Novel Scenes & Classes: Towards Adaptive Open-set Object Detection

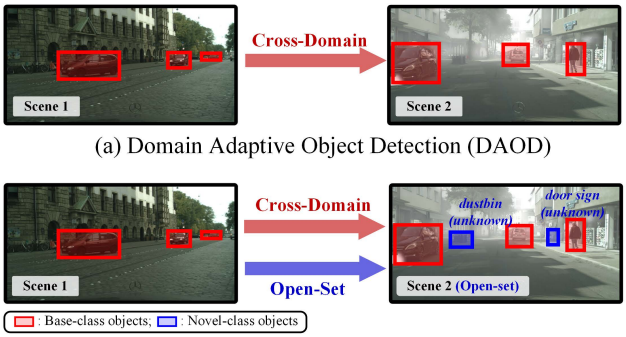
Wuyang Li[1] Xiaoqing Guo[1,2] Yixuan Yuan[1,3,*]
[1]City University of Hong Kong [2]University of Oxford [3]The Chinese University of Hong Kong
wuyangli2-c@my.cityu.edu.hk xiaoqing.guo@eng.ox.ac.uk yxyuan@ee.cuhk.edu.hk

Abstract

Domain Adaptive Object Detection (DAOD) transfers an object detector to a novel domain free of labels. However, in the real world, besides encountering novel scenes, novel domains always contain novel-class objects de facto, which are ignored in existing research. Thus, we formulate and study a more practical setting, Adaptive Open-set Object Detection (AOOD), considering both novel scenes and classes. Directly combing off-the-shelled cross-domain and open-set approaches is sub-optimal since their low-order dependence, e.g.*, the confidence score, is insufficient for the AOOD with two dimensions of novel information. To address this, we propose a novel Structured Motif Matching (SOMA) framework for AOOD, which models the high-order relation with motifs,* i.e.*, statistically significant subgraphs, and formulates AOOD solution as motif matching to learn with high-order patterns. In a nutshell, SOMA consists of Structure-aware Novel-class Learning (SNL) and Structure-aware Transfer Learning (STL). As for SNL, we establish an instance-oriented graph to capture the class-independent object feature hidden in different base classes. Then, a high-order metric is proposed to match the most significant motif as high-order patterns, serving for motif-guided novel-class learning. In STL, we set up a semantic-oriented graph to model the class-dependent relation across domains, and match unlabelled objects with high-order motifs to align the cross-domain distribution with structural awareness. Extensive experiments demonstrate that the proposed SOMA achieves state-of-the-art performance. Codes are available at https://github.com/CityU-AIM-Group/SOMA.*

1. Introduction

Domain Adaptive Object Detection (DAOD) studies robust object detection in cross-domain scenarios, where the independent and identically distributed constraint [17] is no longer applicable due to the domain shift [7]. With tailor-designed techniques [7, 22, 5], the object detector trained in a labeled domain can be generalized to a novel one free of labels (Figure 1(a)), pushing forward its real-world grounding, *e.g.*, self-driving with novel weather and street scenes.



(a) Domain Adaptive Object Detection (DAOD)

(b) The proposed Adaptive Open-set Object Detection (AOOD)

Figure 1. Illustration of (a) existing DAOD assumption and (b) the proposed AOOD setting.

While achieving great success, existing DAOD works strongly assume a shared class space between the two domains (see Figure 1(a)), leading to a vast gap between the transferred domain and real-world [24]. Since natural scenes are diverse and always contain objects beyond pre-defined classes, this gap severely limits the scene understanding for industrial usage [19], *e.g.*, confusing the autonomous driving system with wrong judgments. Hence, to overcome this limitation, we relax the assumption and formulate a more practical setting, called Adaptive Open-set Object Detection (AOOD), by allowing the target domain with novel-class objects. As shown in Figure 1(b), in the target domain, besides detecting the base-class *car* and *person* shared with the source domain, we also aim to identify novel-class objects, *e.g.*, *dustbin* and *door sign*, etc. Specifically, the object detector uses the base-class labels in the source domain for training, and aims to detect base-class objects and identify novel-class objects as *unknown* [19] in the target domain. The proposed AOOD[1] is a practical set-

*Corresponding author.

This work was supported by Hong Kong Research Grants Council General Research Fund 11211221, 14204321, 14220622, and Innovation and Technology Commission-Innovation and Technology Fund ITS/100/20.

[1]See Appendix for comparing with other related task settings.

ting in the real world that considers two dimensions of novel information, namely novel scenes and classes.

Aiming to improve the novel-scene robustness, recent DAOD advances delve into semantic-level cues, *e.g.*, classification scores [50, 30, 22, 29] and prototype-based distance [44, 58], to guide the adversarial alignment [44, 29, 22], metric learning [55, 50, 30], and self-training [26, 25, 56], which achieve cross-domain adaptation at the category level. To improve the novel-class discriminability, existing works rely on classification scores [24, 19, 12, 21, 16, 24] to discover informative background objects, and conduct a novel-class synthesis with base-class sample pairs [59, 14] to achieve open-set learning [12]. The two streams use low-order [35] cues, *e.g.*, the pair-wise distance and relation, to achieve reliable learning with novel information.

However, directly combining cross-domain and open-set approaches may lead to a sub-optimal AOOD. The reason is that the novel-class objects out of distribution and the novel-scene objects attacked by the domain shift are both embedded outside of the labeled feature space with low confidence [19, 7, 21], which are difficult to distinguish with low-order cues. Going beyond the low-order cues, the motif [2, 36, 42], *i.e.*, a statistically significant subgraph, has been studied to model the high-order patterns within a graph. Hence, we aim to use motifs to break through this low-order barrier. Instead of modeling the pair-wise relation between *two entities*, a motif assumes the high-order relation among *several entities* (see Figure 2), which can be used to solve AOOD from the following two aspects. Firstly, we observe that novel-class objects inherently contain class-independent features to distinguish from non-informative backgrounds, *e.g.*, relatively complete contours. These features are shared among different classes and should not be overwhelmed by class-specific semantics. Thus, our critical insight is discovering this shared feature among *several class centers* with motifs, achieving motif-guided novel-class learning[2]. Secondly, for the novel scene, the high-order cues of class-dependent distribution, *e.g.*, within-class variance, are crucial for the cross-domain robustness [1, 30]. These observations motivate us to model the high-order patterns among *several class extreme points* with motifs and achieve motif-guided transfer learning.

To address the critical yet unexplored AOOD issue, we propose a novel Structured Motif Matching (SOMA) framework, which models the high-order patterns with motifs to learn in the real world. Specifically, we propose Structure-aware Novel-class Learning (SNL) to empower novel-class detection in the source domain. SNL estimates class centers and extremes with a semantic bank, which is used to construct an instance-oriented graph to capture class-independent object features. Then, with a newly proposed high-order metric, each candidate object is matched with a graph motif to model significant high-order patterns, serving for motif-guided novel-class learning. Moreover, we design a Structure-aware Transfer Learning (STL) in the target domain for cross-domain transfer. STL constructs a semantic-oriented graph with class extremes, in which the motif is obtained as class-dependent high-order patterns. Then, we use motifs to fulfill motif-guided transfer learning with structural awareness. The proposed motif-based learning paradigm explores the high-order structural patterns well-suited for diverse real-world situations [35]. To be summarized, our main contributions are as follows,

- This paper formulates a real-world friendly setting, Adaptive Open-set Object Detection (AOOD), considering both novel scenes and classes. To address AOOD, we propose a novel Structured Motif Matching (SOMA) framework, which models the high-order patterns with graph motifs for reliable learning.
- We propose a Structure-aware Novel-class Learning (SNL) module, which models the shared object features through motif matching for novel-class learning.
- We design Structure-aware Transfer Learning (STL) for novel scenes. STL models the high-order relation with graph motifs, adapting the cross-domain distribution with structural awareness.
- Three AOOD benchmarks are carefully developed, thoroughly considering different base-novel splitting protocols with practical groundings in the real world. Extensive experiments show that our method achieves state-of-the-art performance in various scenarios.

2. Related Work

2.1. Domain Adaptive Object Detection

Domain Adaptive Object Detection (DAOD) transfers an object detector from a labeled source domain to a novel one with a shared class space. From the perspective of semantic discriminability, existing research can be broadly categorized into adapting class-agnostic distribution [39, 7, 22, 55, 45, 23, 10, 41, 32] and class-aware conditional distribution [58, 49, 29, 30, 50, 44, 11, 38]. As for the methodology, some works [29, 58, 50] estimate and adapt the cross-domain prototypes (class centers) to align the class distribution explicitly. MeGA [44] introduces semantic-level guidance in adversarial alignment in a class-aware manner. SIGMA [30, 31] formulates and solves adaptation with graph matching among dense feature points. Moreover, some works conduct self-training [26, 25, 56, 5] to enhance the discrimination of reliable target-domain samples. However, existing works leverage low-order cues to model the semantic relation, which cannot provide sufficient knowledge for the whole distribution. This work

[2]See Sec. 5 for a high-level clarification of the core idea.

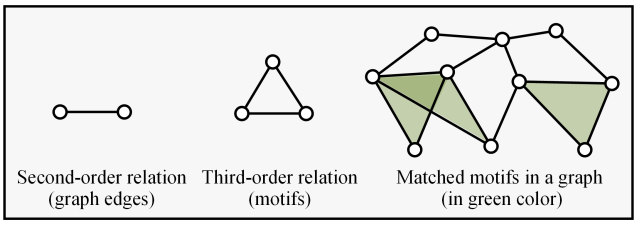


Figure 2. **Left**: the pair-wise low-order relation. Existing OSOD methods aim to discover such knowledge, including the pair-wise distance [19, 24, 12] and pair-wise synthesis [59, 14]; **Middle**: a third-order graph motif; **Right**: matched motifs in a graph.

breaks through this barrier by introducing high-order motifs for robust adaptation.

2.2. Open-Set Object Detection

With labeled base-class objects, Open-Set Object Detection (OSOD) [57, 34, 19, 18, 24, 21, 16, 14, 13] aims to train an object detector to detect both base and novel objects. The authors in [12] benchmark the OSOD problem and then discuss various open-set detectors according to the classifier. Based on Region Proposal Network (RPN) [37], ORE [24] selects non-overlapped proposals with larger objectness scores as potential novel objects and then conducts energy-based novel-class learning. OpenDet [19] optimizes the unknown probability on both base-class and RPN-selected novel-class objects. OW-DETR [18] extends the open-set learning into detection transformer [60] and selects the activated and unmatched object queries for self-training. Moreover, some works [59, 14] use two base-class samples to synthesize novel-class representation for model optimization. Existing works [19, 24, 18, 59, 14] only consider low-order evidence to identify novel objects (Figure 2 **Left**), which cannot model shared object features well. Differently, we model such features with graph motifs to identify novel classes with structural awareness.

2.3. Graph Motif

The usage of graphs in computer vision can be generally categorized into two types. **1) Graph neural networks** [4] model long-range dependencies among pixels, objects, and images, for better model capacity and effective optimization. **2) Graph theories** [47], *e.g.,* graph matching, searching, and clustering introduce theoretically-grounded graphical algorithms in computer vision to solve practice issues and boost the vanilla backpropagation-based optimization.

Beyond the second-order pair-wise relation (Figure 2 **Left**), graph motif [2] is a statistically significant subgraph (Figure 2 **Middle**) with higher-order connectivity patterns, which is essential to understanding the fundamental structures of graphs (Figure 2 **Right**). As a local high-order pattern, graph motif has been studied to improve the graph representation with contrastive learning [42], graph convolution [51], graph attention [36], graph clustering [2]. Graph motifs can model the high-order relation among several graph nodes, yielding better downstream graph learning with structural robustness [2]. Moreover, MotifNet [53] first leverages the graph motif to model the semantic-level structure [53] to address the scene graph parsing. More discussion about the graph in computer vision [4, 47] can be found in Appendix. This work uses graph motifs to model the high-order patterns among categorical knowledge and candidate objects, empowering the novel-class discriminability and novel-scene robustness in object detection.

3. Structured Motif Matching (SOMA)

Problem Formulation for AOOD. We have a labeled source-domain dataset $\mathcal{D}_s = \{I_s^i, Y_s^i\}_{i=1}^{n_s}$ and unlabeled target-domain dataset $\mathcal{D}_t = \{I_t^i\}_{i=1}^{n_t}$ drawn from inconsistent data distribution $P(I_s) \neq P(I_t)$. The source label of each image consists of a set of objects $Y_s = \{(y_s, b_s^x, b_s^y, b_s^w, b_s^h)\}$, where $y_s \in L_s$ represents the object class, and $(b_s^x, b_s^y, b_s^w, b_s^h)$ are the coordinates of bounding boxes. Different from the DAOD assumption that the source and target domain share the same class space, AOOD considers a set of base classes $\Omega_b = \{1, 2, 3, ..., K\}$ and novel classes $\Omega_n = \{K+2, K+3, ..., K+K'\}$ with the following constraints. 1) The labeled objects in the source domain are only in the base class: $L_s \subseteq \Omega_b$. 2) The objects appearing in the unlabeled target domain may be in both base and novel classes, satisfying objects $\subseteq \Omega_b \cup \Omega_n$.

AOOD aims to use $\mathcal{D}_{s/t}$ for training, making it correctly detect base-class objects (objects $\in \Omega_b$) and identify novel-class objects (objects $\in \Omega_n$) as a single *unknown* class (denoted as class $K+1$ [18]) in the target domain $\mathcal{D}_t$.

3.1. Overview

The workflow of SOMA is shown in Figure 3, based on Deformable DETR [60]. With batch-wise labelled source images $\{I_s^i, Y_s^i\}_{i=1}^B$ and unlabeled target images $\{I_t^i\}_{i=1}^B$, we use a feature extractor to extract image-level features $\{X_{s/t}^i\}_{i=1}^B$, and send them to the transformer encoder and decoder [60] to obtain $N = 100$ decoded object queries $Q_{s/t}$. Then, Q_s is sent to SNL (Figure 3(a)) and Q_t is sent to STL (Figure 3(b)). In SNL, we first use ground-truth-matched [3] object queries $Q_m \subseteq Q_s$ to estimate the class space with a semantic bank $\mathbf{Q} = \{\mathbf{Q}_{ctr}, \mathbf{Q}_{ext}\}$, which is used to construct an instance-oriented graph $\mathcal{G}_{IG}$. Then, for each unmatched query $q_{um} \in Q_{um}$, we traverse graph motifs and match the one with minimum high-order metric Γ_{IG}. After that, the matched motifs $\mathcal{M}_{IG}^{mat}$ serve as high-order patterns for motif-guided novel-class learning $\mathcal{L}_{SNL}$, and are used to enrich the semantic bank $\mathbf{Q}$ in turn. In STL, we use activated queries $Q_{act} \subseteq Q_t$ and class extremes $\mathbf{Q}_{ext}$ to establish a semantic-oriented graph $\mathcal{G}_{SG}$. Then, the

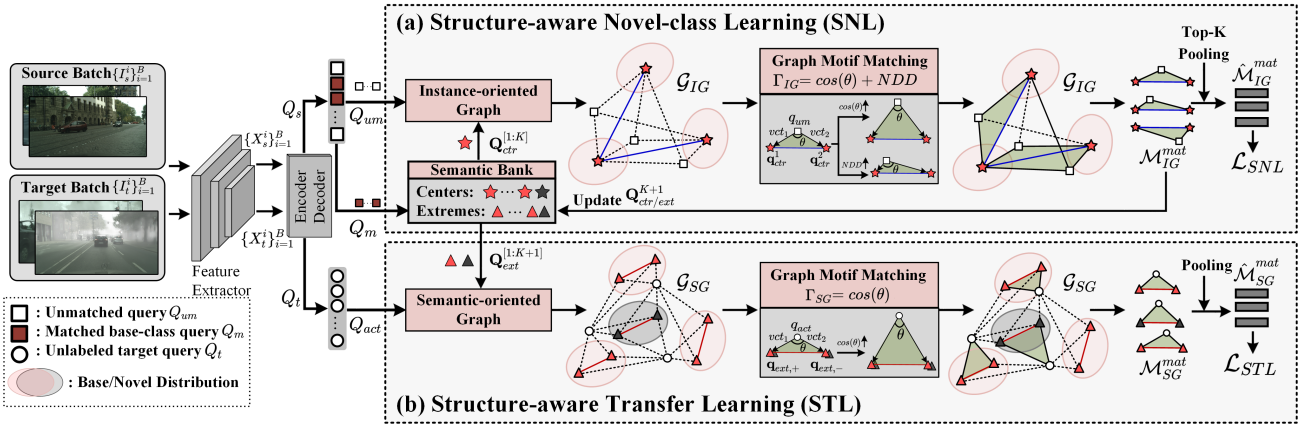


Figure 3. Overview of the proposed SOMA framework for AOOD, consisting of SNL for detecting novel objects and STL for adapting to novel scenes. SOMA is used in the model training and is eliminated in the inference stage.

graph motif is matched $\mathcal{M}_{SG}^{mat}$ with the high-order metric Γ_{SG}, achieving a motif-guided transfer learning with $\mathcal{L}_{STL}$.

3.2. Structure-aware Novel-class Learning

Given batch-wise source $\{I_s^i, Y_s^i\}_{i=1}^B$ and target $\{I_t^i\}_{i=1}^B$ samples with K base classes, the feature extractor and transformer are deployed to obtain decoded object queries $Q_{s/t} \in \mathbb{R}^{BN\times D}$ [60]. With object labels and predictions, some queries $Q_m \subseteq Q_s$ are matched to the ground-truth via bipartite matching [3], while the rest queries $Q_{um} \subseteq Q_s$ may contain novel-class and background information [18].

Considering that the randomness of batch-level sampling may lead to a biased observation [30, 28], we first collect the ground-truth-matched queries Q_m to estimate a semantic-complete class space with a semantic bank $\mathbf{Q}$. Specifically, the bank saves class centers $\mathbf{Q}_{ctr} \in \mathbb{R}^{(K+1)\times D}$ to model class prototypes [58], and preserves the information of the class extreme pairs $\mathbf{Q}_{ext} = \{\mathbf{Q}_{ext,+}, \mathbf{Q}_{ext,-}\}$, $\mathbf{Q}_{ext,-/+} \in \mathbb{R}^{(K+1)\times D}$ to estimate the scale of the class distribution, both of which are randomly initialized at the training start. Since the extreme points are difficult to estimate directly, we introduce and save the standard deviation as the intermediate variable $\mathbf{Q}_{std}$ for a batch-level estimation, following the assumption of a formalized deep feature space [1]. To this end, for the learning of $\mathbf{Q}$, we conduct an Exponential Moving Average (EMA) based update strategy for each base-class $k \in \{1, 2, ..., K\}$ with matched queries Q_m^k:

$$\begin{aligned} \mathbf{Q}_{ctr}^k &\leftarrow \alpha f_{\text{mean}}(Q_m^k) + (1-\alpha)\mathbf{Q}_{ctr}^k, \\ \mathbf{Q}_{std}^k &\leftarrow \alpha f_{\text{std}}(Q_m^k) + (1-\alpha)\mathbf{Q}_{std}^k, \\ \mathbf{Q}_{ext,+}^k = \mathbf{Q}_{ctr}^k + \beta\mathbf{Q}_{std}^k, &\mathbf{Q}_{ext,-}^k = \mathbf{Q}_{ctr}^k - \beta\mathbf{Q}_{std}^k, \end{aligned} \tag{1}$$

where $f_{\text{mean/std}}(\cdot)$ indicates the statistical mean and standard deviation of the observed object queries, $\alpha \in (0, 1)$ controls the update speed, and β scales the estimated distribution. For the novel-class $\mathbf{Q}_{ctr/std}^{K+1}$, we generate a class-independent placeholder by averaging all base-class representations $\frac{1}{K}\sum \mathbf{Q}_{ctr/std}^K$ as the update item, since novel objects contain the shared object features among all classes.

With base-class centers $\mathbf{Q}_{ctr}^{[1,K]}$ and unmatched queries Q_{um}, we construct an instance-oriented graph to model the relation among different classes. Then, object queries will be matched to the most informative motif, serving as high-order patterns for motif-guided novel-class learning.

Instance-oriented Graph. To model the structural relation for novel-class learning, we establish an instance-oriented graph $\mathcal{G}_{IG}$ with class centers $\mathbf{Q}_{ctr}^{[1,K]}$ and unmatched queries Q_{um}, which discovers the shared instance-level knowledge among different classes. Specifically, we first connect each base-class center $\mathbf{q}_{ctr}^1 \in \mathbf{Q}_{ctr}^{[1,K]}$ with its farthest counterpart[3] $\mathbf{q}_{ctr}^2 \in \mathbf{Q}_{ctr}^{[1,K]}$ according to $L2$ distance (blue edges in Figure 3(a)) to establish the pair-wise relation between the two most different classes, *e.g. car* and *person*. The built relation is critical in discovering the shared class-independent knowledge, which can identify object instances from non-informative backgrounds. Then, each query $q_{um} \in Q_{um}$ is linked with all base-class centers $\mathbf{Q}_{ctr}^{[1,K]}$ to model inherent relation candidates (dotted edges in Figure 3(a)). This graph $\mathcal{G}_{IG}$ is established across different images within a batch, capturing the long-distance dependence and modeling inherent class-independent relations among nodes.

Graph Motif Matching. Aiming at extracting high-order patterns in the graph $\mathcal{G}_{IG}$, we match each query q_{um} with the most *informative* motif M_{IG}^{mat} satisfying *topological evidence*. As for the *topological evidence*, we define the fully-connected sub-graph with three nodes, *i.e.*, a *triangle*-like structure, as the available third-order motif [2]. Then, with the well-established graph structure, we traverse available

[3] We prevent two similar classes (*e.g.*, *car* and *truck*) overwhelmed by class-dependent patterns. See Appendix for justifications.

graph motifs for each unmatched query $q_{um} \in Q_{um}$ and generate a motif candidate set $\mathcal{M}_{IG} = \{M_1, M_2, ...M_n\}$, where $M_i = \{q_{um}, \mathbf{q}^1_{ctr}, \mathbf{q}^2_{ctr}\}$. After that, the optimal motif $M^{mat}_{IG} \in \mathcal{M}_{IG}$ is selected as novel-class representation.

To find the most *informative* motif M^{mat}_{IG} in each candidate set $\mathcal{M}_{IG}$, we propose a high-order metric Γ_{IG} that measures the quality of class-independent knowledge in a motif. Formally, given the motif with three node variables $M_i = \{q_{um}, \mathbf{q}^1_{ctr}, \mathbf{q}^2_{ctr}\}$, the high-order metric is defined as $\Gamma_{IG} = cos(\theta) + NDD$ with the followed items,

$$\begin{aligned} cos(\theta) &:= \frac{vct_1 \cdot vct_2}{||vct_1||_2 \cdot ||vct_2||_2}, \\ NDD &:= |\frac{||vct_1||_2 - ||vct_2||_2}{||\mathbf{q}^1_{ctr} - \mathbf{q}^2_{ctr}||_2}|, \end{aligned} \tag{2}$$

where $vct_{1/2} := \mathbf{q}^{1/2}_{ctr} - q_{um}$ are two intermediate vectors to measure the motif structure (Figure 3(a) Middle), $cos(\theta) \in [-1, 1]$ measures the angle between vct_1 and vct_2, and $NDD \in [0, 1]$ is normalized distance difference, reflecting the q_{um}'s affinity difference between $\mathbf{q}^{1/2}_{ctr}$. The insight of the two items is as follows. If $cos(\theta)$ gives a small value, *i.e.*, a large θ, the distance between node q_{um} and the edge connecting $\mathbf{q}^1_{ctr}$ and $\mathbf{q}^2_{ctr}$ will be small, indicating that q_{um} is likely to be similar to $\mathbf{q}^{1/2}_{ctr}$. If NDD is small, then, the node q_{um} tends to be embedded near the central axis between $\mathbf{q}^1_{ctr}$ and $\mathbf{q}^2_{ctr}$ instead of $\mathbf{q}^1_{ctr}$ or $\mathbf{q}^2_{ctr}$ itself.

Hence, we select the motif $M^{mat}_{IG} = \text{argmin}_{\Gamma_{IG}} \mathcal{M}_{IG}$ with minimum Γ_{IG}. Minimizing $cos(\theta)$ encourages the motif to contain more class-independent knowledge. Since $\mathbf{q}^1_{ctr}$ and $\mathbf{q}^2_{ctr}$ represent the most inconsistent semantics, *e.g.*, *car* and *person*, q_{um} is likely to contain vital class-independent information if it shows sufficient similarity on both $\mathbf{q}^{1/2}_{ctr}$ in this min-max game [17]. Meanwhile, minimizing NDD encourages a better class-independent property, avoiding the query q_{um} overwhelmed by a specific base class[4]. Thus, we combine the two structural cues with orthogonal effects to select informative motifs for novel-class learning, yielding a matched motif set $\mathcal{M}^{mat}_{IG} = \{M^{mat}_{IG}\}$. Please refer to Figure 5 for the clear justification of Γ_{IG}.

Motif-guided Novel-class Learning. With matched motifs $\mathcal{M}^{mat}_{IG}$, we further conduct motif-guided novel-class learning with extracted high-order patterns. Considering that the background knowledge is hidden in $\mathcal{M}^{mat}_{IG}$, we rank the high-order metric Γ_{IG} for matched motifs and select Top-K minimums to optimize the novel-class posterior:

$$\mathcal{L}_{SNL} = -\frac{1}{\mathrm{K}} \sum_{i=1}^{\mathrm{K}} log(p(f_{cls}(\hat{M}^{mat}_{IG,i}) = K+1 | \hat{M}^{mat}_{IG,i})),$$

where $f_{cls}(\cdot)$ is the classifier in the object detector, $\hat{M}^{mat}_{IG} = \frac{1}{3}\sum_{n=1}^{3} M^{mat}_{IG,n}$ is the abstracted representation for selected motifs. In addition to optimizing the novel-class discriminability with Q_{um} [18], our motif-based strategy empowers the novel-class learning for base-class $\mathbf{Q}^{[1,K]}_{ctr}$ [19], being aware of the class-independence among varied classes. Finally, we use selected motifs to enrich the semantic bank in $\mathbf{Q}^{K+1}_{ctr/ext}$ by updating the novel-class item as Eq. 1.

3.3. Structure-aware Transfer Learning

With class extremes $\mathbf{Q}^{[1,K+1]}_{ext}$ and target-domain queries Q_t, we construct a semantic-oriented graph to model the relation within the class. Then, with the graph structure, unlabeled nodes are matched to the graph motif, serving to remedy the domain gap with motif-guided learning.

Semantic-oriented Graph. As the domain shift severely attacks the class distribution [58, 30] with weak semantic discriminability, we establish a semantic-oriented graph $\mathcal{G}_{SG}$ to model the semantic-level relation with base-class extremes $\mathbf{Q}^{[1,K+1]}_{ext}$ and activated object queries $Q_{act} \subseteq Q_t$. Specifically, considering the one-vs-all property of the sigmoid classifier, we select activated object queries $q_{act} \in Q_{act}$ according to the pseudo classification scores $Q_{act} = \{q_t | \sum_{k=1}^{K+1} f_{cls}(q^k_t) > \epsilon\}$ [5]. Then, we connect the two extreme points $\mathbf{Q}^k_{ext,+}$ $\mathbf{Q}^k_{ext,-}$ in the same class $k = \{1, 2, ..., K+1\}$ to model the class distribution (red edges in Figure 3(b)), and connect each object query q_{act} with all $\mathbf{Q}^{[1,K+1]}_{ext}$ (dotted edges in Figure 3(b)) to model the candidate relations. Thus, the edges contain abundant class-specific relation, serving for a structure-aware adaptation.

Graph Motif Matching. With the semantic-oriented graph $\mathcal{G}_{SG}$, each activated query q_{act} will be matched to the optimal motif M^{mat}_{SG} as the high-order pattern for robust domain adaptation. Specifically, for each $q_{act} \in Q_{act}$, we traverse all available graph motifs to generate a motif candidate set $\mathcal{M}_{SG} = \{M_1, M_2, ...M_n\}, M_i = \{q_{act}, \mathbf{q}_{ext,+}, \mathbf{q}_{ext,-}\}$. Similarly, the intermediate vectors $vct_{1/2} := \mathbf{q}_{ext,+/-} - q_{act}$ are defined (Figure 3(b) Middle) and used to match the motif with minimum metric $\Gamma_{SG} := cos(\theta)$, where $cos(\theta) := \frac{vct_1 \cdot vct_2}{||vct_1||_2 \cdot ||vct_2||_2}$. Unlike the motif matching in $\mathcal{G}_{IG}$, we only use the angle item for $\mathcal{G}_{SG}$ (see Table 5), since a large NDD, *i.e.*, q_{act} is near one of the extremes, can still ensure the correct adaptation with the same class distribution. Thus, with structural cues, each activated query q_{act} is matched to a high-order motif $M^{mat}_{SG} = \text{argmin}_{\Gamma_{SG}} \mathcal{M}_{SG}$ drawn from a specific class distribution.

Motif-guided Transfer Learning. With matched high-order graph motifs $\mathcal{M}^{mat}_{SG} = \{M^{mat}_{SG}\}$, each activated query q_{act} is matched to a specific class with corresponding pseudo-label $\tilde{y}_3$, serving for a robust cross-domain transfer. Considering that the second-order knowledge $\tilde{y}_2$ is also critical for semantic discriminability (see Table 6), we conduct a cross-order mixup to obtain the ensembled pseudo-labels:

[4] These overwhelmed queries may be caused by the heavy overlapping [27, 61] (see Figure 4) and should not be treated as novel-class.

[5] ϵ is empirically set 0.5 as [30] to satisfy the sigmoid activation.

Method	Set	num. novel-class: 3				num. novel-class: 4				num. novel-class: 5			
		mAP_b↑	AR_n↑	WI↓	AOSE↓	mAP_b↑	AR_n↑	WI↓	AOSE↓	mAP_b↑	AR_n↑	WI↓	AOSE↓
DDETR [60]$_{ICLR'21}$	**het-sem**	47.52	0.00	0.341	459	45.24	0.00	0.506	1028	42.38	0.00	0.659	1968
PROSER [59]$_{CVPR'21}$		46.92	1.80	0.271	218	44.19	2.02	0.415	531	41.99	2.00	0.584	1127
OpenDet [19]$_{CVPR'22}$		47.04	1.92	0.269	221	45.71	1.89	0.499	511	42.09	1.70	0.579	922
OW-DETR [18]$_{CVPR'22}$		43.31	1.84	0.432	192	42.52	2.10	0.619	451	39.92	1.98	0.684	814
SOMA (ours)		**50.87**	**3.78**	**0.268**	**139**	**48.06**	**4.41**	**0.412**	**340**	**45.55**	**4.08**	**0.526**	**649**
DDETR [60]$_{ICLR'21}$	**hom-sem**	44.62	0.00	1.860	2937	43.55	0.00	2.000	3565	40.18	0.00	2.462	6770
PROSER [59]$_{CVPR'21}$		43.15	4.59	1.842	2146	43.31	4.99	2.018	2641	39.99	5.99	2.563	4963
OpenDet [19]$_{CVPR'22}$		45.51	5.28	1.336	1458	44.02	5.67	1.653	1798	40.87	6.58	2.303	3416
OW-DETR [18]$_{CVPR'22}$		43.22	3.15	1.355	1076	42.83	3.46	1.593	1320	39.45	4.38	2.384	3399
SOMA (ours)		**48.67**	**6.96**	**1.257**	**915**	**47.02**	**7.42**	**1.527**	**1232**	**43.37**	**8.42**	**2.281**	**2886**
DDETR [60]$_{ICLR'21}$	**freq-dec**	56.99	0.00	0.579	1240	55.02	0.00	0.835	2136	53.89	0.00	0.93	2625
PROSER [59]$_{CVPR'21}$		55.70	6.68	0.589	**536**	54.51	7.88	0.780	952	53.43	8.22	0.943	**1072**
OpenDet [19]$_{CVPR'22}$		57.28	9.35	0.519	720	54.89	10.59	0.781	1251	53.51	10.37	0.839	1470
OW-DETR [18]$_{CVPR'22}$		56.63	6.61	0.585	698	55.45	7.90	0.745	**930**	53.60	7.90	0.807	1105
SOMA (ours)		**59.18**	**11.41**	**0.507**	669	**56.85**	**12.47**	**0.723**	1140	**55.63**	**12.36**	**0.759**	1315
DDETR [60]$_{ICLR'21}$	**freq-inc**	44.72	0.00	2.862	2859	43.91	0.00	3.270	4907	41.12	0.00	3.609	8291
PROSER [59]$_{CVPR'21}$		44.23	2.94	2.881	1090	42.47	2.98	2.745	1866	39.11	3.01	3.119	3242
OpenDet [19]$_{CVPR'22}$		44.85	3.23	2.579	1700	42.92	3.30	2.741	2835	40.34	3.44	2.970	4965
OW-DETR [18]$_{CVPR'22}$		43.92	3.85	2.032	1377	43.01	3.99	2.219	1891	40.21	2.98	2.184	2293
SOMA (ours)		**46.62**	**8.32**	**1.452**	**733**	**47.30**	**8.43**	**1.566**	**1166**	**44.45**	**7.95**	**1.792**	**1974**

Table 1. Comparison results on Cityscapes→Foggy Cityscapes under AOOD setting.

Method	$\|\Omega_n\|$	mAP_b ↑	AR_n ↑	WI↓	AOSE↓
DDETR [60]$_{ICLR'21}$	6	19.78	0.00	8.95	6347
PROSER [59]$_{CVPR'21}$		18.23	32.37	9.87	5853
OpenDet [19]$_{CVPR'22}$		20.57	41.15	8.93	4295
OW-DETR [18]$_{CVPR'22}$		20.31	35.48	10.26	5184
SOMA (ours)		**21.70**	**43.15**	**7.32**	**4278**
DDETR [60]$_{ICLR'21}$	8	19.31	0.00	9.58	7402
PROSER [59]$_{CVPR'21}$		18.37	33.07	10.40	6636
OpenDet [19]$_{CVPR'22}$		20.84	41.58	9.53	**4919**
OW-DETR [18]$_{CVPR'22}$		21.01	36.53	10.52	5981
SOMA (ours)		**21.69**	**43.40**	**8.24**	5016
DDETR [60]$_{ICLR'21}$	10	19.12	0.00	10.06	9198
PROSER [59]$_{CVPR'21}$		16.80	33.74	11.06	8065
OpenDet [19]$_{CVPR'22}$		18.87	41.50	10.24	6103
OW-DETR [18]$_{CVPR'22}$		18.42	36.50	11.06	7018
SOMA (ours)		**20.09**	**43.73**	**8.88**	**6092**

Table 2. Comparison results on Pascal VOC→ Clipart with AOOD setting. $|\Omega_n|$ indicates the number of novel classes.

$\tilde{y} = 0.5 \cdot \tilde{y}_3 + 0.5 \cdot \tilde{y}_2$, where $\tilde{y}_2 = \text{argmin}(||q_{act} - \mathbf{Q}_{ctr}||_2)$ is the low-order evidence. Then, the motif-guided learning is implemented with the following loss function,

$$\mathcal{L}_{STL} = -\frac{1}{|\mathcal{M}_{SG}^{mat}|} \sum_{i=1}^{|\mathcal{M}_{SG}^{mat}|} \tilde{y} log(p(f_{cls}(\hat{M}_{SG,i}^{mat}) | \hat{M}_{SG,i}^{mat})),$$

where $\hat{M}_{SG}^{mat} = \frac{1}{3}\sum_{n=1}^{3} M_{SG,n}^{mat}$ indicates the pooled motif representation with high-order patterns. The high-order relation between the source and target domain can be modeled in each graph motif, encouraging the robust alignment of per-class distribution with structural awareness.

3.4. Model Optimization

During the training stage of SOMA framework, we implement the overall optimization objective as follows,

$$\mathcal{L} = \lambda_1 \mathcal{L}_{SNL} + \lambda_2 \mathcal{L}_{STL} + \mathcal{L}_{BASE}, \quad (3)$$

where $\mathcal{L}_{SNL}$ is the proposed structure-aware novel-class learning loss, $\mathcal{L}_{STL}$ is used for structure-aware transfer learning, and $\mathcal{L}_{BASE}$ is term shared in all benchmark counterparts, including DETR detection loss [60] and global alignment loss [7]. $\lambda_{1/2}$ are set to 0.1 and 0.01, respectively.

4. Experiments

4.1. Benchmark Setup

Dataset Settings. We develop three AOOD benchmarks using the datasets [7, 39] with the severe domain gap. 1) *Cityscapes→Foggy Cityscapes.* Cityscapes [8] is a street scene dataset captured under the dry weather condition, including *train* set (2975 images) and *val* set (500 images) with 8 classes. *Foggy Cityscapes* [40] is a synthesized dataset based on the Cityscapes, whose 0.02 foggy-level sub-set is used for comparison. 2) *Pascal VOC→Clipart.* Pascal VOC [15] is a real-world dataset with annotated commonly seen objects. Following [39], we use Pascal VOC 2007/2012 *trainval* split with 16,551 images for training. Clipart [39] are collected from the website with images in abstract styles, containing 1,000 images for training and test [6]. 3) *Cityscapes→BDD100k.* BDD100k [52] is a large-scale landscape dataset with 100K videos. We follow [48, 45] to use the daytime subset with 36,728 for training and 5,258 images for evaluation.

For Cityscapes→Foggy Cityscapes/BDD100K, we split the base-/novel-class along two dimensions considering diverse real-world situations [43] with **12** sub-tasks. Firstly, we consider different base/novel-class splittings about *semantic overlapping* and *instance frequency*, including **4** sub-tasks: 1) heterogeneous semantics (**het-sem**): no semantic overlap between base and novel classes (*e.g.*, car and person), 2) homogeneous semantics (**hom-sem**): with semantic overlaps (*e.g.*, car and truck), 3) frequency decrease (**freq-dec**): more base objects than novel counterparts, and 4) frequency increase (**freq-inc**): more novel objects than

Method	Set	num. novel-class: 3				num. novel-class: 4				num. novel-class: 5			
		mAP_b ↑	AR_n ↑	WI↓	AOSE↓	mAP_b ↑	AR_n ↑	WI↓	AOSE↓	mAP_b ↑	AR_n ↑	WI↓	AOSE↓
DDETR [60]$_{ICLR'21}$	het-sem	13.48	0.00	0.153	1448	13.49	0.00	0.164	1604	13.52	0.00	0.227	2378
PROSER [59]$_{CVPR'21}$		13.32	1.53	0.148	910	13.35	1.48	0.163	1032	13.37	1.60	0.218	1466
OpenDet [19]$_{CVPR'22}$		13.70	1.20	0.135	836	13.71	1.17	0.150	992	13.75	1.27	0.209	1244
OW-DETR [18]$_{CVPR'22}$		13.15	1.27	0.129	792	13.15	1.27	0.157	908	13.50	1.30	0.201	1168
SOMA (ours)		**14.11**	**1.86**	**0.127**	**614**	**14.10**	**1.90**	**0.145**	**732**	**14.13**	**2.01**	**0.197**	**1074**
DDETR [60]$_{ICLR'21}$	hom-sem	10.31	0.00	2.846	25530	10.32	0.00	2.873	26488	10.56	0.00	3.003	29812
PROSER [59]$_{CVPR'21}$		9.17	2.38	2.525	13200	9.19	2.41	2.458	13684	9.40	2.58	3.067	15962
OpenDet [19]$_{CVPR'22}$		10.50	3.26	2.308	9760	10.54	3.28	2.327	10126	10.84	3.41	2.861	11776
OW-DETR [18]$_{CVPR'22}$		9.45	1.45	2.255	**6236**	9.47	1.46	2.372	9440	10.52	1.64	2.780	10088
SOMA (ours)		**11.51**	**3.97**	**2.251**	7670	**11.53**	**4.01**	**2.312**	**8054**	**11.83**	**4.13**	**2.611**	**9968**
DDETR [60]$_{ICLR'21}$	freq-dec	15.91	0.00	0.908	7402	15.88	0.00	0.952	8166	15.86	0.00	1.258	13044
PROSER [59]$_{CVPR'21}$		15.98	12.92	0.949	4320	15.76	12.54	0.987	4886	12.88	15.57	1.286	7504
OpenDet [19]$_{CVPR'22}$		16.01	14.87	0.948	4254	16.04	14.36	0.932	4942	16.11	14.69	1.250	7988
OW-DETR [18]$_{CVPR'22}$		15.80	9.68	0.963	4294	15.76	9.31	1.021	4756	15.81	9.60	1.379	7738
SOMA (ours)		**16.81**	**15.67**	**0.869**	**4220**	**16.55**	**15.05**	**0.915**	**4654**	**16.63**	**15.59**	**1.181**	**7230**
DDETR [60]$_{ICLR'21}$	freq-inc	10.02	0.00	3.054	22108	10.02	0.00	3.08	23060	10.18	0.00	3.219	25684
PROSER [59]$_{CVPR'21}$		9.02	1.71	3.995	24118	8.95	1.72	4.019	25366	9.80	1.77	4.202	28170
OpenDet [19]$_{CVPR'22}$		10.47	1.68	3.228	13578	10.30	1.70	3.282	14210	10.46	1.73	3.393	15928
OW-DETR [18]$_{CVPR'22}$		8.11	1.75	2.785	9602	8.12	1.75	2.787	9960	8.34	1.76	2.867	11034
SOMA (ours)		**11.17**	**4.56**	**2.556**	**7420**	**11.08**	**4.56**	**2.577**	**7762**	**11.71**	**4.53**	**2.713**	**8844**

Table 3. Comparison results on Cityscapes→ BDD100k with AOOD setting.

base ones. Secondly, we also follow [19] to consider the number of novel classes with **3** sub-tasks $\{3, 4, 5\}$. For Pascal VOC→Clipart, we follow existing work [19] by considering $\{6, 8, 10\}$ novel classes according to the alphabetical order, yielding **3** sub-tasks for comparison. See Appendix for more splitting details and discussions.

Evaluation Metrics. We use mean Average Precision with a 0.5 IoU threshold (mAP_b) to evaluate the base-class performance. For the novel-class evaluation, we follow the main stream [24] to use Average Recall (AR_n), Wilderness Impact (WI) [12], and Absolute Open-Set Error (AOSE) [12] for strict evaluation. WI and AOSE measure the confusion in predicting a novel objects as base classes.

Implementation Details. Benchmarked methods are implemented on Deformable DETR [60] with backbone alignment [17]. We adopt ResNet-50 [20] feature extractor pretrained with DINO [54] to avoid the novel-class leakage [18] of the ImageNet [9] pretraining. Our model is trained with the AdamW optimizer [33] with a 0.0002 learning rate, four batch-size, and weight decay of 5×10^{-4} on two NVIDIA V100 GPUs. We use 3 transformer encoder and decoder layers and implement SOMA on the last decoder layer for computation efficiency on small-scale datasets [46]. The training schedule is as [18] with extra epochs of warm-up. During optimizing $\mathcal{L}_{SNL}$, we select Top-5 motifs to generate novel-class signals and encourage the unknown score of matched object queries to $\epsilon = 0.5$ in classification to learn the unknown probability of base-class objects. The update factor α and scaling factor β in the semantic bank are set 0.01 and 1.0, respectively.

4.2. Benchmark Comparison

Cityscapes→Foggy Cityscapes. The comparison is shown in Table 1. Compared with DDETR [60] baseline, we observe that existing works are prone to sacrifice large performance on base class precision, *e.g.*, OW-DETR [18] gives 39.92% (42.38%), 39.45% (40.18%), 53.6% (53.89%), and 39.89% (41.1%) mAP_b from set 1 to set 4 with 5 novel classes. Differently, the proposed SOMA gives significant gains with 45.55%, 42.71%, 55.61%, and 41.58% mAP_b, outperforming existing works by a large margin. For the novel-class evaluation, the proposed SOMA achieves the best AR_n on all 12 sub-tasks, verifying the effectiveness and robustness of the proposed method in various settings.

SNL	STL	mAP_b↑	AR_n↑	WI↓	AOSE↓
✗	✗	42.38	0.00	0.659	1968
✓	✗	42.88	3.31	0.608	834
✗	✓	44.70	0.88	0.658	1791
✓	✓	**45.55**	**4.08**	**0.526**	**649**

Table 4. Ablation study on Cityscapes→Foggy Cityscapes under het-sem setting (5 novel classes).

	$cos(\theta)$	NDD	mAP_b↑	AR_n↑	WI↓	AOSE↓
Γ_{IG}	✓		44.79	3.17	0.562	766
		✓	43.85	1.37	0.529	1158
	✓	✓	**45.55**	**4.08**	**0.526**	**649**
Γ_{SG}	✓		**45.55**	4.08	**0.526**	**649**
		✓	44.03	3.92	0.597	949
	✓	✓	45.01	**4.11**	0.531	672

Table 5. Comparison on Cityscapes→Foggy Cityscapes (het-sem) with varied high-order metric $\Gamma_{IG/SG}$ designs.

	sets	mAP_b↑	AR_n↑	WI↓	AOSE↓
SNL	w/o. SNL	44.70	0.88	0.658	1791
	SNL w. Top-3	**46.00**	3.80	0.687	789
	SNL w. Top-5	45.55	**4.08**	**0.526**	649
	SNL w. Top-7	45.31	4.03	0.579	**582**
STL	w/o. STL	42.88	3.31	0.608	834
	STL w. $\tilde{y}_3$	44.05	3.91	0.599	801
	STL w. $\tilde{y}_2$ & $\tilde{y}_3$	**45.55**	**4.08**	**0.526**	**649**

Table 6. Comparison results on Cityscapes→Foggy Cityscapes (het-sem) with different hyper-parameter settings.

Figure 4. Comparison on Cityscapes→Foggy Cityscapes (freq-dec) among (a) DDETR [60], (b) OW-DETR [18], (c) the proposed SOMA.

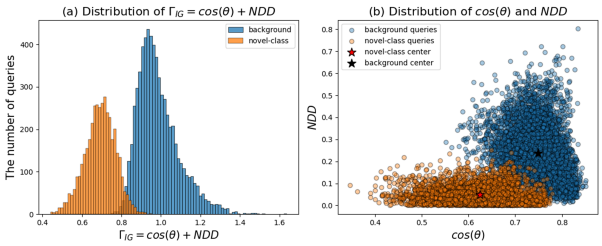


Figure 5. Distribution of novel classes and backgrounds about (a) the high-order metric Γ_{IG} and (b) $cos(\theta)$ and NDD.

Pascal VOC→Clipart. As shown in Table 2, SOMA achieves the best mAP$_b$ and AP$_n$ on all sub-tasks on this semantic-rich benchmark. For the novel-class comparison, SOMA gives 43.15%, 43.4% and 43.73% AP$_n$, surpassing state-of-the-art OW-DETR [18] (35.48%, 36.53% and 36.5%) by a large margin. For the base-class comparison, SOMA gives 21.7%, 21.69% and 20.09% mAP$_b$, outperforming state-of-the-art OW-DETR [18] (20.31%, 21.69% and 18.42%) comprehensively and significantly.

Cityscapes→BDD100k. As shown in Table 3. the proposed SOMA achieves satisfactory and robust performance on all four settings with different numbers of novel-class objects, yielding the best mAP$_b$ and AP$_n$ for all 12 sub-tasks. Compared with state-of-the-art method OW-DETR [18] (1.3%, 1.64%, 9.6% and 1.76% AR$_n$), SOMA surpasses it by a large margin with 2.01%, 4.13%, 15.59% and 4.53% AR$_n$ from set 1 to set 4 under 5 novel-class evaluation.

4.3. Quantitative Analysis

Ablation Study. The detailed ablation studies are shown in Table 4. Compared with baseline [60], introducing SNL significantly boosts the novel-class recall (from 0.0% to 3.31% AR$_n$), and reduces WI (from 0.659 to 0.608) and AOSE (from 1968 to 834). Introducing STL improves cross-domain adaptation with a better 44.7% mAP$_b$, surpassing the baseline (42.38% mAP$_b$) with 2.32% points. Moreover, we observe that STL is able to empower the novel-class detection with a 0.88% AR$_n$ due to the matched motifs with novel-class extremes. After integrating the two parts, the proposed SOMA performs best and outperforms baseline [60] with 3.17% mAP$_b$ and 4.08% AR$_n$, verifying the complementary benefits between them.

High-order Metric Design. The two items in high-order metric $\Gamma_{IG/SG}$ are analyzed in Table 5. As for Γ_{IG}, introducing $cos(\theta)$ and NDD together gives the best open-set results (4.08% AR$_n$, 0.526 WI, and 649 AOSE), encouraging a better feature space for the novel-class learning. Moreover, using $cos(\theta)$ for Γ_{SG} achieves the best 45.5% mAP$_b$ with the optimal adaptation, verifying our practical design.

Hyper-parameter Sensitivity. As shown in Table 6, we analyze the hyper-parameters in each module. In SNL, reducing the number of selected motifs (Top-3) gives a better 46.0% mAP$_b$, while increasing the motif number (Top-7) provides a better AOSE (582). Our optimal setting with Top-5 achieves a balanced performance on all evaluation metrics. For SNL, we compare the pseudo-labeling with distance-based $\tilde{y}_2$ and motif-based $\tilde{y}_3$ cues. Combining cross-order cues can give the best performance with 45.55% mAP$_b$ and 4.08% AR$_n$, verifying our effective design.

4.4. Qualitative Analysis

Comparing AOOD Predictions. We make a comparison[6] with DDETR baseline [60] and OW-DETR [18] in Figure 4. Compared with OW-DETR [18], SOMA not only detects high-quality novel objects beyond the Cityscapes [8] label space, *e.g.*, *dustbin* (1^{st} image), but also identifies predefined novel objects, *e.g.*, *bus* (2^{nd} image) and *truck* (3^{rd}

[6]We remove the open-set predictions sharing queries with selected base-class predictions as a post-processing procedure.

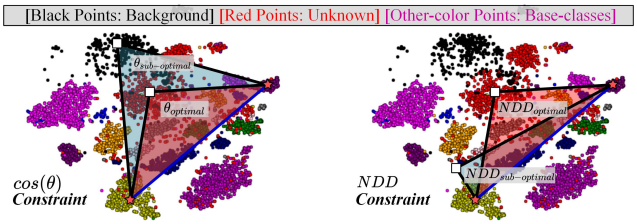


Figure 6. T-SNE feature visualization on the Pascal→Clipart.

image). Moreover, we observe that OW-DETR [18] tends to wrongly consider local parts inside a base-class object as unknown, *e.g.*, the *car* in 4^{th} image, which can be relieved by our method due to the effective NDD constraint in Γ_{IG}.

Justifying the High-order Metric. In Figure 5(a), we randomly sample background and novel-class queries and plot the distribution of the high-order metric Γ_{IG}. We observe that novel-class objects can be successfully separated from backgrounds via a smaller Γ_{IG}, verifying the practical design in selecting the motifs with smaller Γ_{IG}. To delve into the metric design, we further plot $cos(\theta)$ (x-axis) and NDD (y-axis) in Figure 5(b). The novel class distribution is at the left-bottom of the background, revealing a smaller $cos(\theta)$ (*i.e.*, more significant affinity with two most different base classes) and a smaller NDD (*i.e.*, preventing being too near one base-class). Hence, the rationality of the proposed high-order terms (Sec. 3.2) can be justified.

T-SNE Feature Visualization. As shown in Figure 6, we further conduct T-SNE feature visualization and highlight the optimal and sub-optimal motifs with red and blue triangles, respectively. We observe that background regions can be successfully distinguished with a larger $cos(\theta)$ (*left*), and ambiguous objects can be identified with a larger NDD (*right*), as our practical design. Moreover, it can be observed that the sub-optimal solution can be easily found if we do not connect the farthest counterparts for the graph establishment, verifying the practical design of our graph establishment. Kindly note that the justification of the angle term is not rigorous since the T-SNE dimension reduction cannot fully preserve the angle relationship in the high-dimensional space.

5. Discussion

To clarify the advantage of the proposed high-order framework in open-set learning, we thoroughly compare the methodology design in Table 7. Existing works [18, 24, 19, 59, 14] can be categorized into two types. Firstly, most works [18, 24, 19] discover potential novel objects hidden in the background. OW-DETR [18] leverages the intensity of the feature activation to identify novel objects, capturing and modeling 1^{st} order cues. Moreover, some works [19, 24] use 2^{nd} order evidence, *e.g.*, the distance between a sample and a class center [24] to measure the novel-class uncertainty. Secondly, some works [59, 14] aim to synthesize novel-class samples by randomly sampling base-class image pairs (*i.e.*, the mixup of two base-class images), which also consider 2^{nd} order evidence, *i.e.*, the relation between two samples. Differently, we break through the low-order barrier and use graph motifs, *i.e.*, statistically significant subgraphs, to model the high-order (K) relation among several nodes, which has the following advantages.

Method	Ord.	Methodology
OW-DETR [18]	1	discovering in the background
ORE [24]	2	
OSD [12]	2	
Opendet [19]	2	
PROSER [59]	2	synthesis with two base samples
VOS [14]	2	
Ours	K	synthesis with K base samples selecting via topological evidence discovering in the background

Table 7. Methodology comparison of existing methods. Ord. indicates the order of considered discriminative knowledge.

1) Subgraph achieves a high-order novel-sample synthesis with enriched feature space. Different from [59, 14] using the mixup of two base-class images as synthesized novel samples, we go beyond its 2^{nd} order and achieve high-order synthesis among several samples in subgraphs. Moreover, we use topological evidence to select informative subgraphs instead of randomly selecting image pairs [59, 14]. **2) Subgraph encourages the interaction between real and synthesized novel samples with better discriminability.** Directly optimizing unmatched queries Q_{um} [18] (potential novel-class objects) in the background is sub-optimal since many unseen novel samples are not in Q_{um} [18, 24]. Differently, we incorporate Q_{um} into subgraphs and optimize the real and synthesized novel knowledge collaboratively.

6. Conclusion

This paper formulates and studies a real-world friendly setting, Adaptive Open-set Object Detection (AOOD), and develops three benchmarks, which consider both novel scenes and novel classes in object detection. To address this issue, we propose a Structured Motif Matching (SOMA) framework, which delves into the high-order patterns with graph motifs. SOMA adopts a Structure-aware Novel-class Learning (SNL) for novel-class detection, which conducts a motif-matching on the instance-oriented graph and performs motif-guided novel-class learning. Moreover, it uses Structure-aware Transfer Learning (STL) to adapt to novel scenes, which matches the motif on the semantic-oriented graph to align the per-class distribution with structural awareness. Extensive experiments show that the proposed method outperforms existing approaches significantly.

References

[1] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021. 2, 4

[2] Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016. 2, 3, 4

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 3, 4

[4] Chaoqi Chen, Yushuang Wu, Qiyuan Dai, Hong-Yu Zhou, Mutian Xu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. A survey on graph neural networks and graph transformers in computer vision: A task-oriented perspective. *arXiv preprint arXiv:2209.13232*, 2022. 3

[5] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *CVPR*, pages 8869–8878, 2020. 1, 2

[6] Chaoqi Chen, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. I3net: Implicit instance-invariant network for adapting one-stage object detectors. In *CVPR*, pages 12576–12585, 2021. 6

[7] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, pages 3339–3348, 2018. 1, 2, 6

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 6, 8

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 7

[10] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, pages 4091–4101, June 2021. 2

[11] Jinhong Deng, Xiaoyue Zhang, Wen Li, and Lixin Duan. Cross-domain detection transformer based on spatial-aware and semantic-aware token alignment. *arXiv preprint arXiv:2206.00222*, 2022. 2

[12] Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. The overlooked elephant of object detection: Open set. In *WACV*, pages 1021–1030, 2020. 2, 3, 7, 9

[13] Xuefeng Du, Gabriel Gozum, Yifei Ming, and Yixuan Li. Siren: Shaping representations for detecting out-of-distribution objects. In *Neurips*, 2022. 3

[14] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022. 2, 3, 9

[15] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 6

[16] Dario Fontanel, Matteo Tarantino, Fabio Cermelli, and Barbara Caputo. Detecting the unknown in object detection. *arXiv preprint arXiv:2208.11641*, 2022. 2, 3

[17] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015. 1, 5, 7

[18] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *CVPR*, 2022. 3, 4, 5, 6, 7, 8, 9

[19] Jiaming Han, Yuqiang Ren, Jian Ding, Xingjia Pan, Ke Yan, and Gui-Song Xia. Expanding low-density latent regions for open-set object detection. In *CVPR*, 2022. 1, 2, 3, 5, 6, 7, 9

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 7

[21] Yusuke Hosoya, Masanori Suganuma, and Takayuki Okatani. More practical scenario of open-set object detection: Open at category level and closed at super-category level. *arXiv preprint arXiv:2207.09775*, 2022. 2, 3

[22] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *ECCV*, pages 733–748, 2020. 1, 2

[23] Wei-Jie Huang, Yu-Lin Lu, Shih-Yao Lin, Yusheng Xie, and Yen-Yu Lin. Aqt: Adversarial query transformers for domain adaptive object detection. In *IJCAI*, 2022. 2

[24] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *CVPR*, pages 5830–5840, 2021. 1, 2, 3, 7, 9

[25] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G. Macready. A robust learning approach to domain adaptive object detection. In *ICCV*, October 2019. 2

[26] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *ICCV*, pages 6092–6101, 2019. 2

[27] Wuyang Li, Zhen Chen, Baopu Li, Dingwen Zhang, and Yixuan Yuan. Htd: Heterogeneous task decoupling for two-stage object detection. *TIP*, 30:9456–9469, 2021. 5

[28] Wuyang Li, Jie Liu, Bo Han, and Yixuan Yuan. Adjustment and alignment for unbiased open set domain adaptation. In *CVPR*, pages 24110–24119, June 2023. 4

[29] Wuyang Li, Xinyu Liu, Xiwen Yao, and Yixuan Yuan. Scan: Cross domain object detection with semantic conditioned adaptation. In *AAAI*, volume 6, page 7, 2022. 2

[30] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *CVPR*, pages 5291–5300, June 2022. 2, 4, 5

[31] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. *TPAMI*, pages 1–18, 2023. 2

[32] Xinyu Liu, Wuyang Li, Qiushi Yang, Baopu Li, and Yixuan Yuan. Towards robust adaptive object detection under noisy annotations. In *CVPR*, pages 14207–14216, June 2022. 2

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[34] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *ICRA*, 2018. 3

[35] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. *ICCV*, 2021. 2

[36] Hao Peng, Jianxin Li, Qiran Gong, Senzhang Wang, Yuanxing Ning, and Philip S Yu. Graph convolutional neural networks via motif-based attention. *arXiv preprint arXiv:1811.08270*, 2018. 2, 3

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. 3

[38] Farzaneh Rezaeianaran, Rakshith Shetty, Rahaf Aljundi, Daniel Olmeda Reino, Shanshan Zhang, and Bernt Schiele. Seeking similarities over differences: Similarity-based domain alignment for adaptive object detection. In *ICCV*, pages 9204–9213, 2021. 2

[39] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019. 2, 6

[40] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126(9):973–992, 2018. 6

[41] Wenxu Shi, Lei Zhang, Weijie Chen, and Shiliang Pu. Universal domain adaptive object detector. In *ACM MM*, pages 2258–2266, 2022. 2

[42] Arjun Subramonian. Motif-driven contrastive learning of graph representations. In *AAAI*, volume 35, pages 15980–15981, 2021. 2, 3

[43] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. *arXiv preprint arXiv:2110.06207*, 2021. 6

[44] Vibashan VS, Vikram Gupta, Poojan Oza, Vishwanath A. Sindagi, and Vishal M. Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *CVPR*, pages 4516–4526, June 2021. 2

[45] Wen Wang, Yang Cao, Jing Zhang, Fengxiang He, Zheng-Jun Zha, Yonggang Wen, and Dacheng Tao. Exploring sequence feature alignment for domain adaptive detection transformers. In *ACM MM*, pages 1730–1738, 2021. 2, 6

[46] Wen Wang, Jing Zhang, Yang Cao, Yongliang Shen, and Dacheng Tao. Towards data-efficient detection transformers. In *ECCV*, pages 88–105, 2022. 7

[47] Douglas Brent West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001. 3

[48] Aming Wu, Rui Liu, Yahong Han, Linchao Zhu, and Yi Yang. Vector-decomposed disentanglement for domain-invariant object detection. *ICCV*, 2021. 6

[49] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *CVPR*, pages 11724–11733, 2020. 2

[50] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *CVPR*, pages 12355–12364, 2020. 2

[51] Carl Yang, Mengxiong Liu, Vincent W Zheng, and Jiawei Han. Node, motif and subgraph: Leveraging network functional blocks through structural convolution. In *ASONAM*, pages 47–52. IEEE, 2018. 3

[52] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, June 2020. 6

[53] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, pages 5831–5840, 2018. 3

[54] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 7

[55] Yixin Zhang, Zilei Wang, and Yushi Mao. Rpn prototype alignment for domain adaptive object detector. In *CVPR*, pages 12425–12434, June 2021. 2

[56] Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin. Collaborative training between region proposal localization and classification for domain adaptive object detection. In *ECCV*, pages 86–102. Springer, 2020. 2

[57] Jiyang Zheng, Weihao Li, Jie Hong, Lars Petersson, and Nick Barnes. Towards open-set object detection and discovery. In *CVPR*, pages 3961–3970, 2022. 3

[58] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *CVPR*, pages 13766–13775, 2020. 2, 4, 5

[59] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *CVPR*, 2021. 2, 3, 6, 7, 9

[60] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ICLR*, 2020. 3, 4, 6, 7, 8

[61] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 5