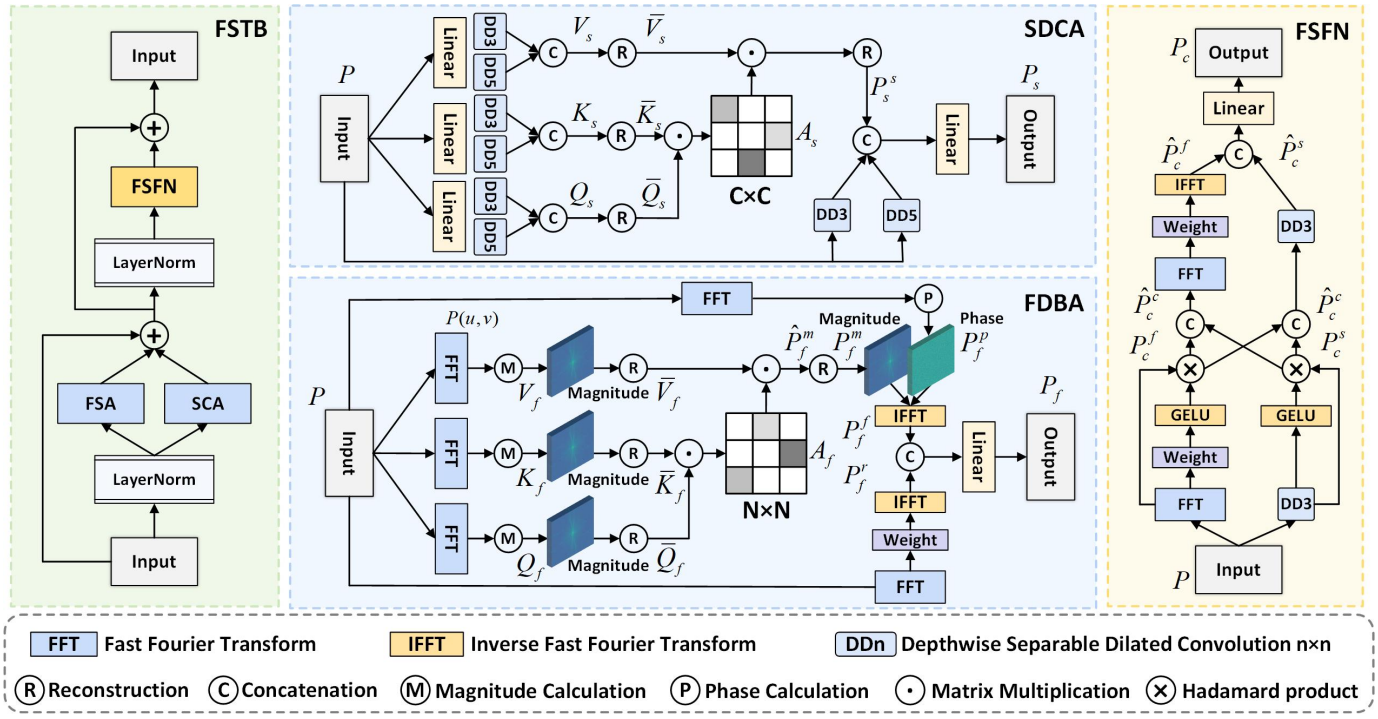

Fig. 4. The detailed flowchart of the proposed core building module FSTB, consisting of three components: Spatial Domain Channel Self-Attention (SDCA), Frequency Domain Bands Self-Attention (FDBA), and Frequency Spatial Feed-forward Network (FSFN).

Finally, to enhance the local information in the spatial domain, we use depthwise separable dilated convolutions to modify the input feature P, and then concatenate the mappings to generate the spatial domain information $P_s \in \mathbb{R}^{C\times H\times W}$, as shown in:

$$P_s = \text{Conv}_1^C(\text{Cat}(P_s^s, P_s^r)) \in \mathbb{R}^{C\times H\times W}, \tag{15}$$

$$P_s^s = \text{Reshape}(\text{Cat}(A_s \odot \overline{V}_s) \in \mathbb{R}^{C\times H\times W}, \tag{16}$$

$$P_s^r = Cat(\text{DD}_3^{\frac{C}{2}}(P), \text{DD}_5^{\frac{C}{2}}(P)) \in \mathbb{R}^{C\times H\times W}, \tag{17}$$

where Reshape$(\cdot)$ refers to the inverse flattening operation, which adjusts the last dimension of the features into its original two-dimensional form suitable for convolution operations. P_s^r is the spatial residual connection feature for providing and supplementing the original spatial information.

Compared to the input feature P, the output feature P_s of SDCA block captures long-range channel dependencies and contains more informative content [27], while also preserving the local details perceived in the spatial domain. This enhancement allows P_s to provide a richer representation that effectively combines both broader contextual information and fine-grained local features.

2) FDBA: Specifically, as shown in Fig. 4, given the input feature $P \in \mathbb{R}^{C\times H\times W}$, the Fourier transform is first applied to obtain its magnitude spectrum, which is then used as the *query* Q_f, *key* K_f, and *value* V_f, as follows:

$$Q_f = K_f = V_f = \mathcal{M}(\mathcal{F}(P)) \in \mathbb{R}^{C\times H\times W}, \tag{18}$$

Here, $\mathcal{F}(\cdot)$ and $\mathcal{M}(\cdot)$ represent the Fast Fourier Transform (FFT) and the magnitude computation formula, as shown in Eq. (1) and Eq. (3). Unlike SDCA, we use the flattened *query* $\overline{Q}_f \in \mathbb{R}^{C\times N}$, *key* $\overline{K}_f \in \mathbb{R}^{C\times N}$, and *value* $\overline{V}_f \in \mathbb{R}^{C\times N}$, where $N = H \times W$, to obtain the spatial attention map A_f, as follows:

$$A_f = Sof(\overline{Q}_f^T \odot \overline{K}_f) \in \mathbb{R}^{N\times N}, \tag{19}$$

Subsequently, we use the spatial attention map to optimize the response values of different frequency bands in the magnitude map $\overline{V}_f$, adaptively enhancing the high-frequency features responsible for textures, as follows:

$$\hat{P}_f^m = A_f \odot \overline{V}_f^T \in \mathbb{R}^{N\times C}. \tag{20}$$

Furthermore, to facilitate the inverse Fourier transform, we reconstruct it to obtain $P_f^m \in \mathbb{R}^{C\times H\times W}$.

On the other hand, as shown in Fig. 2, the phase spectrum is not sensitive to texture and contour information, but it contains shape and structural information and is numerically sensitive [31]. Therefore, we choose to retain it without additional feature extraction modules or other corrective operations, in order to accelerate the convergence of the model, as follows:

$$P_f^p = \mathcal{P}(\mathcal{F}(P)) \in \mathbb{R}^{C\times H\times W}, \tag{21}$$

where $\mathcal{P}(\cdot)$ represents the phase computation operator, as shown in Eq. (4). Then, we perform the inverse Fourier transform using the corrected magnitude spectrum and the original phase spectrum to obtain the corrected frequency domain feature P_f^f, as follows:

$$P_f^f = \mathcal{F}^{-1}(P_f^m \cdot \exp(jP_f^p)) \in \mathbb{R}^{C\times H\times W}, \tag{22}$$

where $\mathcal{F}^{-1}(\cdot)$ denotes the Invert Fast Fourier Transform (IFFT), as shown in Eq. (2). In addition, we propose a

frequency residual connection to enhance the frequency information, and combine the features to obtain the final frequency domain feature P_f, as follows:

$$P_f = \text{Conv}_1^C(\text{Cat}(P_f^f, P_f^r)) \in \mathbb{R}^{C\times H\times W}, \tag{23}$$

$$P_f^r = \mathcal{F}^{-1}(\sigma(\mathcal{F}(P))) \in \mathbb{R}^{C\times H\times W}, \tag{24}$$

where $\sigma(\cdot)$ represents a sequence of operations, including a convolution, batch normalization, a ReLU function, another convolution, and a sigmoid function, utilized to obtain the frequency residual connection feature P_f^r for providing and supplementing the original frequency information.

Compared to the input feature P, the output feature P_f of FDBA block adaptively refines the magnitude spectrum through the self-attention mechanism, with the potential to enhance high-frequency information such as texture and contours in the foreground, while avoiding the introduction of background noise in the high-frequency domain.

3) FSFN: Frequency and spatial features typically focus on different aspects. The frequency domain focuses on the global energy distribution and variations of the signal, while spatial information deals with local pixel-level details and spatial structures. Both provide valuable insights and clues for the overlapping object perception task, thus, the integration method of the two is crucial. In our FSFN, these features are considered as two distinct states, which can undergo entanglement learning during the process to obtain more robust and powerful representations.

Specifically, FSFN consists of two parts. The first stage maps the input feature P to both the frequency and spatial domains, enhancing the model's nonlinear representation capability using the GELU function, as well as employing a gating mechanism to retain global frequency features and local spatial information. The process is as follows:

$$P_c^f = GE(||\sigma(\mathcal{F}(P)\otimes\mathcal{F}(P))||)\otimes||\sigma(\mathcal{F}(P)\otimes\mathcal{F}(P))||, \tag{25}$$

$$P_c^s = GE(\text{DD}_3^C(P))\otimes\text{DD}_3^C(P). \tag{26}$$

Here, $GE(\cdot)$ denotes the GELU function, $\otimes$ represents the Hadamard product, and $||\cdot||$ denotes the modulus operation.

In the second stage, the frequency domain and spatial domain features from the first stage are first concatenated to obtain the joint feature $\hat{P}_c^c$ for feature interaction learning, and the process is as follows:

$$\hat{P}_c^c = \text{Cat}(P_c^s, P_c^f). \tag{27}$$

The feature $\hat{P}_c^c$ is fed into two branches for frequency domain learning and spatial domain learning, respectively. Information from both domains is interactively integrated from the perspectives of global energy perception and local detail perception, resulting in integrated features $\hat{P}_c^f$ and $\hat{P}_c^s$, which emphasize frequency domain and spatial domain information, respectively. Finally, these features are aggregated and the channels are reduced to form the combined feature P_c, as follows:

$$P_c = \text{Conv}_1^C(\text{Cat}(\hat{P}_c^f, \hat{P}_c^s)), \tag{28}$$

$$\hat{P}_c^f = ||\mathcal{F}^{-1}(\sigma(\mathcal{F}(\hat{P}_c^c))\otimes\mathcal{F}(\hat{P}_c^c))||, \tag{29}$$

$$\hat{P}_c^s = \text{DD}_3^C(\hat{P}_c^c). \tag{30}$$

Compared to the input feature P, the output feature P_c of FSFN performs entanglement learning on both spatial and frequency domain features, adaptively selecting and integrating them. It leverages the advantages of global frequency and local spatial information, resulting in a more comprehensive feature representation.

IV. Experiments

A. Implementation Details

For fair comparisons, we train all models under the same conditions with the ImageNet [63] pretrained backbones, including ResNet-50 [64], ResNet-101 [64], ResNeXt-101 [65], and Swin-L [50]. CNN-based models are trained with SGD optimizer, using a learning rate of 0.01, momentum of 0.9, and weight decay of 0.1. Transformer-based models use the AdamW optimizer with a learning rate of 0.0001 and weight decay of 0.0001. All models are trained for 12 epochs and implemented using the MMDetection3.1.0 framework [66], with an image size of 320×320. All training is performed on a consistent computing platform equipped with an NVIDIA GeForce RTX 4090 GPU, an Intel Core i9-13900K CPU, 64 GB of memory, Windows 10 OS, and PyTorch 1.13.1. For the warm-up scheme in convolutional models, the learning rate increases linearly over the first 500 iterations with a warm-up ratio of 0.001. After this warm-up phase, the learning rate decreases stepwise, with adjustments made at the 8th and 11th epochs. For DETR-like models, following the approach of Deformable-DETR, the learning rate is reduced by a factor of 0.1 at the 11th epoch.

B. Datasets and Evaluation Metrics

1) Datasets: The PIXray [1] dataset is capable of performing both object detection and instance segmentation tasks, referred to as PIXray-det and PIXray-seg, respectively. It includes 5046 X-ray images of prohibited items, divided into 4046 training images and 1000 testing images. It encompasses 15 categories of prohibited items, including Gun, Knife, Lighter, Battery, Pliers, Scissors, Wrench, Hammer, Screwdriver, Dart, Bat, Fireworks, Saw blade, Razor blade, and Pressure vessel.

The OPIXray [2] is a fine-grained prohibited item dataset for sharp-edged tools, comprising 8885 X-ray images of prohibited items, allocated into 7019 training images and 1776 testing images. It includes five types of knives: Folding Knife (FO), Straight Knife (ST), Scissor (SC), Utility Knife (UK), Multi-tool Knife (MU).

The RSNA Lung Opacities (RLP) dataset is a subset of the pneumonia category data with fine-grained location labels that we have filtered from the RSNA Pneumonia Detection Challenge dataset [3]. The original dataset had non-standardized labels. The RLP dataset is suitable for object detection training tasks. The dataset contains 6,011 X-ray images of pulmonary pneumonia, with 4,009 images used for training and 1,202 images used for testing.

TABLE I
GENERALIZATION ANALYSIS FOR FOAM ON PIXRAY-DET [1] DATASET FOR PROHIBITED ITEM DETECTION.

Method	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{box}_{S}	AP^{box}_{M}	AP^{box}_{L}
Deformable-DETR [45]	36.6	66.4	37.1	5.6	24.2	45.8
$\mathcal{F}$-Deformable-DETR	**42.3**	70.3	45.9	5.5	27.8	50.8
DINO [57]	64.3	86.5	71.0	19.3	48.9	73.9
$\mathcal{F}$-DINO	**67.2**	88.3	74.8	21.8	52.6	76.5
RT-DETR [69]	61.4	84.0	68.5	20.0	48.0	70.1
$\mathcal{F}$-RT-DETR	**61.7**	84.3	68.6	22.5	48.1	70.9
Mask2Former† [49]	41.8	61.8	45.3	8.2	25.6	51.7
$\mathcal{F}$-Mask2Former†	**43.0**	63.6	46.6	8.0	25.1	53.0
CondInst† [46]	53.4	80.6	60.6	12.7	39.3	62.4
$\mathcal{F}$-CondInst†	**54.4**	80.8	61.7	12.5	38.1	63.5
Cascade-Mask-R-CNN† [47]	70.2	88.9	78.9	21.2	58.5	78.1
$\mathcal{F}$-Cascade-Mask-R-CNN†	**70.7**	88.8	79.9	20.1	58.8	78.6
Mask-R-CNN (X-101)† [70]	65.2	88.6	76.6	11.5	54.4	73.2
$\mathcal{F}$-Mask-R-CNN(X-101)†	**65.6**	88.6	77.4	11.5	54.5	73.7
DINO (Swin-L) [57]	73.3	90.2	80.7	39.4	58.7	80.9
$\mathcal{F}$-DINO (Swin-L)	**73.7**	90.8	80.3	39.5	60.8	81.5

† indicates models trained by both "bounding box" and "segmentation" labels, following [47], [70].
The default backbone of models is ResNet-50 [64]. X-101 and Swin-L stand for ResNeXt-101 [65] and Swin-Transformer-Large [50] backbone, respectively.

2) Evaluation Metrics: For the PIXray-det dataset and the RLP dataset, we apply the COCO [67] evaluation metrics. The main challenge metric is the box average precision (AP^{box}), which is calculated across 10 Intersection over Union (IoU) thresholds from 0.5 to 0.95, with a step of 0.05. Specifically, AP^{box}_{50} indicates the mean average precision at an IoU threshold of 0.5, while AP^{box}_{75} refers to the mean average precision at an IoU threshold of 0.75. In addition, AP^{box}_{S}, AP^{box}_{M}, and AP^{box}_{L} correspond to the mean average precision for small objects (area $< 32^2$), medium objects ($32^2 <$ area $< 96^2$), and large objects ($96^2 <$ area), respectively.

For the OPIXray dataset, we use the VOC [68] evaluation metric. The average precision (AP) for each category is derived from the area under the Precision-Recall curve at an IoU threshold of 0.5. The mean average precision (mAP) is calculated by averaging the AP values across all categories. This mAP acts as a comprehensive evaluation metric, reflecting both the accuracy and recall of the detector, and offering a holistic view of its overall performance, including its strengths and weaknesses.

For the PIXray-seg dataset, the primary challenge metric is mask average precision (AP^{mask}), which is utilized to assess the performance of instance segmentation models comprehensively. It calculates the average precision (AP) based on different Intersection over Union (IoU) thresholds by computing the IoU between the predicted masks and the ground-truth masks. The distinction between it and AP^{box} is merely in the objects that the IoU threshold is applied to. While AP^{box} calculates the average precision using the IoU between bounding boxes, AP^{mask} uses the IoU between predicted and ground truth masks, calculated pixel-wise.

C. Generalization

1) Models and Backbones: In this part, we first demonstrate the powerful architectural and model generalization capabilities of FOAM by applying it to various object detectors under two advanced architectures. We then select representative models with strong backbones, including the convolutional-based ResNeXt-101 [65] and the transformer-based Swin-L [50], to further validate the backbone generalization ability of FOAM. As shown in Tab. I, FOAM achieves box AP gains of 4.3%, 2.9%, 0.3%, and 1.2% for Deformable-DETR [45], DINO [57], RT-DETR [69], and Mask2Former [49] on the PIXray-det [1] dataset, respectively, highlighting its effectiveness for emerging and advanced Deformable-DETR-based models. Similarly, FOAM improves the box AP by 0.5%, 0.4%, and 1.0% for Cascade-Mask-R-CNN [47], Mask-R-CNN [70], and CondInst [46], respectively, demonstrating its effectiveness for traditional models based on the fully convolutional architecture. Finally, FOAM boosts the box AP by 2.9% and 0.4% for DINO with ResNet-50 and Swin-L backbones, and by 0.4% for Mask-R-CNN with ResNeXt-101, illustrating its strong generalization capability across both CNN- and Transformer-based backbones.

D. Comparison with SOTA Models

To further validate the improvement in overlapping object perception brought by the supplementary frequency domain information from FOAM, we leverage FOAM across four datasets to challenge state-of-the-art algorithms on three overlapping object perception tasks, including Prohibited Item Detection, Prohibited Item Segmentation, and Pneumonia Detection.

1) Experiments over Prohibited Item Detection: We challenge state-of-the-art models in the prohibited item detection domain on the PIXray-det [1] and OPIXray [2] datasets.

The quantitative results for the PIXray-det [1] dataset are presented in Tab. II. Notably, with the same ResNet-50 backbone, $\mathcal{F}$-DINO, which is the combination of FOAM and DINO, outperforms other DINO-based improved prohibited item detectors, including AO-DETR [6], $\mathcal{M}$-DINO [58], and $\mathcal{C}$-DINO [59], in terms of accuracy. This demonstrates that the FOAM architecture enhances the model's ability to perceive foreground features from overlapping scenes more effectively than other methods in the prohibited item detection domain, such as CSA [6], MMCL [58], and CSPCL [59]. The version of the small-scale model, $\mathcal{F}$-DINO (ResNet-50) achieves the best performance with a box AP of 67.2%, among models with comparable parameters and FLOPs. For the version of the large-scale model, $\mathcal{F}$-DINO (Swin-L) achieves 73.7% box AP, surpassing both general object detectors and specialized prohibited item detectors.

In addition, we evaluate the performance of FOAM on the fine-grained sharp-edged tools dataset, OPIXray [2]. As shown in Tab. III, $\mathcal{F}$-DINO achieves higher mAP (79.8%) than AO-DETR, $\mathcal{M}$-DINO, and $\mathcal{C}$-DINO, demonstrating that the additional frequency domain cues provided by FOAM enable the model to more effectively understand and distinguish subtle foreground differences in overlapping scenes compared

TABLE II
COMPARISON WITH STATE-OF-THE-ART OBJECT DETECTORS ON PIXRAY-DET [1] DATASET.

Method	Backbone	FPS	PARAMs (M)	FLOPs (G)	# queries	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{box}_{S}	AP^{box}_{M}	AP^{box}_{L}
General Object Detectors											
Faster R-CNN [71]	ResNeXt-101	70	59.83	28.35	*	53.6	82.3	60.8	3.9	37.7	62.7
Cascade R-CNN [72]	ResNet-50	39	69.20	60.99	*	61.0	83.9	69.0	10.4	46.8	69.7
ATSS [73]	ResNet-101	66	51.14	27.82	*	52.8	80.8	60.2	7.0	37.4	63.6
GFLv1 [74]	ResNeXt-101	66	50.70	28.51	*	57.5	82.8	66.0	9.1	42.0	67.4
DINO [57]	ResNet-50	54	58.38	26.89	30	64.3	86.5	71.0	19.3	48.9	73.9
RT-DETR [69]	ResNet-50	64	42.81	17.07	60	62.3	85.3	69.9	25.6	48.0	70.9
DINO [57]	Swin-L	40	229.0	156.0	30	73.3	90.2	**80.7**	39.4	58.7	80.9
Prohibited Item Detectors											
AO-DETR [6]	ResNet-50	54	58.38	26.89	30	65.6	86.1	72.0	23.9	50.7	74.8
$\mathcal{M}$-DINO [58]	ResNet-50	54	58.38	26.89	30	66.7	87.5	74.4	23.5	50.7	75.5
$\mathcal{M}$-RT-DETR [58]	ResNet-50	64	42.81	17.07	60	63.6	85.9	71.4	24.0	49.9	72.6
$\mathcal{C}$-DINO [59]	ResNet-50	54	58.38	26.89	30	66.4	86.8	73.6	25.7	50.9	75.6
$\mathcal{C}$-RT-DETR [59]	ResNet-50	64	42.81	17.07	30	61.8	84.3	68.7	25.2	47.7	70.6
$\mathcal{F}$-DINO (Ours)	ResNet-50	38	59.75	30.79	30	67.2	88.3	74.8	21.8	52.6	76.5
$\mathcal{F}$-DINO (Ours)	Swin-L	29	230.37	171.93	30	**73.7**	**90.8**	80.3	**39.5**	**60.8**	**81.5**

TABLE III
COMPARISON WITH STATE-OF-THE-ART OBJECT DETECTORS ON OPIXRAY [2] DATASET. FO, ST, SC, UT, AND MU REPRESENT FOLDING KNIFE, STRAIGHT KNIFE, UTILITY KNIFE, AND MULTI-TOOL KNIFE, RESPECTIVELY.

Method	Backbone	FPS	PARAMs (M)	FLOPs (G)	# queries	mAP	FO	ST	SC	UT	MU
General Object Detectors											
Faster R-CNN [71]	ResNeXt-101	70	59.83	28.35	*	73.4	80.6	45.4	89.1	69.1	83.1
Cascade R-CNN [72]	ResNet-50	39	69.20	60.99	*	76.9	83.8	58.8	90.0	73.2	78.8
ATSS [73]	ResNet-101	66	51.14	27.82	*	67.5	72.8	38.0	88.6	58.0	80.2
GFLv1 [74]	ResNeXt-101	66	50.70	28.51	*	75.6	80.0	53.6	89.3	71.7	83.4
DINO [57]	ResNet-50	54	58.38	30.79	30	78.2	83.2	58.8	89.4	72.7	86.7
RT-DETR [69]	ResNet-50	64	42.81	17.07	320	61.8	61.1	26.0	88.6	56.4	76.8
DINO [57]	Swin-L	40	229.0	156.0	30	80.0	84.2	61.1	89.0	**78.9**	86.6
Prohibited Item Detectors											
AO-DETR [6]	ResNet-50	54	58.38	26.89	30	79.2	83.8	60.5	90.1	74.7	87.1
$\mathcal{M}$-DINO [58]	ResNet-50	54	58.38	26.89	30	78.6	83.9	57.2	**90.4**	74.2	87.1
$\mathcal{M}$-RT-DETR [58]	ResNet-50	64	42.81	17.07	320	62.5	65.9	22.3	86.4	57.1	80.7
$\mathcal{C}$-DINO [59]	ResNet-50	54	58.38	26.89	30	77.9	82.8	56.0	89.9	74.2	86.7
$\mathcal{C}$-RT-DETR [59]	ResNet-50	64	42.81	17.07	30	70.1	76.0	34.4	88.6	67.4	84.3
$\mathcal{F}$-DINO (Ours)	ResNet-50	38	59.75	30.79	30	79.8	84.3	62.3	89.9	74.9	87.5
$\mathcal{F}$-DINO (Ours)	Swin-L	29	230.37	171.93	30	**81.7**	**86.4**	**65.2**	89.3	78.6	**89.0**

to other anti-overlapping strategies. For the version of the large-scale model, $\mathcal{F}$-DINO (Swin-L) achieves an mAP of 81.7%, exceeding other state-of-the-art models in both the fields of general object detection and prohibited item detection.

2) Experiments over Prohibited Item Segmentation: To explore the effectiveness of FOAM in instance segmentation tasks under overlapping scenes. We challenge state-of-the-art models in prohibited item segmentation on the PIXray-seg [1] dataset. As shown in Tab. IV, Mask-R-CNN (ResNeXt-101) [70] achieves a mask AP of 55.2%, outperforming other instance segmentation models, such as Cascade-Mask-R-CNN [47]. Building upon this, $\mathcal{F}$-Mask-R-CNN further improves the mask AP to 55.8%, surpassing other instance segmentation models, including DETR-based Mask2Former [49] and fully convolutional models such as SOLO [75], SOLOv2 [76], and CondInst [46].

3) Experiments over Pneumonia Detection: To further explore the application of FOAM, we are the first to apply anti-overlapping detection techniques in the medical diagnostic field. Specifically, we applied FOAM to the pneumonia detection dataset, RLP, based on X-ray images. As shown in Tab. V, traditional multi-stage or two-stage object detectors, such as Cascade-R-CNN and Faster-R-CNN, achieve better accuracy compared to single-stage detectors like ATSS and GFLv1, and outperform Deformable-DETR-based models like RT-DETR and DINO, which perform better on the general object detection dataset COCO [67]. Therefore, we utilize the Cascade-R-CNN as the baseline model, and observe that under the influence of FOAM, $\mathcal{F}$-Cascade-R-CNN improves the box AP from 19.7% to 20.2%, demonstrating the strong generalization ability of FOAM.

E. Ablation Study

In this part, to optimize the proposed method FOAM, we conducted extensive ablation experiments using DINO as the baseline on the PIXray-det dataset.

TABLE IV
COMPARISON WITH STATE-OF-THE-ART INSTANCE SEGMENTATION MODELS ON PIXRAY-SEG [1] DATASET.

Instance Segmentation	Backbone	FPS	PARAMs (M)	FLOPs (G)	# queries	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}_{S}	AP^{mask}_{M}	AP^{mask}_{L}
Mask2Former† [49]	ResNet-50	12	176.11	26.79	100	37.7	66.1	37.2	6.3	18.8	50.2
SOLO [75]	ResNet-50	38	36.15	51.23	*	33.5	65.0	31.0	0.8	15.2	44.3
SOLOv2 [76]	ResNet-50	24	46.29	86.37	*	35.2	68.5	32.3	0.1	16.6	47.2
CondInst† [46]	ResNet-50	23	34.01	33.60	*	30.3	62.5	26.8	1.9	14.9	37.8
Cascade-Mask-R-CNN† [47]	ResNet-50	44	77.08	1271.85	*	55.1	85.7	60.2	**9.5**	37.3	62.8
Mask-R-CNN† [70]	ResNeXt-101	30	107.31	110.12	*	55.2	85.8	61.1	3.9	38.0	63.5
$\mathcal{F}$-Mask-R-CNN† (Ours)	ResNeXt-101	26	109.48	112.08	*	**55.8**	**86.8**	**61.5**	5.3	**39.2**	**63.9**

† indicates models trained by both "bounding box" and "segmentation" labels, following [70], [72].

TABLE V
COMPARISON WITH STATE-OF-THE-ART OBJECT DETECTORS ON RLP DATASET.

Method	Backbone	FPS	PARAMs (M)	FLOPs (G)	# queries	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{box}_{S}	AP^{box}_{M}	AP^{box}_{L}
RT-DETR [69]	ResNet-50	64	42.81	17.07	60	14.7	38.9	8.0	2.6	12.7	21.3
DINO [57]	ResNet-50	54	58.38	26.89	30	15.6	44.7	6.6	5.4	13.1	25.0
DINO [57]	Swin-L	40	229.0	156.0	30	16.3	44.6	7.8	4.4	12.7	25.9
ATSS [73]	ResNet-101	66	51.14	27.82	*	16.1	48.9	5.0	0.3	12.7	25.0
GFLv1 [74]	ResNeXt-101	66	50.70	28.51	*	10.2	35.4	1.7	0.0	9.5	14.8
Faster R-CNN [71]	ResNeXt-101	70	59.83	28.35	*	18.6	55.7	7.0	**4.1**	15.9	26.3
Cascade-R-CNN [72]	ResNet-50	39	69.20	60.99	*	19.7	**56.8**	8.0	2.5	16.3	28.3
$\mathcal{F}$-Cascade-R-CNN (Ours)	ResNet-50	36	70.36	72.96	*	**20.2**	56.7	**8.6**	2.1	**16.4**	**29.2**

TABLE VI
ABLATION STUDY OF HDC AND FSTB RESULTS. PARAMS, FLOPS, AND FPS REPRESENT THE TOTAL NUMBER OF PARAMETERS, FLOATING POINT OPERATIONS, AND THE NUMBER OF INFERENCES THE MODEL CAN PERFORM PER SECOND, RESPECTIVELY.

N	FSTB	HDC	PARAMs(M)	FLOPs(G)	FPS	AP^{box}
0	✗	✗	58.380	26.820	54	64.3
1	✓	✗	59.746	30.791	38	66.0
	✓	✓	59.746	30.791	38	67.2
2	✓	✗	61.111	34.762	30	66.4
	✓	✓	61.111	34.762	30	67.3
3	✓	✗	62.477	38.732	23	66.9
	✓	✓	62.477	38.732	23	67.7

TABLE VII
COMPARISON OF GAUSSIAN BLURRING (GB), DOWNSAMPLING AND UPSAMPLING (DU), AND GAUSSIAN NOISE (GN) CORRUPTION STRATEGIES FOR THE HDC MECHANISM. KS MEANS THE KERNEL SIZE OF GAUSSIAN NOISE

Method	Parameter	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{box}_{S}	AP^{box}_{M}	AP^{box}_{L}
–	–	64.3	86.5	71.0	19.3	48.9	73.9
(a) DU	×2	65.6	87.1	72.9	18.7	50.9	75.4
	×3	65.3	86.9	72.5	20.6	50.2	74.7
	×4	**66.5**	88.0	73.3	25.0	51.9	75.9
	×5	64.8	86.6	71.8	19.5	50.2	74.3
(b) GN	0.1	65.7	87.3	73.5	21.1	51.4	75.1
	0.2	**66.4**	87.9	73.8	23.0	52.3	75.6
	0.5	66.1	87.7	72.7	21.7	50.8	75.9
	1	65.4	87.5	71.9	21.5	51.5	74.8
(c) GB (ks=3)	0.1	66.6	87.9	73.5	21.2	51.9	75.7
	1	66.9	88.5	73.8	23.6	51.1	76.4
	5	**67.2**	88.3	74.8	21.8	52.6	76.5
	10	66.5	88.2	73.9	23.0	50.8	76.2
(d) GB (ks=5)	0.1	63.9	86.0	70.6	21.8	48.4	73.0
	1	66.4	87.5	73.6	22.0	51.9	75.9
	5	**66.5**	87.7	74.2	21.3	51.6	76.1
	10	65.5	86.9	72.0	22.1	50.4	75.0

The studied hyperparameters that control the degradation level of each corruption strategy are (a) the Downsampling scale factor, (b) the Gaussian Noise sigma, and (c) the Gaussian Blurring sigma.

1) Ablation study for HDC and FSTB: Tab. VI presents a complex ablation study that thoroughly evaluates the effects of the Frequency Spatial Transformer Block (FSTB), the number of iterations N of cascaded FSTBs for corruption and base branches, and the Hierarchical De-Corrupting (HDC) mechanism on the model's performance. When the number of iterations $N = 1$, the FSTB is able to increase the model's box AP from 64.3% to 66.0%. Furthermore, when both the HDC and FSTB are utilized together, the box AP reaches 67.2%, indicating good compatibility between the two methods and demonstrating that the HDC mechanism effectively guides the FSTB to achieve a stronger anti-overlapping detection capability. Additionally, the PARAMs and FLOPs only increased by 1.366 M and 3.971 G, while FPS decreased by 16 frames. This suggests that the method is not demanding in terms of computational resources. Similarly, when we increase the number of iterations N to 2 and 3, FSTB and HDC continue to show good compatibility, and the model's box AP improved further, albeit with diminishing returns. To balance computational complexity and performance, we set N to 1 for subsequent experiments.

2) Ablation study for corruption strategies: As shown in Tab. VII, to further investigate the effects of different corruption strategies, such as Gaussian Blurring (GB), Downsampling and Upsampling (DU), and Gaussian Noise (GN), on

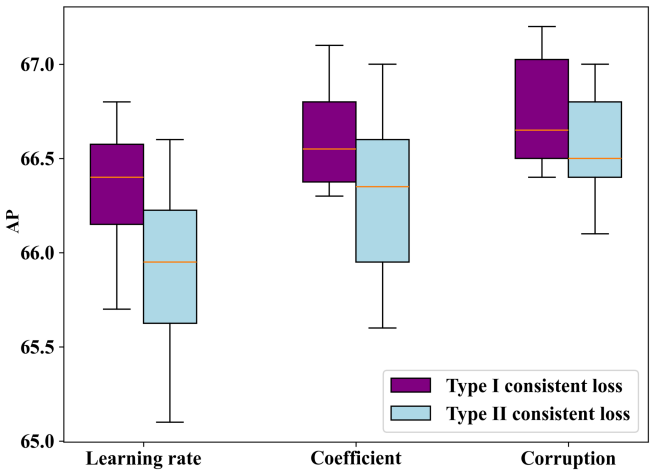

Fig. 5. Comparison of Type I consistent Loss and Type II consistent Loss. The standard boxplot illustrates the relationship between box AP and variations in the learning rates, coefficients of consistent loss, and corruption strategies.

TABLE VIII
ABLATION STUDY OF THE TARGET LAYER SET $\mathcal{L}$ UNDER THE GAUSSIAN BLURRING STRATEGY (KERNEL SIZE IS 5, SIGMA IS 5).

$l=1$	$l=2$	$l=3$	$l=4$	AP^{box}	AP^{box}_{50}	AP^{box}_{75}
✗	✗	✗	✗	64.3	86.5	71.0
✗	✗	✗	✓	67.1	88.1	74.8
✗	✗	✓	✓	**67.2**	88.3	74.8
✗	✓	✓	✓	66.9	88.0	74.3
✓	✓	✓	✓	65.2	87.8	72.7

TABLE IX
ABLATION STUDY OF SDCA AND FDBA.

ID	SDCA	FDBA		AP^{box}	AP^{box}_{50}	AP^{box}_{75}
	Shape of A_s	Shape of A_f	Target			
(a)	C×C	N×N	Magnitude	**67.2**	88.3	74.8
(b)	N×N	N×N	Magnitude	67.0	87.8	74.5
(c)	C×C	C×C	Magnitude	66.0	87.2	74.6
(d)	C×C	C×C	Phase	65.2	86.5	73.7
(e)	C×C	N×N	Phase	65.8	87.2	74.2

A_s means channel attention map in Eq. (14), and A_f means spatial attention map in Eq. (19).

the guidance provided by the HDC mechanism for enhancing the texture and contour feature perception capabilities of FSTB, we compare the performance of the $\mathcal{F}$-DINO model on the PIXray-det dataset under various hyperparameters for each strategy. For the DU strategy, we adjust the Downsampling scale factor and find that a factor of "×4" achieves the highest 66.5% box AP compared to other coefficients. Under the Gaussian Noise (GN) strategy, when the Gaussian Noise sigma is set to 0.2, the model gains an increase of 2.1% box AP. In the case of the GB strategy, the models' overall performances are superior when the Gaussian kernel size (ks) is set to 3 as opposed to 5. Furthermore, when $ks = 3$ and the Gaussian Blurring sigma is set to 5, the model achieved its highest box AP of 67.2%. Overall, the GB strategy provides the most positive impact for the anti-overlapping feature-awareness capability of the FSTB module.

3) Ablation study for consistent losses: As shown in Fig. 5, we conduct three sets of experiments, each using different learning rates, loss weighting coefficients, and corruption strategies to compare the impact of Type I consistent loss, as defined in Eq. (8), versus Type II consistent loss, as defined in Eq. (11), on the box AP of $\mathcal{F}$-DINO on the PIXray dataset, whose results are presented as the boxplot. All experiments are conducted with the same batch size, learning strategy, data augmentation strategy, and optimizer to exclude potential interference from other factors. The experimental results are progressively improved, as each subsequent experimental group uses the best parameter settings from the previous group. For example, the corruption control group uses the optimal learning rate and coefficient settings from the learning rate and coefficient control groups, respectively. It can be observed that although both loss functions positively impact the model's accuracy, which exceeds the baseline box AP of 64.3%, the mean and maximum of the outputs for Type I consistent loss based on KL divergence are higher, while its variance is overall lower compared to Type II consistent loss based on MSE. This discrepancy is attributed to the fact that Type I loss emphasizes the relative difference between F_C^{n+1} and F_O^n, whereas Type II loss focuses on the absolute difference between the two. Therefore, Type I loss is more robust and effective for fine-grained feature-level alignment tasks.

4) Ablation study for target layer set: Since $F_C^{n+1} = \{F_{C,l}^{n+1}\}_{l=0}^{L}$ and $F_O^n = \{F_{O,l}^n\}_{l=0}^{L}$ are multi-scale features with $L = 4$ layers, we conduct experiments on $\mathcal{F}$-DINO to investigate which feature layers in Eq. (8) need to be aligned using the consistent loss mechanism in the HDC mechanism to better guide the FSTB. The results show that the model achieves higher accuracy when consistency guidance is applied for higher-level features. This can be attributed to the fact that higher-level features primarily capture global information, which is more adept at encoding semantic information. In contrast, lower-level features tend to focus on local information, often accompanied by redundant data and noise. The presence of ineffective information from these lower-level features disrupts the FSTB's ability to comprehend the reverse process of corruption during consistency training, thereby reducing its ability to extract features in the presence of overlap.

5) Ablation study of SDCA and FDBA: To enable the FSTB to optimize and maximize the integration of local spatial features and global frequency domain information, we conduct a series of ablation experiments, as shown in Tab. IX. By comparing Tab. IX(a) and Tab. IX(b), we observe that the Spatial Domain Channel Self-Attention (SDCA) mechanism is better suited for using a channel attention matrix of the form $A_s \in \mathbb{R}^{C\times C}$, rather than the classical spatial domain feature reorganization with an attention map of the form $A_s \in \mathbb{R}^{N\times N}$ [41], where $N = H \times W$ represents the flattened spatial scale. By comparing Tab. IX(c) and Tab. IX(d), we find that the Frequency Domain Bands Self-Attention (FDBA) mechanism is better suited for adaptively correcting the magni-

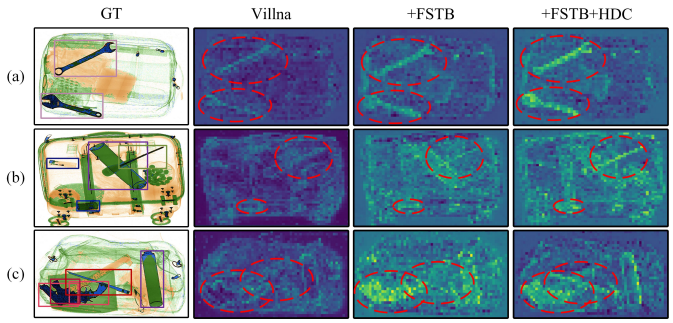

Fig. 6. Visualization of feature maps of "DINO", "DINO+FSTB", and "DINO+FSTB+HDC" ($\mathcal{F}$-DINO).

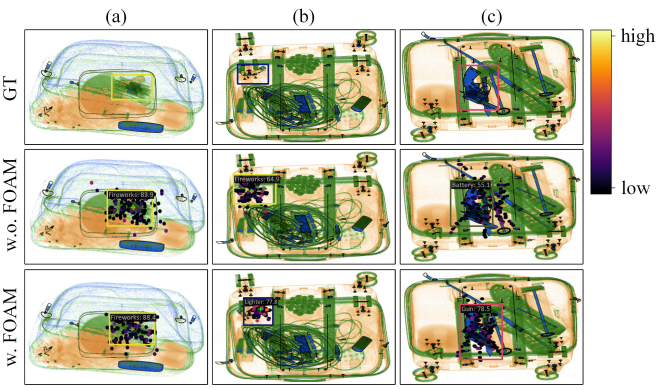

Fig. 7. Visualization of deformable attention sampling points, reference points, and prediction results for corresponding content query in the last decoder layer. Row "GT", "w.o. FOAM", and "w. FOAM" refers to the ground truth, and results of DINO and $\mathcal{F}$-DINO. Column "(a)-(c)" represents images of PIXray-det [1] dataset. Each sampling point is shown as a filled circle, with color indicating its attention weight, and the reference point is marked by a green cross.

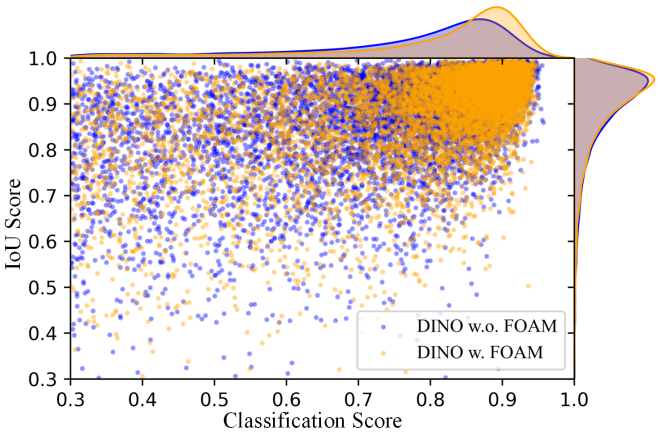

Fig. 8. The scatter plot and Kernel Density Estimation (KDE) joint distribution plot of the prediction results from the final decoder layer. Blue and Orange refer to the results of DINO and $\mathcal{F}$-DINO, respectively.

tude spectrum, which is more effective at representing texture and contour information, rather than the phase spectrum, which excels at capturing shape and structural information. Finally, by comparing Tab. IX(d) with Tab. IX(e), or Tab. IX(a) with Tab. IX(c), we observe that for FDBA, the classical spatial attention mechanism with $A_f \in \mathbb{R}^{N\times N}$ is more suitable. This is because this approach is better at integrating effective information from both the high-frequency and low-frequency bands of the magnitude spectrum.

Overall, the configuration in Tab. IX(a), as described in Sec. III-C, achieves the best accuracy. In this configuration, SDCA preserves the local features that spatial domain features excel at, while also capturing long-range channel dependencies and containing more informative content [27]. Meanwhile, FDBA integrates both high-frequency and low-frequency information from the magnitude spectrum, particularly regarding texture and contour. By leveraging the complementary characteristics of both feature domains, the system incorporates both global and local features, providing high-quality input features for the subsequent coupling learning in FSFN.

F. Visualization and Analysis

In this part, we first use the state-of-the-art model, DINO, on PIXray-det as the baseline and conduct a progressive analysis of the impact of FOAM across three levels: feature extraction by the backbone network, feature extraction in the decoder, and high-order statistical analysis of the final inference results. This includes visualizing feature maps to assess how FOAM influences the feature extraction process of the backbone, examining the decoder of the last layer via sampling and reference points, and utilizing scatter plots to analyze classification and localization results. Finally, we visualize the prediction results of the SOTA models on four datasets for the three tasks, qualitatively comparing the impact of FOAM on the prediction outcomes of models.

1) Feature maps: Fig. 6 compares the feature maps of the DINO, DINO+FSTB, and DINO+FSTB+HDC ($\mathcal{F}$-DINO) models, all of which use the Swin-L backbone network. In row (a), the feature map generated by DINO captures incomplete "Wrench" features, with low response values and unclear contours. In contrast, the DINO+FSTB model demonstrates a higher feature response to the "Wrench". Furthermore, the feature map of $\mathcal{F}$-DINO shows greater contrast between the foreground and background, with clearer contours, indicating the positive guidance effect of the Hierarchical De-Corrupting (HDC) mechanism in enhancing FSTB's ability to perceive features in the presence of overlap. Similarly, in row (b), the feature map of DINO fails to focus more on the prohibited items, such as the "Battery" and "Screwdriver", compared to the background. However, in the DINO+FSTB and $\mathcal{F}$-DINO feature maps, this focus gradually improves. Finally, in row (c), DINO+FSTB demonstrates better attention to the heavily overlapping "Gun" and "Hammer" features compared to the baseline model. Additionally, $\mathcal{F}$-DINO, building upon this, further enhances attention to the "Saw" while reducing the focus on background objects.

Overall, the frequency domain features introduced by FSTB help the backbone network to more comprehensively perceive foreground features. The HDC mechanism essentially further directs FSTB to prioritize foreground feature attention while suppressing background feature focus, manifested as an enhancement in the perception of foreground contour, thereby

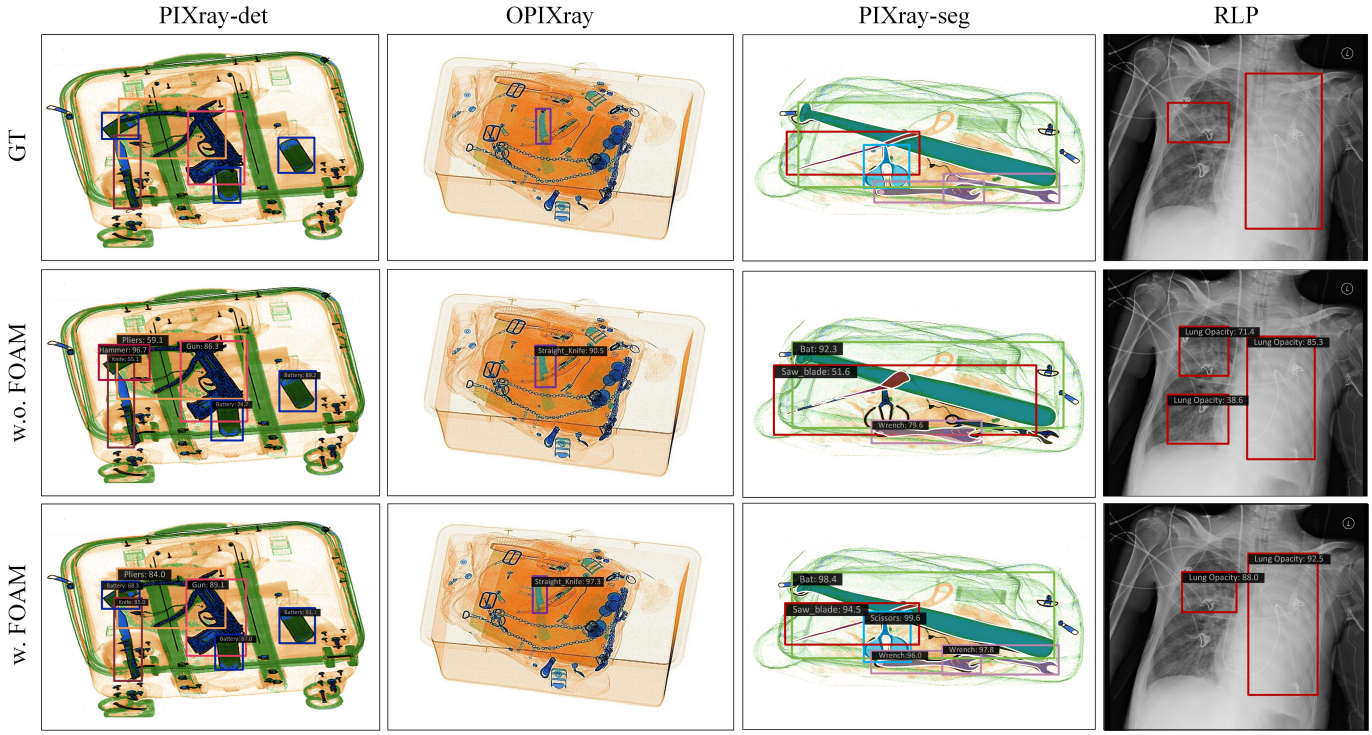

Fig. 9. Prediction results on PIXray-det, OPIXray, RLP, PIXray-seg datasets. We select state-of-the-art models from four datasets across three tasks (row two) and apply FOAM to enhance their performance (row three). These models include DINO (Swin-L) on PIXray-det and OPIXray, Mask-R-CNN on PIXray-seg, and Cascade-R-CNN on RLP. In comparison, the models enhanced by FOAM achieve superior predictive accuracy.

improving the model's resistance to feature overlap.

2) Sampling points and reference points: We take DINO as a representative of the Deformable-DETR series models to explore how the reference point and sampling points of the final decoder layer, directly related to the final detection results, change in response to specific images under the influence of the proposed FOAM. As shown in Fig. 7, column (a) indicates that the sampling points of $\mathcal{F}$-DINO are more concentrated on the prohibited item "Fireworks", while in the baseline model DINO, the high-confidence sampling points focus more on background features. Further, in column (b), when faced with the weakly featured prohibited item "Lighter", severely disrupted by the background, the reference point of DINO accurately locates the target. However, the sampling points are not sufficiently concentrated, capturing a large amount of background features, which leads to the misclassification of the "Lighter" as "Fireworks". In contrast, $\mathcal{F}$-DINO's sampling points are not only more focused on the "Lighter" itself but also exhibit higher sampling confidence. This means that the "Lighter" features contribute more significantly to the model's final decision, resulting in a correct classification with high confidence and accurate localization [45]. In column (c), when faced with the overlapping and closely positioned "Gun" and "Knife", DINO's reference point and sampling points tend to focus on both the "Knife" and "Gun", leading to an incorrect detection result of "Battery". In contrast, reference point and sampling points of $\mathcal{F}$-DINO are focused on the "Gun" itself, resulting in the correct detection of the "Gun".

Overall, the frequency domain features supplemented by the FOAM backbone help the decoder focus on and extract foreground features from overlapping scenes, leading to more accurate predictions.

3) The scatter diagram of prediction results: To assess the effectiveness of the FOAM mechanism for prediction results, we visualize the IoU and classification scores of DINO and $\mathcal{F}$-DINO (DINO with our FOAM) predictions on the PIXray-det dataset, as shown in Fig. 8. We plot a scatter diagram of prediction results with classification and IoU scores above 0.3, along with Kernel Density Estimation (KDE) curves. Blue and orange represent the results of DINO and $\mathcal{F}$-DINO, respectively. The orange points are more concentrated, significantly shifted further to the right, and slightly moved upwards compared to the blue points, indicating that under FOAM, the model leverages both frequency and spatial domain cues to capture more classification semantic information, such as textures, as well as localization information, such as contours. This improvement enables the model to more effectively perceive and extract foreground information from overlapping features in complex scenes, thereby enhancing prediction accuracy for overlapping object perception.

4) Prediction results: Fig. 9 illustrates the qualitative impact of FOAM on the prediction results of SOTA models across four datasets for three tasks.

Prohibited Item Detection Task: On the PIXray-det dataset, DINO (Swin-L) misclassifies the "Battery" in the top-left corner as a Hammer, and the localization result of the Pliers is disrupted by interference from the "Gun". Similarly, on the OPIXray dataset, the localization of the "Straight Knife" is compromised by background features. In contrast, under the influence of FOAM, $\mathcal{F}$-DINO demonstrates more

accurate localization and classification for the prohibited item detection task, with higher confidence in its predictions.

Prohibited Item Segmentation Task: On the PIXray-seg dataset, Mask-R-CNN fails to detect the "Wrench" and "Scissors" and provides incomplete segmentation for the "Saw". In contrast, the instance segmentation results of $\mathcal{F}$-Mask-R-CNN are nearly identical to the ground truth.

Pneumonia Detection Task: On the RLP dataset, Cascade-R-CNN erroneously detects the background as "Lung Opacity", and its localization is imprecise. In comparison, $\mathcal{F}$-Cascade-R-CNN is less affected by interference from background elements [4] such as potential EKG leads, external tubes, artifacts, overlapping devices, bones, and healthy tissues, and more accurately detects the pathological boundaries of "Lung Opacity".

Overall, FOAM improves the accuracy of prediction results of SOTA models across multiple tasks and datasets, demonstrates its strong generalization ability, and enhances the perception of foreground features in overlapping scenes.

V. Conclusion

In this paper, we attempt to address the critical challenges of foreground-background feature coupling in overlapping object perception tasks. Instead of adhering to mainstream spatial domain learning methods, we explore and leverage the advantages of frequency domain learning, designing a highly compatible joint perception approach in both the frequency and spatial domains, named the Frequency-Optimized Anti-Overlapping Framework (FOAM), which enhances the model's ability to perceive foreground textures and contours. Extensive experimental results demonstrate that FOAM exhibits superior performance and a wide range of applications, significantly improving the accuracy of existing SOTA models across four datasets for at least three overlapping object perception tasks: Prohibited Item Detection, Prohibited Item Segmentation, and Pneumonia Detection.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant U22A2063, 62173083, 62276186, and 62206043; the China Postdoctoral Science Foundation under No.2023M730517 and 2024T170114; the Liaoning Provincial "Selecting the Best Candidates by Opening Competition Mechanism" Science and Technology Program under Grant 2023JH1/10400045; the Fundamental Research Funds for the Central Universities under Grant N2424022; the Major Program of National Natural Science Foundation of China (71790614) and the 111 Project (B16009).

References

[1] B. Ma, T. Jia, M. Su, X. Jia, D. Chen, and Y. Zhang, "Automated segmentation of prohibited items in x-ray baggage images using dense de-overlap attention snake," *IEEE Transactions on Multimedia*, 2022. 1, 2, 4, 8, 9, 10, 11, 13

[2] Y. Wei, R. Tao, Z. Wu, Y. Ma, L. Zhang, and X. Liu, "Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 138–146. 1, 2, 4, 8, 9, 10

[3] G. Shih, C. C. Wu, S. S. Halabi, M. D. Kohli, L. M. Prevedello, T. S. Cook, A. Sharma, J. K. Amorosa, V. Arteaga, M. Galperin-Aizenberg *et al.*, "Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia," *Radiology: Artificial Intelligence*, vol. 1, no. 1, p. e180041, 2019. 1, 8

[4] M. Gambato, N. Scotti, G. Borsari, J. Zambon Bertoja, J.-D. Gabrieli, A. De Cassai, G. Cester, P. Navalesi, E. Quaia, and F. Causin, "Chest x-ray interpretation: detecting devices and device-related complications," *Diagnostics*, vol. 13, no. 4, p. 599, 2023. 1, 15

[5] M. Li, B. Ma, H. Wang, D. Chen, and T. Jia, "Gadet: A geometry-aware x-ray prohibited items detector," *IEEE Sensors Journal*, vol. 24, no. 2, pp. 1665–1678, 2024. 2, 4

[6] M. Li, T. Jia, H. Wang, B. Ma, H. Lu, S. Lin, D. Cai, and D. Chen, "Ao-detr: Anti-overlapping detr for x-ray prohibited items detection," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2024. 2, 4, 9, 10

[7] A. Chang, Y. Zhang, S. Zhang, L. Zhong, and L. Zhang, "Detecting prohibited objects with physical size constraint from cluttered x-ray baggage images," *Knowledge-Based Systems*, vol. 237, p. 107916, 2022. 2, 4

[8] C. Zhao, L. Zhu, S. Dou, W. Deng, and L. Wang, "Detecting overlapped objects in x-ray security imagery by a label-aware mechanism," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 998–1009, 2022. 2, 4

[9] C. Miao, L. Xie, F. Wan, C. Su, H. Liu, J. Jiao, and Q. Ye, "Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2119–2128. 2, 4

[10] M. Li, B. Ma, H. Wang, Y. Li, D. Chen, and T. Jia, "Pid-yolox: An x-ray prohibited items detector based on yolox," in *2023 IEEE 13th International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*. IEEE, 2023, pp. 413–418. 2, 3

[11] Z. Liu, X. Deng, P. Jiang, C. Lv, G. Min, and X. Wang, "Edge perception camouflaged object detection under frequency domain reconstruction," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2

[12] H. Liao, B. Huang, and H. Gao, "Feature-aware prohibited items detection for x-ray images," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 1040–1044. 2, 3, 6

[13] J. W. Cooley, P. A. W. Lewis, and P. D. Welch, "The fast fourier transform and its applications," *IEEE Transactions on Education*, vol. 12, no. 1, pp. 27–34, 1969. 2, 3, 4

[14] Z. Zhu, Y. Zhu, H. Wang, N. Wang, J. Ye, and X. Ling, "Fdtnet: Enhancing frequency-aware representation for prohibited object detection from x-ray images via dual-stream transformers," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108076, 2024. 2, 3, 6

[15] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Transactions on information Forensics and Security*, vol. 7, no. 3, pp. 868–882, 2012. 2, 3

[16] M. Fritsche, S. Gu, and R. Timofte, "Frequency separation for real-world super-resolution," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2019, pp. 3599–3608. 3

[17] P. Vandewalle, S. Süsstrunk, and M. Vetterli, "A frequency domain approach to registration of aliased images with application to super-resolution," *EURASIP journal on advances in signal processing*, vol. 2006, pp. 1–14, 2006. 3

[18] S. Grgic, M. Grgic, and B. Zovko-Cihlar, "Performance analysis of image compression using wavelets," *IEEE Transactions on industrial electronics*, vol. 48, no. 3, pp. 682–695, 2001. 3

[19] V. Velisavljevic, B. Beferull-Lozano, and M. Vetterli, "Space-frequency quantization for image compression with directionlets," *IEEE Transactions on Image Processing*, vol. 16, no. 7, pp. 1761–1773, 2007. 3

[20] K. Xu, M. Qin, F. Sun, Y. Wang, Y.-K. Chen, and F. Ren, "Learning in the frequency domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[21] M. Zhou, J. Huang, K. Yan, D. Hong, X. Jia, J. Chanussot, and C. Li, "A general spatial-frequency learning framework for multimodal image fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[22] Y. Sun, C. Xu, J. Yang, H. Xuan, and L. Luo, "Frequency-spatial entanglement learning for camouflaged object detection," in *European Conference on Computer Vision*. Springer, 2025, pp. 343–360. 3

[23] M. Ding, A. Qu, H. Zhong, Z. Lai, S. Xiao, and P. He, "An enhanced vision transformer with wavelet position embedding for histopathological image classification," *Pattern Recognition*, vol. 140, p. 109532, 2023. 3

[24] Z. Liu, X. Deng, P. Jiang, C. Lv, G. Min, and X. Wang, "Edge perception camouflaged object detection under frequency domain reconstruction," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 3

[25] M. Zhou, J. Huang, K. Yan, H. Yu, X. Fu, A. Liu, X. Wei, and F. Zhao, "Spatial-frequency domain information integration for pan-sharpening," in *European conference on computer vision*. Springer, 2022, pp. 274–291. 3

[26] Y. Cui, W. Ren, X. Cao, and A. Knoll, "Image restoration via frequency selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141. 3, 7, 13

[28] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 783–792. 3

[29] G. Yun, J. Yoo, K. Kim, J. Lee, and D. H. Kim, "Spanet: Frequency-balancing token mixer using spectral pooling aggregation modulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6113–6124. 3

[30] Y. Zhong, B. Li, L. Tang, S. Kuang, S. Wu, and S. Ding, "Detecting camouflaged object in frequency domain," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4504–4513. 3

[31] V. Britanak, P. C. Yip, and K. R. Rao, *Discrete cosine and sine transforms: general properties, fast algorithms and integer approximations*. Elsevier, 2010. 3, 7

[32] Y. Sun, C. Xu, J. Yang, H. Xuan, and L. Luo, "Frequency-spatial entanglement learning for camouflaged object detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 343–360. 3, 4

[33] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," *Advances in neural information processing systems*, vol. 31, 2018. 3

[34] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19. 3

[35] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 713–13 722. 3

[36] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018. 3

[37] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154. 3

[38] Z. Gao, J. Xie, Q. Wang, and P. Li, "Global second-order pooling convolutional networks," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2019, pp. 3024–3033. 3

[39] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 510–519. 3

[40] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha *et al.*, "Resnest: Split-attention networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2736–2746. 3

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. 3, 6, 12

[42] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 558–567. 3, 4

[43] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578. 3

[44] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229. 4

[45] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable {detr}: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=gZ9hCDWe6ke 4, 9, 14

[46] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 282–298. 4, 9, 10, 11

[47] Z. Cai and N. Vasconcelos, "Cascade r-cnn: High quality object detection and instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1483–1498, 2021. 4, 9, 10, 11

[48] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in neural information processing systems*, vol. 34, pp. 17 864–17 875, 2021. 4

[49] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299. 4, 9, 10, 11

[50] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022. 4, 6, 8, 9

[51] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, "Efficientvit: Memory efficient vision transformer with cascaded group attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14 420–14 430. 4

[52] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational visual media*, vol. 8, no. 3, pp. 415–424, 2022. 4

[53] R. Tao, Y. Wei, X. Jiang, H. Li, H. Qin, J. Wang, Y. Ma, L. Zhang, and X. Liu, "Towards real-world x-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10 923–10 932. 4

[54] L. Zhang, L. Jiang, R. Ji, and H. Fan, "Pidray: A large-scale x-ray benchmark for real-world prohibited item detection," *arXiv preprint arXiv:2211.10763*, 2022. 4

[55] B. Ma, T. Jia, M. Li, S. Wu, H. Wang, and D. Chen, "Towards dual-view x-ray baggage inspection: A large-scale benchmark and adaptive hierarchical cross refinement for prohibited item discovery," *IEEE Transactions on Information Forensics and Security*, pp. 1–1, 2024. 4

[56] S. Lin, T. Jia, H. Wang, B. Ma, M. Li, and D. Chen, "Detection of novel prohibited item categories for real-world security inspection," *Engineering Applications of Artificial Intelligence*, vol. 144, p. 110110, 2025. 4

[57] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *arXiv preprint arXiv:2203.03605*, 2022. 4, 9, 10, 11

[58] M. Li, T. Jia, H. Lu, B. Ma, H. Wang, and D. Chen, "Mmcl: Boosting deformable detr-based detectors with multi-class min-margin contrastive learning for superior prohibited item detection," *arXiv preprint arXiv:2406.03176*, 2024. 4, 9, 10

[59] ——, "Cspcl: Category semantic prior contrastive learning for deformable detr-based prohibited item detectors," *arXiv preprint arXiv:2501.16665*, 2025. 4, 9, 10

[60] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*. USA: Prentice-Hall, Inc., 2006. 4, 5

[61] M. Zhou, J. Huang, K. Yan, D. Hong, X. Jia, J. Chanussot, and C. Li, "A general spatial-frequency learning framework for multimodal image fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2024. 4

[62] R. Cong, M. Sun, S. Zhang, X. Zhou, W. Zhang, and Y. Zhao, "Frequency perception network for camouflaged object detection," in *Proceedings of the 31st ACM international conference on multimedia*, 2023, pp. 1179–1189. 6

[63] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255. 8

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 8, 9

[65] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500. 8, 9

[66] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019. 8

[67] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755. 9, 10

[68] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–338, 2010. 9

[69] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," *arXiv preprint arXiv:2304.08069*, 2023. 9, 10, 11

[70] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969. 9, 10, 11

[71] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 06, pp. 1137–1149, 2017. 10, 11

[72] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 10, 11

[73] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9759–9768. 10, 11

[74] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 002–21 012, 2020. 10, 11

[75] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "Solo: Segmenting objects by locations," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 2020, pp. 649–665. 10, 11

[76] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "Solov2: Dynamic and fast instance segmentation," *Advances in Neural information processing systems*, vol. 33, pp. 17 721–17 732, 2020. 10, 11