Decompose and Attribute: Boosting Generalizable Open-Set Object Detection via Objectness Score

Yuxuan Yuan[1, 2*], **Lichen Wei**[1, 2*], **Luyao Tang**[4], **Chaoqi Chen**[5], **Zheyuan Cai**[1, 3], **Yue Huang**[3, 6†], **Xinghao Ding**[1, 3]

[1] Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, China
[2]Institute of Artificial Intelligence, Xiamen University, China
[3]School of Informatics, Xiamen University, China
[4]The University of Hong Kong
[5]Shenzhen University
[6]The National Key Laboratory of Infrared Detection Technologies
{yuanyuxuan0908, weilichen26}@stu.xmu.edu.cn, yhuang2010@xmu.edu.cn

Abstract

Open-set object detection (OSOD) aims to recognize known object categories while localizing previously unseen instances. However, real-world scenarios often involve co-occurring domain shifts and novel object categories. Existing OSOD methods typically overlook domain shifts, relying on source-trained representations that entangle domain-specific style with semantic content, thereby hindering generalization to both unseen domains and novel categories. To address this challenge, we propose a unified framework, termed DecOmpose and ATtribute (DOAT), which disentangles domain-specific style from semantic structure, thereby facilitating generalizable object detection. DOAT employs wavelet-based feature decomposition to separate style information from high-frequency structural details, thus enabling an explicit separation of domain and category shifts. To account for domain shift, the low-frequency components are perturbed within a style subspace to simulate diverse domain appearances. For unknown object discovery, the high-frequency components are utilized to estimate objectness scores via an attribution mechanism that fuses wavelet energy with semantic distance to known-category prototypes. Extensive experiments on standard open-set benchmarks have demonstrated the superior generalization performance of DOAT.

Introduction

The ability to generalize to unseen environments is a critical requirement for object detectors deployed in real-world scenarios. Although modern detectors have achieved remarkable success on large-scale benchmarks with fixed domains and closed label sets, their performance often degrades significantly under real-world variations. Real-world environments are inherently non-stationary: detectors frequently encounter both domain shifts (Lin et al. 2021; Wu et al. 2025) in visual appearance (e.g., lighting and weather conditions)

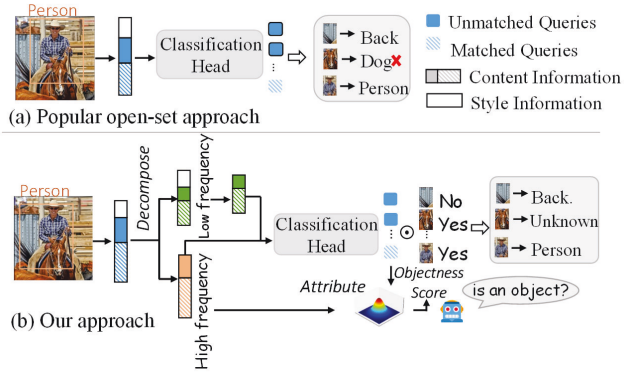

Figure 1: Motivation of DOAT. (a) Popular open-set detectors use entangled features mixing style and content, often misclassifying unknown objects under domain shift. (b) Our approach employs frequency-based decomposition and objectness attribution: low-frequency features suppress domain-specific style, while high-frequency cues estimate class-agnostic objectness to discover unknown instances.

and unseen object categories (Gupta et al. 2022; Sun, Li, and Mu 2024), which are neither annotated nor available during training. These two challenges often emerge simultaneously in practical settings and collectively result in significant degradation in detection performance.

To address the challenge of unknown categories, a variety of open-set object detection (OSOD) methods have been proposed in recent years (Gupta et al. 2022; Zheng et al. 2022; Fontanel et al. 2022; Zhao et al. 2023), aiming to detect and localize objects beyond the training label space. The prevailing approaches include energy-based scoring (Wu and Deng 2023; Wu, Chen, and Deng 2023), pseudo-labeling strategies (Gupta et al. 2022; Mullappilly et al. 2024), uncertainty estimation (Harakeh and Waslander 2021), and prototype-based method (Li, Guo, and Yuan 2023), among others. These methods typically leverage fea-

*These authors contributed equally.
†Corresponding Author.

ture representations of known classes to generate virtual representations for unknown categories, thereby facilitating the detection of unknown instances.

While existing methods have made significant progress in detecting unknown categories, they often struggle under domain shift due to the entangled representations of domain-specific style and semantic content. This leads to degraded performance and increased confusion between known and unknown objects in unseen domains. To this end, a critical question remains open in the field: *How to preserve the ability to detect unknown objects under distribution shift?* In this work, we posit that the effective discovery of unknown objects under domain shift requires an explicit disentanglement of domain-specific variations from semantic content. Building on this idea, we introduce a unified framework, DecOmpose and ATtribute (DOAT), which leverages frequency-aware modeling to separate these factors. The core motivation of this approach is visually depicted in Fig. 1.

To separate domain-sensitive style from semantic information, we decompose feature maps in the frequency domain (Liu et al. 2024). Unlike prior works that rely on Fast Fourier Transform (FFT) (Lin et al. 2023), DOAT employs wavelet transform, which preserves spatial locality and directional information. This design facilitates more precise modeling of localized object features. We observe that low-frequency components primarily encode style-related variations sensitive to domain shifts, whereas high-frequency components preserve domain-invariant structural features (Fig. 3). Motivated by this, the low-frequency stream is perturbed within a style subspace to simulate domain variations and improve the model's adaptability to new domains. To facilitate the discovery of unknown objects, a simple yet effective objectness attribution mechanism is introduced. Specifically, category-agnostic scores are computed by integrating coarse structural cues derived from wavelet energy with refined semantic deviations from known-category prototypes. This formulation of objectness eliminates the need to explicitly model unknown classes, enabling robust detection across both known and unknown categories.

The main contributions of this paper are summarized as follows:

- We propose DOAT, a unified framework for generalizable object detection. It disentangles style and structure through frequency decomposition, thereby enabling robust detection of unknown objects under domain shift.
- A novel objectness attribution mechanism is proposed, which assigns category-agnostic scores to queries by coarse structural cues from energy-guided attribution and refined semantic cues from semantic-guided attribution.
- The effectiveness of DOAT is demonstrated by extensive experiments.

Related Work

Domain Generalization (DG) aims to train models on data drawn from one or more source domains, with the objective of achieving robust performance on unseen target domains (Wang et al. 2021, 2025b). We categorize existing methods into four groups. **(1) Domain-generalized object detection** (Wang et al. 2023, 2024a,b, 2025a,c; Wu et al. 2024b) has attracted growing interest for its ability to operate without target domain data. **(2) Open-set DG.** Recently, some works have delved into the exploration of DG in open-set scenarios (Chen et al. 2023b,a). For example, (Shu et al. 2021) proposes a framework based on domain augmentation and meta-learning. (Katsumata et al. 2021) borrows metric learning to diffuse the feature representations of unknown classes but relies on existing DG baselines to acquire domain-invariant features. **(3) DG by style variation.** Style features are widely explored in DG (Zhong et al. 2022; Zhao et al. 2024; Li et al. 2024), which aims at changing the image style but maintaining the content. **(4) Frequency-based DG** (Bi, You, and Gevers 2024; Bi et al. 2024; Liu et al. 2024; Wu et al. 2024a) exploits the frequency domain to disentangle style and content, typically via FFT-based amplitude and phase decomposition. While FFT is widely used, Haar wavelet transform remains underutilized, which offers stronger feature decorrelation and better separation of frequency components, potentially benefiting generalization.

Open-Set Object Detection (OSOD) (Zhao et al. 2023; Zohar, Wang, and Yeung 2023; Ma et al. 2023) aims to train object detectors capable of identifying both known and unknown objects, drawing significant attention due to its potential for real-world applications. The concept of the OSOD problem was first introduced in (Dhamija et al. 2020). To improve unknown object discovery, some methods (Han et al. 2022; Mullappilly et al. 2024) emphasize spatial cues to produce high-quality pseudo labels, OW-DETR (Gupta et al. 2022) adapts detection transformers (Zhu et al. 2020) by reusing unmatched queries in a self-training scheme. Other approaches (Du et al. 2022) synthesize unknown representations from known base classes. PROB (Zohar, Wang, and Yeung 2023) improves unknown recall by modeling them via class-agnostic Gaussian distributions, whereas CAT (Ma et al. 2023) mimics human perception to better detect unknowns in open-world scenarios.

Preliminary and Motivation

Problem Formulation

Let a labeled source domain $\mathcal{D}_s = \{(x_s^i, y_s^i)\}_{i=1}^{n_s}$ be given, where each $x_s^i \in \mathcal{X}$ denotes an image, and $y_s^i \in \mathcal{Y}_s$ represents a set of bounding boxes and associated labels drawn from a known class set $\mathcal{C}_s$. Generalizable object detection addresses two fundamental challenges. **Domain shift**: The model must generalize to an unlabeled target domain $\mathcal{D}_t = \{x_t^j\}_{j=1}^{n_t}$, whose distribution differs from that of the source. **Category shift**: The target domain contains instances from a superset of classes $\mathcal{C}_t \supset \mathcal{C}_s$, where the unknown class set is defined as $\mathcal{C}_u = \mathcal{C}_t \setminus \mathcal{C}_s$. These classes are never labeled or exposed during training. The objective of generalizable object detection is to train a detector using only the labeled source domain such that it can accurately localize and recognize both known classes ($\{1, 2, \ldots, |\mathcal{C}_s|\}$) and unknown objects ($|\mathcal{C}_s| + 1$) (Gupta et al. 2022).

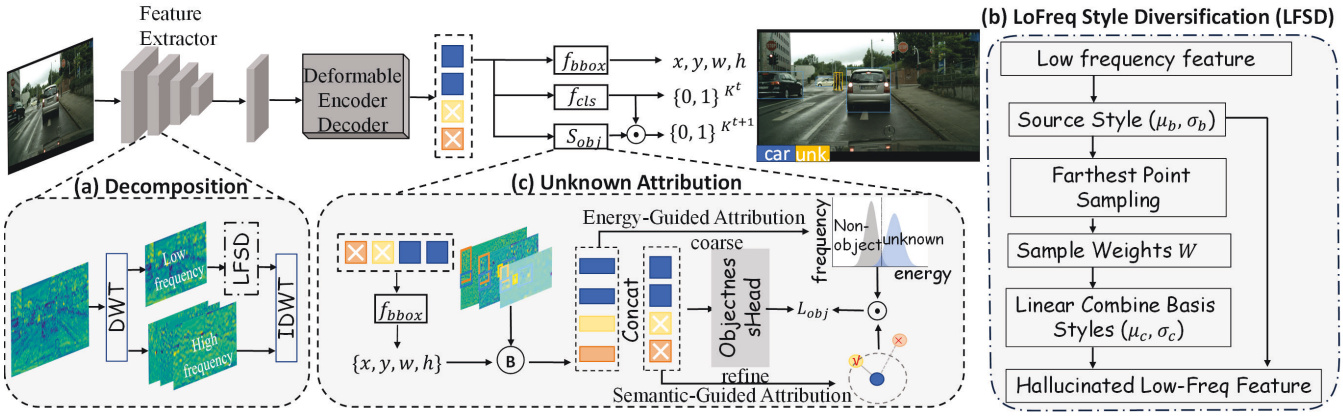

Figure 2: Overview of the proposed DOAT framework, which consists of three key modules: (a) Decomposition: Apply wavelet transform to separate features into low- and high-frequency components, disentangling domain-sensitive style and semantic structure; (b) LFSD: Perturb low-frequency features within a learned style subspace to simulate diverse domain shifts; (c) Unknown Attribution: Estimate category-agnostic objectness scores by integrating wavelet energy and semantic deviation from known-class prototypes, guiding robust unknown discovery.

Motivation of Objectness Score

Detecting unknown objects is inherently challenging due to the absence of supervision. These unknowns encompass both novel categories that have never been encountered during training and unlabeled objects present within the source domain. Most existing methods rely on complex pseudo-labeling strategies to train a $|\mathcal{C}_s| + 1$-class classification head (Li, Guo, and Yuan 2023; Yuan et al. 2025), which is unreliable under domain shift. In contrast, prior studies have noted that both known and unknown objects tend to exhibit general object-like characteristics that distinguish them from background regions (Zohar, Wang, and Yeung 2023).

Motivated by this observation, we introduce an **objectness head** that explicitly models objectness as a category-agnostic signal of object presence, thereby providing a more reliable cue for the discovery of unknown instances. Formally, given a decoder query $\mathbf{q}_i$, we estimate its objectness probability $p(o \mid \mathbf{q}_i)$, where $o = 1$ indicates that $\mathbf{q}_i$ corresponds to a real object. The final classification probability is then defined as:

$$p(c \mid \mathbf{q}_i) = p(c \mid o = 1, \mathbf{q}_i) \cdot p(o = 1 \mid \mathbf{q}_i), \tag{1}$$

where $p(c \mid o = 1, \mathbf{q}_i)$ denotes the conditional probability of assigning class c, given the presence of an object. This formulation reduces the likelihood of misclassifying unknown objects as known classes and eliminates the need for generating explicit labels for unknowns. The objectness head is solely responsible for estimating object presence in a category-agnostic manner.

Proposed Method

Fig. 2 provides an overview of the proposed DOAT framework, which adapts the deformable DETR (D-DETR) (Zhu et al. 2020). The framework comprises three key components: (1) **Frequency-based feature decomposition**, which separates low- and high-frequency components via wavelet transform; (2) **style diversification for unseen perception**, in which the low-frequency features are perturbed to enhance adaptability to novel environments; (3) **objectness attribution or unknown discovery**, which utilizes high-frequency cues for recognizing previously unseen objects.

Frequency-based Feature Decomposition

Visual features contain both low- and high-frequency components that encode different aspects of an image. The low-frequency components are often associated with coarse structure, while the high-frequency components preserve finer details and object boundaries (Bi et al. 2024). Most existing frequency-based methods apply the Fast Fourier Transform (FFT) to convert visual signals into the frequency domain (Yu et al. 2022; Lin et al. 2023; Lee, Bae, and Kim 2023). The basis functions of FFT are globally supported sinusoids, meaning that each frequency component spans the entire spatial domain of the image. Such behavior makes FFT suboptimal for tasks requiring spatially localized feature modulation (e.g., object detection and segmentation). To address this limitation, a spatially localized frequency decomposition based on the Discrete Wavelet Transform (DWT) is adopted, which is better suited for modeling the non-stationary and locally structured characteristics of visual features.

In our framework, the feature $\mathbf{f}_l$ extracted from the l-th backbone block is first decomposed into low- and high-frequency components. Specifically, the decomposition is performed using the LL^T, LH^T, HL^T, and HH^T kernels (Bay, Tuytelaars, and Van Gool 2006), defined as:

$$LL^T = \frac{1}{2}\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}, \quad LH^T = \frac{1}{2}\begin{bmatrix}-1 & 1\\ -1 & 1\end{bmatrix},$$
$$HL^T = \frac{1}{2}\begin{bmatrix}-1 & -1\\ 1 & 1\end{bmatrix}, \quad HH^T = \frac{1}{2}\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}. \tag{2}$$

These filters enable the separation of the feature map into

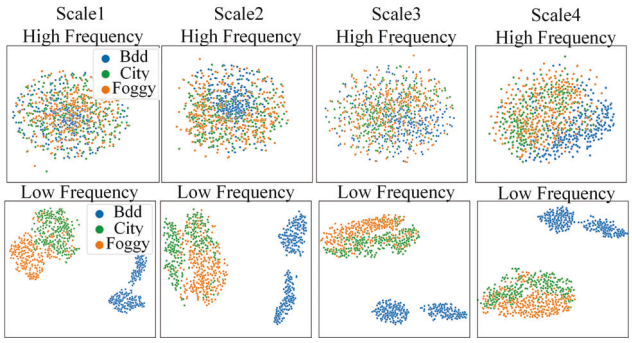

Figure 3: t-SNE (Van der Maaten and Hinton 2008) visualization of the high-frequency and low-frequency features extracted from four different scales of the ResNet backbone. The features are obtained from three domains: Cityscapes, Foggy Cityscapes, and BDD.

the coarse, style-related component and the detailed, object-related components. This decomposition enables independent processing of low- and high-frequency components, allowing style perturbations to be applied in the low-frequency band while preserving structural information in the high-frequency band, thereby enhancing the robustness of unknown object detection. The decomposition is computed as,

$$\begin{aligned} \mathbf{f}_l^{LL} &= \mathbf{f}_l \otimes LL^T, \qquad \mathbf{f}_l^{LH} = \mathbf{f}_l \otimes LH^T, \\ \mathbf{f}_l^{HL} &= \mathbf{f}_l \otimes HL^T, \qquad \mathbf{f}_l^{HH} = \mathbf{f}_l \otimes HH^T, \end{aligned} \tag{3}$$

where $\otimes$ represents the convolution operation, $\mathbf{f}_l^{LL}$ denotes the low-frequency feature after layer l. In contrast $\mathbf{f}_l^{LH}, \mathbf{f}_l^{HL}$ and $\mathbf{f}_l^{HH}$ are the high-frequency feature after layer l.

Style Diversification for Unseen Perception

As illustrated in Fig. 3, low-frequency features exhibit clear domain clustering, suggesting that they encode domain-specific style information. In contrast, high-frequency features display more substantial overlap across domains, indicating a higher degree of domain invariance. This observation supports the intuition that wavelet-based decomposition naturally separates style-sensitive and semantic-rich components. While conventional style augmentation methods typically operate on the entire image or feature space, they implicitly assume that style information is uniformly distributed across frequency bands. Our analysis, however, reveals that style variations are predominantly embedded in low-frequency components, whereas high-frequency features preserve semantic structure and object boundaries.

Motivated by this insight, we propose **LoFreq Style Diversification (LFSD)**, a frequency-aware augmentation strategy that perturbs only the low-frequency features to simulate domain-specific variations. As shown in Fig. 2(b), this targeted perturbation increases style diversity without disrupting semantic content. To ensure the perturbed styles remain realistic, we constrain the perturbation within a learned style subspace. Specifically, we define the style space as a subspace $\mathcal{S} \subseteq \mathbb{R}^C$ of the full C-dimensional feature space. Directly sampling random or orthogonal vectors from the ambient space may yield implausible styles that degrade performance. An effective augmentation strategy should therefore maintain a balance between diversity and realism by operating within the feasible region of $\mathcal{S}$. Specifically, we adopt an AdaIN-style re-normalization strategy (Huang and Belongie 2017),

$$\tilde{\mathbf{f}}^{LL} = \sigma^H \cdot \frac{\mathbf{f}^{LL} - \mu(\mathbf{f}^{LL})}{\sigma(\mathbf{f}^{LL})} + \mu^H, \tag{4}$$

where (μ^H, σ^H) are synthesized statistics sampled from the style subspace.

Naively mixing source statistics, as done in previous works (Nuriel, Benaim, and Wolf 2021; Chen et al. 2022), typically results in dominant-style interpolation that fails to capture the diversity of style modes, particularly under a single source domain. To mitigate this, we construct a compact and expressive style basis $\{(\mu_c^k, \sigma_c^k)\}_{k=1}^K$ by applying Farthest Point Sampling (FPS) (Qi et al. 2017) to the source style pool $\{(\mu_b^j, \sigma_b^j)\}_{j=1}^N$. FPS ensures that the selected basis spans a wide and well-separated region of $\mathcal{S}$, promoting diverse yet realistic style coverage.

New style targets are synthesized by sampling interpolation weights $W \in \Delta^K$ from a Dirichlet distribution (Minka 2000), and computing:

$$\mu^H = \sum_{k=1}^{K} W_k \mu_c^k, \quad \sigma^H = \sum_{k=1}^{K} W_k \sigma_c^k. \tag{5}$$

Following prior work (Zhao et al. 2024), we set $K = C$, where C is the feature dimensionality, ensuring that the basis styles adequately span the style subspace $\mathcal{S}$. By dynamically recombining diverse basis styles, LFSD generates stylistically rich yet semantically consistent perturbations, improving generalization under domain shift.

Objectness Attribution for Unknown Discovery

The second core challenge in generalizable object detection is the discovery of unknown objects, which belong to novel categories not observed or annotated during training. Building on the motivation of objectness score, we further explore how to effectively attribute objectness in a class-agnostic manner. Specifically, we observe that objects frequently exhibit generic object-like characteristics, such as well-defined contours, local contrast, and structural closure—properties that are preserved in the high-frequency subbands of wavelet-decomposed features. To exploit this, a novel Objectness Attribution module is introduced.

To estimate objectness score effectively, we design two complementary attribution mechanisms: (i) **Energy-Guided Attribution** (EGA), which quantifies the structural richness of each query's attended region based on high-frequency wavelet energy; and (ii) **Semantic-Guided Attribution** (SGA), which measures the deviation of a query from the distribution of known-class embeddings in semantic space. These complementary cues jointly supervise the objectness head, enabling category-agnostic objectness estimation.

Energy-Guided Attribution. To capture structural saliency, a high-frequency energy embedding is computed for each query $\mathbf{q}_i \in \mathbb{R}^D$. We first project its predicted bounding box onto each feature scale $l \in \{1, \ldots, L\}$ and extract wavelet subband activations $\mathbf{f}_l^{LH}, \mathbf{f}_l^{HL}, \mathbf{f}_l^{HH}$ from the box region $B_i^{(l)}$. The energy is defined as follows:

$$\mathcal{E}(\mathbf{q}_i) = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{|B_i^{(l)}|}\sum_{(x,y)\in B_i^{(l)}}\Big[\left\|\mathbf{f}_l^{LH}(x,y)\right\|^2 + \left\|\mathbf{f}_l^{HL}(x,y)\right\|^2 + \left\|\mathbf{f}_l^{HH}(x,y)\right\|^2\Big], \tag{6}$$

where each $\|\cdot\|^2$ denotes channel-wise square (no reduction), preserving the feature dimensionality. The resulting energy embedding $\mathcal{E}(\mathbf{q}_i) \in \mathbb{R}^D$ is then concatenated with the query embedding $\mathbf{q}_i$, and the fused representation is fed into the objectness head to predict the objectness score.

Semantic-Guided Attribution. Although energy-based attribution captures structural details, it may wrongly highlight background regions with rich textures, such as buildings or vegetation, as objects. To further refine objectness estimation, we introduce semantic-guided attribution, which exploits the semantic distance to known-class queries as a soft prior for unknown discovery.

Specifically, we maintain a prototype $\hat{\mathbf{q}} \in \mathbb{R}^D$ representing the semantic center of known-object embeddings. During training, $\hat{\mathbf{q}}$ is updated using an Exponential Moving Average (EMA) (Li, Xiong, and Hoi 2020) over matched queries, with an update momentum $\alpha \in (0, 1)$,

$$\hat{\mathbf{q}} \leftarrow \alpha \cdot \hat{\mathbf{q}} + (1-\alpha)\cdot \text{mean}(\mathbf{q}_\text{m}). \tag{7}$$

For each query, we compute the semantic distance to the prototype using normalized embeddings,

$$\text{dist}(\mathbf{q}_i, \hat{\mathbf{q}}) = \left\| \frac{\mathbf{q}_i}{\|\mathbf{q}_i\|} - \frac{\hat{\mathbf{q}}}{\|\hat{\mathbf{q}}\|} \right\|^2. \tag{8}$$

The final *semantic-guided objectness score* is then computed as:

$$s_{\mathbf{q}_i} = \exp\left(-\frac{\text{dist}(\mathbf{q}_i, \hat{\mathbf{q}})}{\tau}\right)\cdot \hat{\mathcal{E}}(\mathbf{q}_i), \tag{9}$$

where $\hat{\mathcal{E}}(\mathbf{q}_i)$ denotes the normalized energy score, and τ is a temperature hyperparameter that controls the decay of semantic influence. The soft score $s_{\mathbf{q}_i} \in [0, 1]$ is used to supervise the training of the objectness head. The loss is formulated as:

$$\mathcal{L}_\text{obj} = -\frac{1}{N}\sum_{i=1}^{N}\left[s_{\mathbf{q}_i}\cdot \log p_i + (1 - s_{\mathbf{q}_i})\cdot \log(1-p_i)\right], \tag{10}$$

where p_i is the predicted objectness probability, and N is the number of decoder queries.

Experiments

Benchmarks and Setup

Datasets. Following (Li, Guo, and Yuan 2023), we evaluate the effectiveness of DOAT on three benchmark scenarios. For Cityscapes → Foggy Cityscapes/BDD100K, category splits are defined under 12 settings along two axes. (1) **Semantic overlap and instance frequency**, including four sub-tasks: heterogeneous semantics (**het-sem**), homogeneous semantics (**hom-sem**), frequency decrease (**freq-dec**), and frequency increase (**freq-inc**); and (2) **Number of unknown classes**, comprising three sub-tasks. For Pascal VOC → Clipart1K, the splits are based solely on the count of unknown categories.

Implementation Details. We adopt Deformable DETR as the base detector and initialize its ResNet-50 backbone with DINO-pretrained weights (He et al. 2016; Zhang et al. 2022). The model is optimized using AdamW (Loshchilov and Hutter 2017) with a batch size of 4 and an initial learning rate of 2×10^{-4}, along with a weight decay of 5×10^{-4}. Training is conducted for 80 epochs on an NVIDIA A40 GPU. The architecture employs three encoder and decoder layers, with 100 object queries and embedding dimension set to 256. We set the momentum α for prototype updates to 0.9, and the temperature parameter τ in Eq. 9 to 1.0. We apply wavelet decomposition to the feature maps extracted from all stages of the backbone.

Evaluation Metrics. We utilize mean Average Precision (mAP_b) to evaluate the base class performance. In line with prior works (Li, Guo, and Yuan 2023; Dhamija et al. 2020), we adapt Average Recall (AR_n), Wilderness Impact (WI) (Dhamija et al. 2020), and Absolute Open-Set Error (AOSE) (Dhamija et al. 2020) for novel class evaluation. WI and AOSE quantify the confusion when predicting a novel object as the base classes.

Baselines. We compare DOAT against six baseline methods. Open-set Detector (**OpenDet**) (Han et al. 2022), Open-World Detection Transformer (**OW-DETR**) (Gupta et al. 2022), Probabilistic Objectness transformer-based open-world detector (**PROB**) (Zohar, Wang, and Yeung 2023), Localization and identification Cascade Detection Transformer (**CAT**) (Ma et al. 2023), Structured Motif Matching (**SOMA**) (Li, Guo, and Yuan 2023), and Semi-Supervised Open-World detector (**SS-OWFormer**) (Mullappilly et al. 2024). All methods are reproduced and evaluated under the generalizable object detection setting for fair comparison.

Main Results

Cityscapes → Foggy Cityscapes. We evaluate our proposed method DOAT under 12 diverse settings on the Cityscapes → Foggy Cityscapes in Tab. 1. Across all 12 settings, DOAT consistently achieves the highest unknown-class detection performance, as indicated by superior values in both mAP_k and AR_u, while maintaining competitive AOSE and WI scores. These results demonstrate the effectiveness of DOAT under different category shifts. Although the relative gain over D-DETR under the freq-dec setting is smaller due to the increased difficulty of discovering rare unknown classes, DOAT still surpasses all open-set baselines by a clear margin, demonstrating its robustness even under severe open-category shifts. Fig. 4 provides the qualitative comparisons, where DOAT accurately identifies unknown instances and yields more precise bounding box predictions.

Task	Method	num.unknown-class: 3				num.unknown-class: 4				num.unknown-class: 5			
		mAP_k↑	AR_u↑	WI↓	AOSE↓	mAP_k↑	AR_u↑	WI↓	AOSE↓	mAP_k↑	AR_u↑	WI↓	AOSE↓
het-sem	D-DETR (Carion et al. 2020)	37.69	0.00	0.376	454	36.67	0.00	0.508	899	33.96	0.00	0.598	1248
	OpenDet (Han et al. 2022)	37.12	2.32	0.371	268	35.54	1.51	0.499	524	33.26	1.98	0.621	1048
	OW-DETR (Gupta et al. 2022)	34.32	2.24	0.482	251	35.38	1.98	0.702	513	30.49	2.03	0.813	922
	PROB (Zohar, Wang, and Yeung 2023)	28.71	2.83	0.393	209	28.45	3.43	0.564	340	33.32	2.72	0.674	777
	CAT (Ma et al. 2023)	33.12	2.12	0.427	202	33.03	2.26	0.581	416	26.11	2.12	0.681	725
	SOMA (Li, Guo, and Yuan 2023)	34.58	1.54	**0.327**	193	33.50	1.84	0.433	371	31.07	1.64	0.655	719
	SS-OWFormer (Mullappilly et al. 2024)	40.52	2.88	0.424	182	38.67	2.19	0.629	375	35.69	2.82	0.697	682
	DOAT (ours)	**42.68**	**4.67**	0.353	**159**	**42.10**	**4.54**	**0.428**	**339**	**39.23**	**4.28**	**0.581**	**662**
hom-sem	D-DETR (Carion et al. 2020)	39.59	0.00	3.148	4283	38.88	0.00	2.700	5135	36.46	0.00	3.110	8959
	OpenDet (Han et al. 2022)	37.17	4.75	2.521	1735	37.80	5.32	2.663	2205	35.24	6.21	3.031	3505
	OW-DETR (Gupta et al. 2022)	36.61	3.75	2.533	1736	36.00	3.59	2.680	1880	33.49	4.95	3.262	3473
	PROB (Zohar, Wang, and Yeung 2023)	25.67	10.12	3.045	2031	25.02	9.92	3.201	2349	24.40	11.30	2.930	4692
	CAT (Ma et al. 2023)	30.90	8.89	3.577	2335	30.50	9.30	3.773	2839	29.01	10.17	4.701	5626
	SOMA (Li, Guo, and Yuan 2023)	40.62	8.76	2.814	2603	39.62	9.66	2.992	3195	37.14	9.89	3.507	5642
	SS-OWFormer (Mullappilly et al. 2024)	42.21	9.37	2.342	1696	41.58	9.79	2.406	**1756**	38.73	10.93	3.448	3567
	DOAT (ours)	**44.43**	**10.62**	**1.735**	**1643**	**43.37**	**10.13**	**2.010**	2066	**39.99**	**11.45**	**2.763**	**3428**
freq-dec	D-DETR (Carion et al. 2020)	52.01	0.00	0.572	1055	50.46	0.00	0.841	1833	49.18	0.00	0.917	2012
	OpenDet (Han et al. 2022)	52.01	9.16	0.582	570	50.01	10.02	0.801	1179	49.10	10.28	0.898	1248
	OW-DETR (Gupta et al. 2022)	51.81	8.06	0.547	575	50.29	9.13	0.792	1074	49.14	9.33	0.870	1219
	PROB (Zohar, Wang, and Yeung 2023)	40.12	9.33	0.423	488	39.38	10.12	0.671	1012	39.76	10.41	0.824	1201
	CAT (Ma et al. 2023)	44.40	6.30	0.448	495	43.52	8.11	0.742	**941**	42.49	8.08	0.844	1168
	SOMA (Li, Guo, and Yuan 2023)	51.03	9.13	0.467	631	49.74	9.85	0.669	1078	47.63	10.72	0.792	1403
	SS-OWFormer (Mullappilly et al. 2024)	51.84	9.21	0.480	**438**	51.94	10.07	0.737	1052	51.03	10.49	0.837	1066
	DOAT (ours)	**52.80**	**10.30**	**0.393**	501	**52.93**	**10.88**	**0.642**	1069	**51.82**	**11.01**	**0.672**	**1016**
freq-inc	D-DETR (Carion et al. 2020)	30.68	0.00	2.393	2719	30.89	0.00	2.417	4499	28.06	0.00	**2.679**	7663
	OpenDet (Han et al. 2022)	28.17	4.75	2.521	1735	27.80	5.32	2.663	2205	25.24	6.21	3.031	4105
	OW-DETR (Gupta et al. 2022)	27.28	5.42	3.343	1806	26.92	5.52	3.428	2941	24.25	5.26	3.698	4798
	PROB (Zohar, Wang, and Yeung 2023)	21.06	8.64	5.637	2289	22.38	9.51	5.571	3478	20.76	8.41	5.612	5757
	CAT (Ma et al. 2023)	25.41	7.91	6.137	2571	26.13	8.03	6.491	4092	23.23	7.60	6.543	6052
	SOMA (Li, Guo, and Yuan 2023)	32.97	3.93	2.673	1158	30.08	6.09	3.428	2654	26.49	5.84	3.436	4290
	SS-OWFormer (Mullappilly et al. 2024)	34.36	5.72	6.316	1322	34.75	6.04	6.520	2552	31.31	5.92	6.435	3940
	DOAT (ours)	**38.28**	**10.16**	**2.307**	**1067**	**38.69**	**10.29**	**1.973**	**2158**	**35.21**	**9.77**	2.757	**3557**

Table 1: Results on Cityscapes → Foggy Cityscapes dataset under 12 different task settings.

Method	num.unknown-class: 6				num.unknown-class: 8				num.unknown-class: 10			
	mAP_k↑	AR_u↑	WI↓	AOSE↓	mAP_k↑	AR_u↑	WI↓	AOSE↓	mAP_k↑	AR_u↑	WI↓	AOSE↓
D-DETR (Carion et al. 2020)	18.34	0.00	6.057	4567	18.17	0.00	6.459	5379	16.39	0.00	6.893	6402
OpenDet (Han et al. 2022)	16.57	21.20	6.472	3608	16.02	21.24	7.558	4122	15.88	17.66	7.598	5248
OW-DETR(Gupta et al. 2022)	16.67	21.78	6.637	3711	16.40	20.00	7.408	4278	15.71	17.75	7.885	5254
PROB (Zohar, Wang, and Yeung 2023)	12.12	24.66	6.101	3091	11.51	25.07	6.826	4213	11.76	21.41	6.968	5921
CAT (Ma et al. 2023)	13.06	22.64	6.617	3289	13.38	24.51	6.571	4351	13.87	20.45	7.012	6016
SOMA (Li, Guo, and Yuan 2023)	15.20	26.64	7.197	3966	15.04	25.50	7.800	4632	14.48	24.27	8.738	6003
SS-OWFormer (Mullappilly et al. 2024)	18.31	25.66	6.642	3820	18.12	26.76	7.533	4296	17.73	22.87	8.284	5369
DOAT (ours)	**23.56**	**34.92**	**5.994**	**2450**	**23.50**	**35.93**	**6.170**	**2940**	**21.53**	**36.88**	**6.176**	**3607**

Table 2: Performance on Pascal VOC → Clipart dataset.

Pascal VOC → Clipart. As shown in Tab. 2, DOAT significantly outperforms all open-set baselines across different numbers of unknown classes. It achieves the highest AR_u and mAP_k, while maintaining the lowest WI and AOSE, highlighting its superior ability to detect novel objects under large domain and style shifts. Notably, DOAT maintains stable performance even as the number of unknown classes increases, demonstrating strong scalability and robustness.

Cityscapes → BDD100K. As shown in Tab. 3, under the het-sem split, DOAT achieves the best performance in most metrics across all settings. While the absolute gains are less pronounced than in other settings, we attribute this to the domain's inherent difficulty, including low lighting, motion blur, and small object sizes, which limit the effectiveness of high-frequency cues in objectness estimation. Notably, most baseline methods also suffer significant performance drops in this setting, underscoring the general challenge posed by BDD100K under heterogeneous category shift.

Ablation Studies

Ablations of key process in DOAT. In this part, we provide the ablation results in Tab. 4, investigating the independent and combined effects of decomposing (Dec) and attributing (Att) proposed in DOAT. Both the decomposition and objectness attribution modules contribute significantly to unknown object detection. Removing either leads to substantial AR_u drops, particularly on Foggy. The SGA further refines the objectness estimation, as omitting it results in noticeable performance degradation. The full DOAT model achieves the best results, confirming the effectiveness of its modular design.

Analysis on style diversification. As shown in Tab. 5 FPS consistently outperforms random interpolation and perturbation techniques across datasets. Notably, applying FPS on

Figure 4: Qualitative comparisons between SOMA (top) and DOAT (bottom).

Method	num.unknown-class: 3				num.unknown-class: 4				num.unknown-class: 5			
	mAP_k↑	AR_u↑	WI↓	AOSE↓	mAP_k↑	AR_u↑	WI↓	AOSE↓	mAP_k↑	AR_u↑	WI↓	AOSE↓
D-DETR (Carion et al. 2020)	13.24	0.00	0.133	956	13.27	0.00	0.170	1054	13.29	0.00	0.184	1598
OpenDet (Han et al. 2022)	13.02	1.10	0.131	712	12.91	1.36	0.197	902	12.92	1.27	0.171	1248
OW-DETR (Gupta et al. 2022)	12.77	1.19	0.129	630	12.77	1.21	0.181	700	12.80	1.32	0.162	862
PROB (Zohar, Wang, and Yeung 2023)	10.71	**1.83**	**0.128**	609	10.29	1.78	0.166	728	9.82	**1.88**	0.178	801
CAT (Ma et al. 2023)	11.12	1.45	0.135	730	11.38	1.51	0.171	778	11.76	1.41	0.182	812
SOMA (Li, Guo, and Yuan 2023)	8.22	1.22	0.163	764	7.15	0.79	0.169	**512**	7.33	0.88	0.198	736
SS-OWFormer (Mullappilly et al. 2024)	12.21	1.61	0.158	782	12.64	1.02	0.157	623	12.34	1.06	0.167	701
DOAT (ours)	**13.58**	1.82	0.129	**608**	**13.72**	**1.91**	**0.146**	616	**13.70**	**1.88**	**0.159**	**660**

Table 3: Performance on Cityscapes → BDD100K dataset under **het-sem** setting.

Set	Foggy		Clipart	
	mAP_k	AR_u	mAP_k	AR_u
D-DETR	30.68	0.00	18.34	0.00
w/o. Att	34.21	0.01	19.26	0.00
w/o. Dec	34.64	7.43	20.12	30.28
Att w/o. SGA	37.58	6.95	22.43	28.95
DOAT (Full)	**38.28**	**10.16**	**23.56**	**34.92**

Table 4: Ablation of DOAT (%).

Strategy	Foggy	Clipart	BDD
Random Interpolation	37.26	22.23	**13.66**
Random Perturbation	36.45	21.58	13.12
FPS (ours)	**38.28**	**23.56**	13.53
FPS on Full Feature	37.42	22.70	13.38

Table 5: Comparison of different feature perturbation strategies. mAP_k (%) is reported.

low-frequency components leads to better performance than on full features, validating the effectiveness of conducting perturbation specifically in the low-frequency domain.

Analysis on objectness attribution. Fig. 5 provides a comprehensive analysis of the proposed objectness attribution. Fig. 5(a) illustrates the score distribution predicted by the objectness head. It can be observed that the objectness scores for foreground and background queries follow distinguishable, approximately Gaussian-like distributions. This

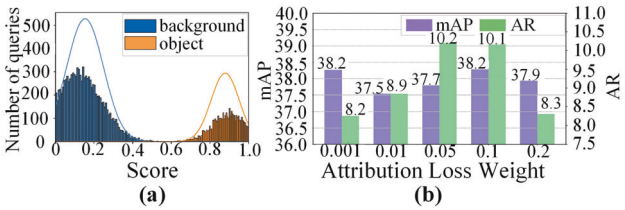

Figure 5: (a) Distribution of objectness head prediction. (b) Loss weight for objectness attribution.

behavior suggests that our model is able to learn a structured and interpretable objectness space. Fig. 5(b) explores the impact of loss weighting on objectness attribution. Varying the attribution loss weight results in relatively stable performance in both mAP_k and AR_u, indicating that the method exhibits low sensitivity to this hyperparameter, thereby highlighting its robustness and practical tunability.

Conclusion

This paper introduces DOAT, a unified framework that addresses the challenge of open-set object detection under domain shift. Leveraging wavelet-based frequency decomposition, DOAT isolates domain-sensitive low-frequency features for perturbation, thereby improving generalization across unseen domains while preserving object semantics. In parallel, the proposed objectness attribution utilizes both high-frequency structural cues and semantic divergence to estimate objectness scores, facilitating the discovery of unknown objects without relying on pseudo-labels.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 82272071, 62271430, 82172073, and 52105126; and in part by the Open Fund of the National Key Laboratory of Infrared Detection Technologies.

References

Bay, H.; Tuytelaars, T.; and Van Gool, L. 2006. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, 404–417. Springer.

Bi, Q.; Yi, J.; Zheng, H.; Zhan, H.; Huang, Y.; Ji, W.; Li, Y.; and Zheng, Y. 2024. Learning frequency-adapted vision foundation model for domain generalized semantic segmentation. *Advances in Neural Information Processing Systems*, 37: 94047–94072.

Bi, Q.; You, S.; and Gevers, T. 2024. Generalized foggy-scene semantic segmentation by frequency decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1389–1399.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, C.; Li, J.; Han, X.; Liu, X.; and Yu, Y. 2022. Compound domain generalization via meta-knowledge encoding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7119–7129.

Chen, C.; Tang, L.; Huang, Y.; Han, X.; and Yu, Y. 2023a. CODA: generalizing to open and unseen domains with compaction and disambiguation. *Advances in Neural Information Processing Systems*, 36.

Chen, C.; Tang, L.; Tao, L.; Zhou, H.-Y.; Huang, Y.; Han, X.; and Yu, Y. 2023b. Activate and reject: towards safe domain generalization under category shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11552–11563.

Dhamija, A.; Gunther, M.; Ventura, J.; and Boult, T. 2020. The overlooked elephant of object detection: Open set. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1021–1030.

Du, X.; Wang, Z.; Cai, M.; and Li, Y. 2022. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*.

Fontanel, D.; Tarantino, M.; Cermelli, F.; and Caputo, B. 2022. Detecting the unknown in object detection. *arXiv preprint arXiv:2208.11641*.

Gupta, A.; Narayan, S.; Joseph, K.; Khan, S.; Khan, F. S.; and Shah, M. 2022. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9235–9244.

Han, J.; Ren, Y.; Ding, J.; Pan, X.; Yan, K.; and Xia, G.-S. 2022. Expanding low-density latent regions for open-set object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9591–9600.

Harakeh, A.; and Waslander, S. L. 2021. Estimating and evaluating regression predictive uncertainty in deep object detectors. *arXiv preprint arXiv:2101.05036*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 1501–1510.

Katsumata, K.; Kishida, I.; Amma, A.; and Nakayama, H. 2021. Open-set domain generalization via metric learning. In *2021 IEEE International Conference on Image Processing (ICIP)*, 459–463. IEEE.

Lee, S.; Bae, J.; and Kim, H. Y. 2023. Decompose, adjust, compose: Effective normalization by playing with frequency for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11776–11785.

Li, J.; Xiong, C.; and Hoi, S. C. 2020. Mopro: Webly supervised learning with momentum prototypes. *arXiv preprint arXiv:2009.07995*.

Li, W.; Guo, X.; and Yuan, Y. 2023. Novel Scenes & Classes: Towards Adaptive Open-set Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15780–15790.

Li, Y.; Zhang, D.; Keuper, M.; and Khoreva, A. 2024. Intra- & extra-source exemplar-based style synthesis for improved domain generalization. *International Journal of Computer Vision*, 132(2): 446–465.

Lin, C.; Yuan, Z.; Zhao, S.; Sun, P.; Wang, C.; and Cai, J. 2021. Domain-invariant disentangled network for generalizable object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8771–8780.

Lin, S.; Zhang, Z.; Huang, Z.; Lu, Y.; Lan, C.; Chu, P.; You, Q.; Wang, J.; Liu, Z.; Parulkar, A.; et al. 2023. Deep frequency filtering for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11797–11807.

Liu, L.; Wang, R.; Wang, Y.; Jing, L.; and Wang, C. 2024. Frequency shuffling and enhancement for open set recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3675–3683.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, S.; Wang, Y.; Wei, Y.; Fan, J.; Li, T. H.; Liu, H.; and Lv, F. 2023. Cat: Localization and identification cascade detection transformer for open-world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19681–19690.

Minka, T. 2000. Estimating a Dirichlet distribution.

Mullappilly, S. S.; Gehlot, A. S.; Anwer, R. M.; Khan, F. S.; and Cholakkal, H. 2024. Semi-supervised open-world object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4305–4314.

Nuriel, O.; Benaim, S.; and Wolf, L. 2021. Permuted adain: Reducing the bias towards global statistics in image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9482–9491.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Shu, Y.; Cao, Z.; Wang, C.; Wang, J.; and Long, M. 2021. Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9624–9633.

Sun, Z.; Li, J.; and Mu, Y. 2024. Exploring orthogonality in open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17302–17312.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, G.; Zhang, X.; Peng, Z.; Tian, S.; Zhang, T.; Tang, X.; and Jiao, L. 2025a. Oral: An observational learning paradigm for unsupervised hyperspectral change detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, G.; Zhang, X.; Peng, Z.; Zhang, T.; Tang, X.; Zhou, H.; and Jiao, L. 2024a. Negative deterministic information-based multiple instance learning for weakly supervised object detection and segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4): 6188–6202.

Wang, K.; Fu, X.; Bao, Y.; Ge, C.; Cao, C.; Zhai, W.; and Zha, Z.-J. 2025b. PAID: Pairwise Angular-Invariant Decomposition for Continual Test-Time Adaptation. *arXiv preprint arXiv:2506.02453*.

Wang, K.; Fu, X.; Ge, C.; Cao, C.; and Zha, Z.-J. 2024b. Towards generalized uav object detection: A novel perspective from frequency domain disentanglement. *International Journal of Computer Vision*, 132(11): 5410–5438.

Wang, K.; Fu, X.; Huang, Y.; Cao, C.; Shi, G.; and Zha, Z.-J. 2023. Generalized uav object detection via frequency domain disentanglement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1064–1073.

Wang, K.; Fu, X.; Lu, X.; Ge, C.; Cao, C.; Zhai, W.; and Zha, Z.-J. 2025c. Efficient Test-time Adaptive Object Detection via Sensitivity-Guided Pruning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10577–10586.

Wang, Z.; Luo, Y.; Qiu, R.; Huang, Z.; and Baktashmotlagh, M. 2021. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 834–843.

Wu, A.; Chen, D.; and Deng, C. 2023. Deep feature deblurring diffusion for detecting out-of-distribution objects. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13381–13391.

Wu, A.; and Deng, C. 2023. TIB: Detecting unknown objects via two-stream information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1): 611–625.

Wu, S.; Chen, H.; Yin, Y.; Hu, S.; Feng, R.; Jiao, Y.; Yang, Z.; and Liu, Z. 2024a. Joint-Motion Mutual Learning for Pose Estimation in Video. In Cai, J.; Kankanhalli, M. S.; Prabhakaran, B.; Boll, S.; Subramanian, R.; Zheng, L.; Singh, V. K.; César, P.; Xie, L.; and Xu, D., eds., *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, 8962–8971. ACM.

Wu, S.; Liu, Z.; Zhang, B.; Zimmermann, R.; Ba, Z.; Zhang, X.; and Ren, K. 2024b. Do as I Do: Pose Guided Human Motion Copy. *IEEE Trans. Dependable Secur. Comput.*, 21(6): 5293–5307.

Wu, S.; Zhang, H.; Liu, Z.; Chen, H.; and Jiao, Y. 2025. Enhancing Human Pose Estimation in Internet of Things via Diffusion Generative Models. *IEEE Internet Things J.*, 12(10): 13556–13567.

Yu, H.; Huang, J.; Liu, Y.; Zhu, Q.; Zhou, M.; and Zhao, F. 2022. Source-free domain adaptation for real-world image dehazing. In *Proceedings of the 30th ACM International Conference on Multimedia*, 6645–6654.

Yuan, Y.; Tang, L.; Chen, Y.; Chen, C.; Huang, Y.; and Ding, X. 2025. ASGS: Single-Domain Generalizable Open-Set Object Detection via Adaptive Subgraph Searching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20911–20921.

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

Zhao, X.; Ma, Y.; Wang, D.; Shen, Y.; Qiao, Y.; and Liu, X. 2023. Revisiting open world object detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhao, Y.; Zhong, Z.; Zhao, N.; Sebe, N.; and Lee, G. H. 2024. Style-hallucinated dual consistency learning: A unified framework for visual domain generalization. *International Journal of Computer Vision*, 132(3): 837–853.

Zheng, J.; Li, W.; Hong, J.; Petersson, L.; and Barnes, N. 2022. Towards open-set object detection and discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3961–3970.

Zhong, Z.; Zhao, Y.; Lee, G. H.; and Sebe, N. 2022. Adversarial style augmentation for domain generalized urban-scene segmentation. *Advances in neural information processing systems*, 35: 338–350.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.

Zohar, O.; Wang, K.-C.; and Yeung, S. 2023. Prob: Probabilistic objectness for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11444–11453.