TUPL: Text-guided Unknown Pseudo-Labeling for Open World Object Detection

Xuefei Wang[1], Dong Xu[1(✉)]

[1] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
dxu@shu.edu.cn

Abstract. Open-world object detection (OWOD) aims to improve model performance in practical settings by enabling the detection of previously unseen objects while continuously learning new known classes. This is done in an environment where the set of labeled classes is constantly expanding. Current models often exhibit suboptimal performance when employing objectness scores for pseudo-labeling unknown objects, primarily due to inherent biases toward the known class set. To address this limitation, our study employs a cross-modal learning framework to integrate high-level semantic information from text into the generation of pseudo-labels for objects currently unseen. We introduce a straightforward yet efficient method called **T**ext-guided **U**nknown **P**seudo-**L**abeling (**TUPL**) for open-world object detection. A module named SRS is proposed to direct the model in learning the detection of previously unseen objects. To enhance the model's ability to distinguish foreground elements, we introduce an ROI Feature Refinement module that improves the model's learning of all distinctive foreground characteristics. Experimental evaluations on the PASCAL VOC and MS-COCO benchmarks demonstrate TUPL's exceptional open-world detection capability. Under the OWOD SPLIT setting, TUPL achieves a UR (Unknown Recall) value of 23.1, which is at least double the performance of existing pseudo-labeling methods based on objectness scores.

Keywords: Open world object detection, Cross-modal learning, Pseudo-labeling.

1 Introduction

Deep learning has made significant advancements in the field of object detection. However, conventional approaches to object detection rely on a strict closed-set assumption, where all target classes encountered during testing must have been previously seen in the training phase. Real-world scenarios are characterized by a wide range of diverse objects and the constant emergence of new categories. This leads to considerable complexity in object detection tasks. While traditional methods excel under the closed-set assumption, they often struggle to adapt to the open-world scenario. In recent years, researchers have made substantial progress in developing methods capable of detecting object classes that have not been encountered before.

The first OWOD method ORE, proposed by Joseph [1], addresses the crucial task of enabling the model to recognize object categories that haven't been encountered before as the "unknown class" while still detecting known categories. Numerous recent studies have made notable advancements in this field, categorizable into two main approaches. One approach focuses on learning the feature-level distribution of both known and unknown object categories during training [2 - 5]. The other approach involves generating pseudo-labels for unknown-class objects during the training phase and treating the unknown class as a distinct "known class" for joint learning [1, 6 - 12]. The former typically requires researchers to carefully set a threshold for classifying predictions as the "unknown class". The latter employs pseudo-labeling for unknown-class objects to aid the model's learning during training, often relying on limited statistics from known classes, such as objectness scores, which can introduce bias towards these known classes.

Our main goal is to address the issue of bias towards known classes in current pseudo-labeling methods. Leveraging the advancements in cross-modal tasks, we can seamlessly integrate information from various modalities, such as images, text, audio, and more. Our proposition is that while images offer rich visual details through shapes, textures, and colors, language conveys a higher level of abstract semantic content that can provide more comprehensive guidance.

Building upon the Featurized Query R-CNN (FQR-CNN) framework [13], we have developed a straightforward yet effective unknown pseudo-labeling strategy for OWOD, which we named **TUPL** (**T**ext-guided **U**nknown **P**seudo-**L**abeling for Open World Object Detection). During training, we employ a unique strategy called Sim-Random-Sim (SRS) to label certain candidate boxes as unknown if they do not match any ground truth. Moreover, we utilize a one-to-many matching tactic during training, combined with Non-Maximum Suppression in the inference stage, as an alternative to the initial one-to-one algorithm in FQR-CNN. This approach helps alleviate the problem of insufficient supervision [14, 15]. High-quality feature representation of foreground objects is essential for training object locators and classifiers. To enhance the model's ability to recognize objects from different categories, we propose the ROI Feature Refinement Module (RRM).

Our main contributions are summarized as follows:

- We introduce TUPL, an open-world object detector based on the Featurized Query R-CNN. TUPL combines the strengths of existing OWOD detectors, specifically Faster R-CNN-style and DETR-style models, resulting in a streamlined and efficient integration.
- We propose SRS, a valid unknown pseudo-labeling strategy. By leveraging cross-modality text data and incorporating random debiasing techniques, we effectively mitigate bias issues towards known classes in current methods.
- We introduce RRM, a strategy aimed at enhancing the ROI features that contain foreground objects. By implementing a modified self-attention mechanism on the Region of Interest (ROI) features, we improve the model's accuracy and reliability.
- Our evaluation results on popular OWOD benchmarks, including PASCAL VOC [16] and MS-COCO [17], demonstrate the effectiveness of TUPL in adapting to the

open environment. TUPL maintains high detection performance on known classes while also showcasing strong capabilities in detecting unknown objects.

2 Related work

2.1 Open world object detection

Joseph pioneered the open-world object detection task, which entails the model's ability to not only recognize unknown objects but also progressively acquire the capability to detect new known objects. They proposed ORE [1] as a solution for this challenging task, building upon the Faster R-CNN [18]. A crucial challenge lies in accurately identifying unknown class objects by the model, primarily due to the absence of labels. The use of pseudo-labels has gained significant attention and has become a frequently employed technique [1, 6 - 12]. Several works, including ORE [1], OW-DETR [7], RandBox [9], UC-OWOD [10], and SA [11], utilize objectness scores of candidate proposals to generate pseudo-labels for unknown instances. Moreover, other works such as Open World DETR [6], CAT [8], and RE-OWOD [12] employ fusion of supplementary proposal generation techniques (e.g., selective search [19], etc.) to facilitate the selection of potential candidates for unknown class. The meticulously crafted pseudo-label methods for unknowns have demonstrated significant effectiveness in real-world applications.

In contrast, other methods like OW-RCNN [2], identify both known and unknown classes by establishing multiple thresholds meticulously. 2B-OCD [3] trains an object-centric calibrator, while OCPL [4] utilizes prototype learning to reduce the overlap between the distributions of known and unknown classes within the feature space. PROB [5] employs a category-independent Gaussian distribution to represent object features and calculates the classification outcome of a query by multiplying its probability of being a foreground object with the probability of the foreground object belonging to a specific class. Ann [20] employs a label transfer learning paradigm to distinguish between the features of known and unknown objects. Our primary focus is on OWOD algorithms based on pseudo-labeling, aiming to develop a simple yet robust approach for annotating pseudo-labels.

2.2 Class-Agnostic Object Detection

In open object detection tasks, it is essential for models to learn how to detect "unknown" objects. The task, class-agnostic object detection, focuses on improving the capability of object detection models to detect objects without considering their specific class. The class-agnostic object detection paradigm relies on a finite set of known class training datasets to develop an object detector capable of identifying all foreground objects within an image, irrespective of their class distinctions. WACV [21] emphasizes that in certain real-world situations, accurately determining the presence and exact location of objects is more important than identifying specific categories. It introduces the challenge of class-agnostic object detection. Some methods, such as OLN [22], SIBGRAPI [23], LDET [24], and GOOD [25], enhance the model's detection capabilities for foreground objects at the image level. MAVL [26] observed that previous methods lack supervision from easily understandable semantic signals. Addressing this issue, a novel approach is introduced by utilizing a multi-modality visual transformer trained on aligned image-text data to achieve superior performance in detecting unknown objects. Inspired by this advancement, we propose utilizing semantic information as a guiding signal to help the model detect objects belonging to unknown class during the training process.

2.3 Pre-trained Visual-Language Models

In recent years, pre-trained vision-language models have attracted significant attention in the fields of computer vision and natural language processing. By training on extensive text and image data, these models acquire comprehensive semantic and visual representations. Consequently, they function as robust feature extractors for a wide range of downstream tasks. Typical work, such as CLIP [27], employs contrastive learning on a dataset of 400 million image-text pairs collected from the internet. By aligning the features of images and corresponding text in the feature space, it yields highly robust image and text encoders and showcases remarkable proficiency across diverse downstream tasks, including semantic segmentation [28], object detection [29], image editing [30], image generation [31], and video comprehension [32], among others. Contemporary computer vision research predominantly utilizes CLIP-based methodologies, which rely on text features from the CLIP text encoder as substitutes for classifiers. In contrast, we employ CLIP to align corresponding visual and text features in the feature space. This utilization of CLIP harnesses the embedded text information as high-level semantic guidance, enabling the model to be directed in an unbiased manner. As a result, it facilitates exceptional detection capabilities for all objects in an image, encompassing both known and unknown classes.

3 Proposed methods

3.1 Preliminary

Formulation for OWOD. OWOD consists of a series of subtasks, represented as $T = \{T_1, T_2, \ldots\}$. The corresponding training data can be divided into $D = \{D_1, D_2, \ldots\}$, where the set of categories for all annotated objects in D is $C = \{C_1, C_2, \ldots\}$, and it satisfies $C_i \cap C_j = \phi, i \neq j$. As the training data is sequentially fed, the model learns each subtask step by step. During the T_i phase, the categories of objects included in the training data D_i are denoted as C_i. The categories in C_i are referred to as the currently known classes. The set $C = \{C_1, \ldots, C_{i-1}\}$ is referred to as the previously known classes, while the set $C = \{C_{i+1}, \ldots\}$ is referred to as the currently unknown class, which will be gradually learned in the subsequent incremental learning process. Following the completion of the learning phase for each subtask T_i, we will assess the performance of the current model $Model_i$ on a test dataset encompassing all categories, including both known and unknown ones.

3.2 Overall architecture

Due to the trade-off advantages of FQR-CNN [13] in terms of accuracy and speed by incorporating the query mechanism from DETR into the R-CNN-like detector, TUPL was implemented based on the FQR-CNN framework in this paper. The overall training process aligns with the prevailing paradigm of existing OWOD methods. After completing each subtask, the model will undergo fine-tuning on a small dataset comprising a few samples of previously known classes.

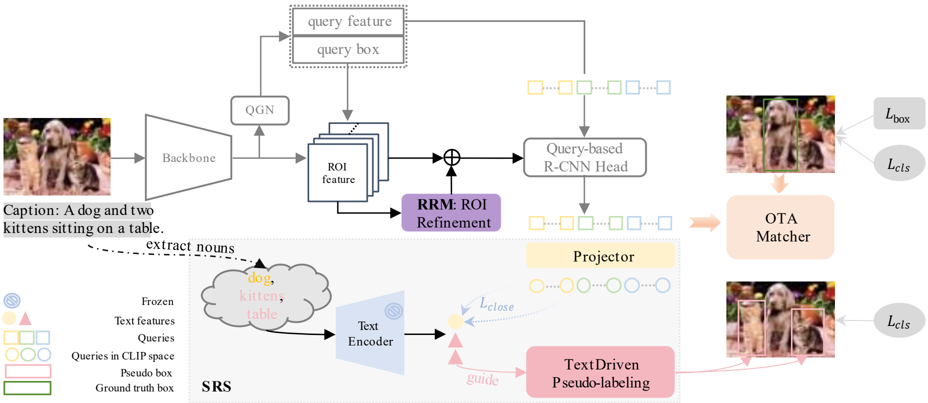

Fig. 1. The architecture of the TUPL framework. TUPL consists of four main steps: (1) extraction of image/text features and generation of queries; (2) interaction between queries and enhanced ROI features; (3) label assignment; and (4) annotation of pseudo-queries for unknown class objects.

For the current subtask learning, **Fig. 1** provides an overview of TUPL. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$ and its corresponding caption Cap, through the backbone, QGN [13] and ROIAlign [33] modules, we can obtain queries $Q = \{q_i\}_1^N$ and ROI features $R = \{r_i\}_1^N$, where N denotes the total number of queries (or ROI features). The ROI features will be enhanced through the RRM module, allowing us to achieve more expressive ROIs R_{aug}. The dynamic interaction between the query features Q and R_{aug} results in the query features denoted as Q_{img}. During the label matching process, a one-to-many label assignment method following OTA [36] is used to provide sufficient supervision information for queries. In this process, queries that match any target are referred to as $Q_{img_m} = \{q_i\}_1^M$, while the others are referred to as $Q_{img_u} = \{q_i\}_1^U$, where $\mathrm{M} + \mathrm{U} = \mathrm{N}$. Moreover, we extract nouns from the caption Cap of I, and use the "a photo of {noun}" template for completion. The templated text is then input into the frozen CLIP text encoder to obtain the corresponding text embedding feature, denoted as E. During this process, the known class embedding features are used to guide the learning of query features with assigned known target labels. The other text embedding features enable the model to identify potential unknown class pseudo-candidates from the remaining set of queries. During the inference stage, the model identifies unknown objects in the same manner as known objects, without requiring any decision threshold boundaries.

Pseudo-labeling for Unknown. Initially, the query features Q_{img} are mapped onto the CLIP text feature space. This mapping aligns the dimensions of the queries with the textual features encoded by CLIP, utilizing the projection operator $Proj$ as:

$$q_{txt_i} = Proj\left(q_{img_i}\right), i \in 1,2, \ldots, N \tag{1}$$

Queries represented in the text feature space is denoted as Q_{txt}. The $Proj$ consists of a compact network made up of linear layers combined with an activation layer. As shown in **Fig. 2**, for queries that match the ground truth of known classes, we minimize the distance between the projected features in the CLIP feature space $Q_{txt_m} = \{q_{txt_i}\}_1^M$, and the textual features of their respective class names $E_m = \{e_i\}_1^M$as:

$$L_{close} = \frac{1}{D}||q_{txt_i} - e_i||_2 + [1 - sim(q_{txt_i}, e_i)], i \in 1,2, \ldots, M \tag{2}$$

D represents the dimension of textual features. L_{close} consists of two parts: the first part reduces the spatial distance between the query features and the features of class names using the L_2 distance, while the second part, represented by $\mathrm{sim}(\cdot)$, increases the similarity between the query features and the corresponding class-specific textual features through cosine similarity distance.

After removing nouns representing known classes from image captions, the remaining nouns may likely describe unknown objects. We refer to them as novel class texts, and their corresponding feature representation is denoted as $E_{novel} = \{e_i\}_1^O$, where O represents the number of novel text embeddings. Our central idea revolves around the premise that if the features of candidate queries Q_{txt_u} exhibit a similarity higher than a certain threshold with any feature representing a novel class text, they may constitute potential candidate boxes for objects of unknown class. Thus, we compute the similarity matrix between Q_{txt_u} and E_{novel} to derive $M \in \mathbb{R}^{U \times O}$. Subsequently, we identify the maximum value along the last dimension of M to yield $Sim \in \mathbb{R}^U$. The method for labeling candidate queries as unknown is determined based on a threshold δ_1:

$$l_{q_i} = \begin{cases} "unknown", & Sim[i] > \delta_1 \\ "bg", & Sim[i] \leq \delta_1 \end{cases}, \ i \in 1,2, \ldots, U \tag{3}$$

and l_{q_i} represents the label for q_i.

SRS: Sim-Rand-Sim. Experimental results (as shown in **Fig. 5**) demonstrate the effectiveness of our ordinary text-guided annotation method for generating pseudo-candidate boxes for unknown class. To reduce the model's bias towards known class features, we propose integrating a random selection approach. This approach aims to decrease the likelihood of mistakenly labeling candidate queries as unknown during the pseudo-annotation process. It specifically targets cases where candidate queries partially contain known class objects. Following the random selection of candidate queries, an additional filtering step is employed. This step applies a higher threshold to enhance the quality of pseudo labels. In summary, the pseudo-labeling procedure combines text guidance with random selection as follows: (1) Utilizing a small threshold δ_1 to filter out most of the background queries, resulting in s_1 candidate queries likely to contain foreground objects; (2) Randomly selecting r candidates from the s_1 queries; (3) Applying a higher threshold δ_2 to further filter low-quality candidates, resulting in s_2 unknown pseudo-queries for annotation.

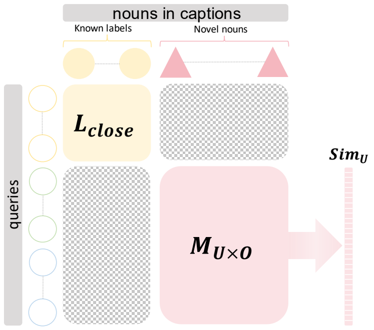

Fig. 2. Pipeline of SRS.

RRM: ROI Refinement by Attention Mechanism. To better capture information pertinent to foreground objects, we introduce the ROI feature refinement module (RRM). Our premise is that the utility of individual ROI features is inherently limited. By comprehensively considering relevant information and enhancing their expressiveness, ROI features can be improved. Therefore, we propose an attention-based approach to enhance ROI features, including those corresponding to foreground objects, by exploiting the similarity between different ROI features. Given the computational expense of directly computing attention maps for ROI features in their original dimension, we employ dimension reduction pooling operations for all ROI features.

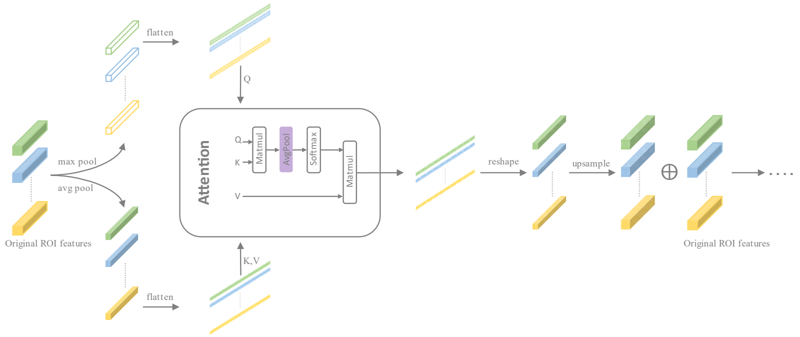

Fig. 3. Structure of ROI Refinement module RRM.

As shown in **Fig. 3**, for ROI features $R \in \mathbb{R}^{N\times dim\times h\times w}$, we obtain $R \in \mathbb{R}^{N\times dim\times\frac{h}{2}\times\frac{w}{2}}$ through max pooling as ROI queries, denoted as Q, and $R \in \mathbb{R}^{N\times dim\times\frac{h}{2}\times\frac{w}{2}}$ through average pooling as ROI keys and ROI values, denoted as K and V respectively. Due to the inevitable introduction of noise when absorbing information from other ROIs, we were inspired by the approach taken by [34], which utilizes the Weights Normalized Convolutional kernel to reduce noise and enhance the attention map. Before applying the softmax operation to the attention map, we utilize average pooling to selectively retain crucial information, effectively filtering out noise. After the modified attention operation, the resulting output will be restored to the original dimension and then combined with the original ROI features using residual connections. The augmented ROIs R_{aug} will then be input into the R-CNN Head along with queries for interaction.

4 Experiments

4.1 Datasets and Metrics

Evaluation experiments were conducted on the widely used OWOD datasets, Pascal VOC [16] and MS-COCO [17]. It was observed that images in the Pascal VOC dataset lack descriptive information, posing a challenge for utilizing text information. To address this limitation, the existing excellent approach, BLIP [35], was employed to automatically generate a caption for each image in Pascal VOC. A comparative analysis between TUPL and existing state-of-the-art methods was conducted using two commonly employed data partitioning criteria. The first criterion, OWOD SPLIT, follows the principles of ORE [1], combining the MS-COCO and Pascal VOC datasets. Within this criterion, all objects belonging to categories in Pascal VOC are considered known classes for Task 1. The sixty additional categories in MS-COCO are divided into three groups, each comprising twenty categories. These groups are assigned as known classes for Task 2, Task 3, and Task 4. The second criterion, proposed by Gupta et al. in Ow-DETR [7], MS-COCO SPLIT, aims to address potential issues of superclass information leakage inherent in the former partitioning pattern. This criterion relies solely on data from the MS-COCO dataset and generates the known class data for each task based on the division of super-classes.

The evaluation criteria align with the mainstream assessment methods employed by existing approaches. The primary metric, mean average precision (mAP), assesses the model's detection performance on known classes, including previously known classes (pre mAP), the known classes specific to the current task (cur mAP), and the cumulative known classes up to the current task (both mAP). To evaluate the model's ability to detect unknown class, the recall rate of unknown class objects (UR) is primarily used. Additionally, two additional metrics are included: the absolute number of unknown class objects mistakenly detected as known class objects (AOSE), and Wilderness Impact (WI), a metric that measures the impact of the model's ability to detect unknown class objects on its detection performance of known class objects. Collectively, these metrics provide a comprehensive assessment of the model's ability to detect both known and unknown class objects in an open environment.

4.2 Implementation details

ResNet50 [37] was employed as the backbone network, initialized with pre-trained weights from ImageNet. Feature Pyramid Network (FPN) was used as the neck, and the cascaded optimization of FQR-CNN for dynamic interaction was employed in the decoding structure. The number of queries N was set to 100 with a dimension of 256. During training, r=5 pseudo candidate boxes for unknown class were randomly selected. The values of δ_1 and δ_2 were empirically set as 0.5 and 0.8, respectively. In the inference stage, the default NMS threshold is set to 0.6.

4.3 Main Results

A comparative analysis was conducted between TUPL and existing open-world object detection algorithms using two different data splitting modes: OWOD SPLIT and MS-COCO SPLIT. To ensure fairness, we exclude the energy-based uncertainty estimation component from ORE (denoted as ORE-EBUI) to eliminate any potential influence of data leakage.

Table 1. Comprehensive evaluation of open-world object detection using OWOD SPLIT dataset.

TaskIDs	Task1				Task2						Task3						Task4		
	WI	AOSE	mAP(↑)	UR	WI	AOSE	mAP(↑)			UR	WI	AOSE	mAP(↑)			UR	mAP(↑)		
	(↓)	(↓)	Cur	(↑)	(↓)	(↓)	Pre	Cur	Both	(↑)	(↓)	(↓)	Pre	Cur	Both	(↑)	Pre	Cur	Both
ORE-EBUI [1]	0.0621	10,459	56.00	4.9	0.0282	10,445	52.70	26.00	39.40	2.9	0.0211	7,990	38.20	12.70	29.70	3.9	29.60	12.40	25.30
OW-DETR [7]	0.0571	10,240	59.2	7.5	0.0278	8,441	53.60	33.50	42.90	6.2	0.0156	6,803	38.30	15.80	30.80	5.7	31.40	17.10	27.80
RandBox [9]	0.0240	4,498	61.80	10.6	0.0078	1,880	-	-	45.30	6.3	0.0054	1,452	-	-	39.40	7.8	-	-	35.40
UC-OWOD [10]	0.0136	9,294	50.66	2.4	0.0117	5,602	33.13	30.54	31.84	3.4	0.0073	3,801	28.80	16.34	24.65	8.7	25.57	15.88	23.14
SA [11]	0.0417	4,889	56.20	-	0.0213	2,546	53.39	26.49	39.94	-	0.0146	2,120	38.04	12.81	29.63	-	30.11	13.31	25.91
CAT [8]	0.0581	7,070	59.90	21.8	0.0263	5,902	54.00	33.60	43.80	18.6	0.0177	5,189	42.10	19.80	34.70	23.9	35.10	17.10	30.60
RE-OWOD [12]	0.0449	-	59.70	9.1	0.0331	-	54.11	37.26	45.64	9.9	0.0241	-	43.06	24.64	37.59	11.4	37.99	28.66	35.66
2B-OCD [3]	0.0481	-	56.37	12.1	0.0160	-	51.57	25.34	38.46	9.4	0.0137	-	37.24	13.23	29.24	11.7	30.06	13.28	25.82
OCPL [4]	0.0423	5,670	56.64	8.3	0.0220	5,690	50.65	27.54	39.10	7.7	0.0162	5,166	38.63	14.74	30.67	11.9	30.75	14.42	26.67
PROB [5]	0.0569	5,195	59.50	19.4	0.0344	6,452	55.70	32.20	44.00	17.4	0.0151	2,641	43.00	22.20	36.00	19.6	35.70	18.90	31.50
Ann-RCNN [20]	0.0604	8,332	56.67	12.8	0.0269	9,454	51.96	29.13	40.55	5.0	0.0157	6,635	40.82	14.56	32.07	9.8	31.68	13.09	27.03
Ann-DETR [20]	0.0564	46,589	59.34	13.6	0.0274	24,709	53.18	37.98	45.58	10.0	0.0194	14,952	43.62	26.66	37.97	14.3	33.54	21.76	30.60
TUPL	0.0620	4,995	60.31	23.1	0.0253	2,598	51.92	34.31	43.12	18.7	0.0157	1,978	41.06	23.10	35.07	22.1	35.06	19.02	31.05

Table 1 compares TUPL with existing OWOD methods using the OWOD SPLIT dataset. The existing methods are classified into three categories based on their use of pseudo-labeling. The first category, methods that annotate unknown class candidates based on the objectness scores, is listed at the top of the table. The second category, methods that annotate unknown class candidates by employing additional techniques, such as the selective search algorithm, is listed in the middle of the table. The third category, methods that do not use pseudo-labeling, is listed at the bottom of the table. Our method either outperforms existing state-of-the-art algorithms or achieves comparable performance across all evaluation metrics. Despite some OWOD methods [5, 7, 8, 20] utilizing advanced and robust object detection frameworks like DDETR [38], TUPL outperforms them in both mAP for known classes and UR for unknown class.

Notably, TUPL achieves a recall rate approximately 2 to 9 times higher than that of existing objectness score-based methods, as shown at the top of the table. Typical methods, such as ORE-EBUI [1] and OW-DETR [7], exhibit a clear bias problem towards known classes, with low unknown recall rates of 4.9 and 7.5, and high AOSE scores. This suggests frequent misclassification of unknown class objects as known class objects. We attribute this issue to the fact that objectness scores' learning exclusively relies on a few labeled known class objects. Consequently, they frequently misclassify boxes containing partial known class objects as unknown. This limitation hampers the ability to differentiate between the two. In contrast, we use high-level textual information as a unbiased guiding factor in learning unknown class objects. Additionally, we incorporate random selection to reduce the chances of mislabeling pseudo-labeled candidate boxes belong to unknown class. These techniques effectively address the known class bias issue observed in existing methods.

Table 2. Comprehensive evaluation of open-world object detection using MS-COCO SPLIT dataset.

TaskIDs	Task1				Task2						Task3						Task4		
	WI	AOSE	mAP(↑)	UR	WI	AOSE	mAP(↑)			UR	WI	AOSE	mAP(↑)			UR	mAP(↑)		
	(↓)	(↓)	Cur	(↑)	(↓)	(↓)	Pre	Cur	Both	(↑)	(↓)	(↓)	Pre	Cur	Both	(↑)	Pre	Cur	Both
ORE-EBUI [1]	-	-	61.40	1.5	-	-	56.50	26.10	40.60	3.9	-	-	38.70	23.70	33.70	3.6	33.60	26.30	31.80
PROB [5]	0.0196	1,915	73.85	17.3	0.0307	3,400	66.15	36.19	50.42	22.1	0.017	1,552	47.72	30.27	41.91	24.5	42.80	31.72	40.03
OW-DETR [7]	0.0458	19,815	71.50	5.7	0.0499	19,749	62.80	27.50	43.80	6.2	0.0248	9,233	45.20	24.90	38.50	6.9	38.20	28.10	33.10
CAT [8]	0.0234	2,126	70.65	24.5	0.0330	4,441	65.83	35.54	50.68	22.2	0.0208	3,545	51.09	32.82	45.00	25.0	45.48	34.90	42.84
TUPL	0.0443	2,490	63.81	29.4	0.0244	1,367	54.23	48.87	51.55	29.0	0.0174	1,405	49.96	43.00	47.64	35.1	47.46	45.02	46.85

Table 2 compares TUPL with existing OWOD algorithms using the MS-COCO SPLIT dataset. The MS-COCO SPLIT aims to address category superclass leakage between different tasks, making model learning more challenging. TUPL achieved significantly greater improvements when faced with the more challenging MS-COCO SPLIT compared to the OWOD SPLIT. TUPL outperformed existing state-of-the-art (SOTA) methods in detecting unknown class, achieving UR rates of 29.4, 29.0, and 35.1 in Tasks 1, 2, and 3, respectively. Evaluation results of Task 1 reveal a significant improvement in the detection accuracy of known classes compared to ORE-EBUI [1]. However, a noticeable disparity in mAP remains in Task 1 compared to the OWOD detectors using DDETR [38]. This disparity is attributed to the underlying framework of object detection. Our approach demonstrated significant improvement as the tasks progressed, reaching optimal levels of accuracy in detecting known classes and recall rates for unknown class in Tasks 2, 3, and 4. Furthermore, it effectively mitigated substantial performance gaps in detecting both previously known and currently known classes, which were common in previous approaches. This further evidences the robust adaptability of TUPL.

4.4 Ablation study

Component of TUPL. To demonstrate the effectiveness of each module in TUPL, ablation experiments are conducted, and the results are presented in Table 3. The baseline in the first row is FQR-CNN without any modifications, indicating that our basic detection model, FQR-CNN, lacks the capability to detect unknown class objects.

Table 3. Ablation experiments of TUPL components.

LineID	baseline	Obj1	Obj5	SRS	OTA	RRM	WI(↓)	AOSE(↓)	mAP(↑)	UR(↑)
1	√						0.0718	79,516	57.06	0.0
2	√	√					0.0761	77,927	56.42	10.0
3	√		√				0.0755	72,358	55.87	13.7
4	√			√			0.0756	15,169	56.33	22.8
5	√			√	√		0.0637	5,143	59.96	22.8
6	√			√	√	√	0.0620	4,995	60.31	23.1

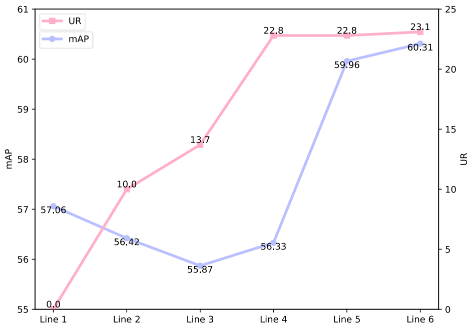

Fig. 4. mAP and UR under different configurations of TUPL components.

"Obj1" represents selecting pseudo-unknown candidates based on the top-1 objectness score, aligning with the approach used in ORE-EBUI. Results in the second row show that "obj1" enables the model to detect unknown class to some extent. However, it falls short in effectively distinguishing between known and unknown classes, as evidenced by the relatively high AOSE score. "Obj5" refers to selecting pseudo-unknown boxes based on the top-5 objectness scores, consistent with the pseudo-labeling approach used in OW-DETR. The increased number of pseudo-labeling candidates enhances the model's capability to recognize unknown class objects to some extent, achieving UR rates from 10.0 to 13.7. However, it also compromises the accuracy of known classes significantly, resulting in a significant decrease in mAP from 0.64 to 1.19.

"SRS" stands for our Text-driven Pseudo-labeling module, which includes random selection. This method significantly improves the recall rate of unknown class objects, surpassing "obj1" by a factor of 2.28. Furthermore, "SRS" not only leads to a moderate decrease in the WI value but also effectively addresses the issue of the model mistakenly identifying unknown class objects as known class objects, resulting in an impressive 80.53% reduction from "obj1" in A-OSE. "OTA" refers to replacing the Hungarian matching algorithm with the OTA matching algorithm. Assigning multiple candidate boxes to each ground-truth during the training process enhances the supervision signal in the model learning process and improves performance in detecting both known and unknown classes. Additionally, integrating the "RRM" module enhances the distinctive features of foreground objects in the images, as illustrated in Line 6, thereby further improving the model's capability to detect all foreground objects. The various components of TUPL are outlined in Line 6. **Fig. 4** provides a visual representation of how the model's performance varies with different combinations of these components.

The Versatility of SRS. We conducted comparative experiments to validate the effectiveness of sub-structures in our pseudo-label method, SRS. As depicted in **Fig. 5**, our method demonstrates exceptional detection performance by selecting pseudo candidates for the unknown class from the top 5 candidates based on similarities, achieving a notable 19.4 unknown recall rate. This performance surpasses the results obtained by existing methods that depend on objectness scores. The inclusion of random selection in our method further enhances UR by an additional 3.5, indicating the random selection further reduces bias towards known classes in our model. Lastly, by applying a higher similarity threshold for secondary filtering, we can obtain higher-quality pseudo labels that are specifically tailored for unknown class. Thus, the model's detection capability is improved for both known and unknown classes simultaneously.

Furthermore, we directly integrate the SRS module into ORE-EBUI, an OWOD method based on the Faster R-CNN framework, to assess its framework-agnostic validity. In the training process, specifically for candidate boxes that do not match any ground truth, we first exclude those with low scores and retain fifty background candidates with the highest objectness score. The primary goal of this step is to reduce noise during the pseudo-labeling process. Subsequently, we apply our SRS without making any modifications. The comparative results for Task 1 are presented in Table 4, where "ORE-EBUI+SRS" denotes our enhancement of ORE through SRS. "ORE-EBUI+obj5" means changing obj1 in the original ORE to obj5. The data reveal that merely augmenting the number of pseudo-labels for unknown class can adversely affect the performance on known classes. The SRS-enhanced model, ORE-EBUI+SRS, achieved a higher unknown class recall rate of 7.6 for unknown class, compared to the original rate of 4.9, without compromising the detection of known classes. Furthermore, the WI metric, which evaluates the model's susceptibility to wild environments, has shown significant improvement. These findings suggest that SRS improves the model's understanding of both known and unknown class objects.

Different ways for ROI Refinement. To assess the effectiveness of our RRM module, we conducted a comparative analysis with "RoIAttn" module proposed by [39], along with two other commonly used feature enhancement methods. The results are summarized in Table 5, where "TUPL-RRM" denotes our model without any ROI enhance ment. LSTM is frequently employed to learn correlations between objects. In this study, we treat all ROI features as a continuous sequence. Subsequently, we average values from the bidirectional LSTM's encoded sequence to obtain the enhanced ROI features, as indicated in the "BiLstm" row. Furthermore, we employ a Graph Convolutional Neural Network (GCN) as another comparative method for our RRM module. In this approach, we consider ROI features as nodes in a graph and establish edges between them based on similarities. To address concerns regarding computational complexity, we utilize the enhanced features obtained after a single update of the graph convolution, as presented in the "GCN" row.

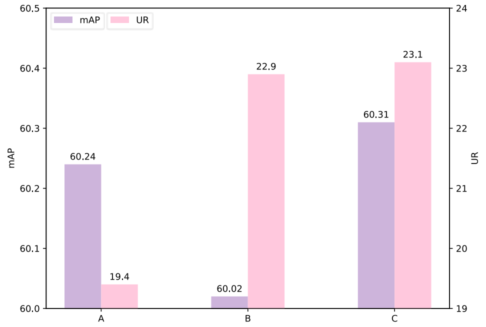

Fig. 5. Ablation of substructures in SRS. 'A': $\delta_1 = 0.5$; 'B': $\delta_1 = 0.5, r = 5$; 'C': $\delta_1 = 0.5, r = 5, \delta_2 = 0.8$.

Table 4. Framework-agnostic validation of SRS on ORE-EBUI.

Strategies	WI(↓)	AOSE(↓)	mAP(↑)	UR(↑)
ORE-EBUI	0.0621	10,459	56.00	4.9
ORE-EBUI +obj5	0.0480	17,345	18.31	7.1
ORE-EBUI +SRS	0.0528	12,120	56.03	7.6

As presented in Table 5, the utilization of the ROI enhancement module has generally yielded a positive impact on the model compared to not using any ROI feature enhancement module. However, RRM consistently exhibits superior performance overall. The BiLstm and GCN approaches result in an increase of 0.26 and 0.1 in mAP, and 0.1 and 0.2 in UR, respectively. In contrast, our RRM module achieve even higher improvements, with a growth of 0.35 and 0.2 in mAP and UR, respectively. It is worth noting that "RoIAttn" has a somewhat inverse effect on the model’s performance when detecting known classes. We attribute this to the nature of "RoIAttn," which involves clustering operations with two additional memory units. In an open environment, labels are unavailable for unknown class, this process is susceptible to a significant amount of unlabeled noise. However, RRM enhances the features of foreground objects based on the similarity of ROI features, thereby circumventing this negative factor.

Hyperparameter Analysis. In our proposed SRS module, there are two critical hyperparameters, denoted as δ_1 and δ_2. We conducted a systematic exploration of these parameters, the outcomes are illustrated in **Fig. 6**. δ_1 is employed to filter out the majority of background queries. Firstly, we conducted an ablation study on δ_1 based on the top-5 similarity scores to selecting pseudo-candidate boxes set. δ_1 is varied in [0.4, 0.5, 0.6, 0.7]. To strike a balance between the model's ability to detect known and unknown class objects, we ultimately settled on $\delta_1 = 0.5$. Building upon this finding, an ablation analysis on δ_2 was further conducted to ascertain its role in the secondary filtering process. Consequently, we determined $\delta_2 = 0.8$ as the optimal value.

Table 5. Comparative experiments of different ROI feature enhancement strategies.

Strategies	WI(↓)	AOSE(↓)	mAP(↑)	UR(↑)
TUPL-RRM	0.0637	5,143	59.96	22.8
BiLstm	0.0637	5,068	60.22	22.9
GCN	0.0634	5,074	60.06	23.0
RoIAttn	0.0633	5,161	59.89	23.2
TUPL	0.0620	4,995	60.31	23.1

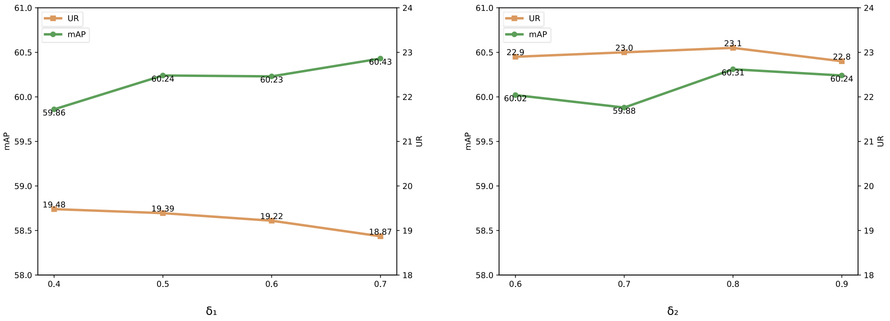

Fig. 6. mAP and UR in Task1 with different hyperparameters.

4.5 Visualization

To more effectively demonstrate the efficacy of TUPL, we present visualizations of some of the test results. As shown in **Fig. 7**, we present the results of our TUPL comparison with ORE-EBUI and OW-DETR after training on task 1. The test results of ORE-EBUI, OW-DETR, and TUPL are displayed in rows 1, 2, and 3, respectively.

Superior Performance. As depicted in **Fig. 7**, TUPL exhibits superior capability in detecting both known and unknown objects in images. Firstly, TUPL shows remarkable capability in accurately localizing and predicting the categories of known objects with high confidence, even in the presence of severe occlusion. For example, TUPL successfully identifies and predicts the correct categories in the top right corner of the images in the third column, where only a chair leg is visible, and in the top left corner of the images in the fourth column, where a small car is heavily occluded. In contrast, both

ORE-EBUI and OW-DETR fail to detect these objects. Secondly, TUPL also demonstrates superior capability in detecting unknown class. For instance, in the first column, ORE-EBUI fails to detect any objects of the unknown class, while OW-DETR incorrectly identifies the background as an unknown object with two nearly identical bounding boxes. In contrast, TUPL explicitly detects unknown objects such as a skateboard, shoes, and a board in the image. Although OW-DETR detects more objects of unknown class in the third column (broccoli, carrot, etc.), it fails to accurately delineate appropriate borders for unknown objects and to detect the known class objects in the picture. Additionally, it incorrectly identifies two objects that do not belong to the known classes. In contrast, TUPL accurately locates objects of unknown class in the image (such as a cup and rice) and precisely identifies all known class objects (dining table and chair).

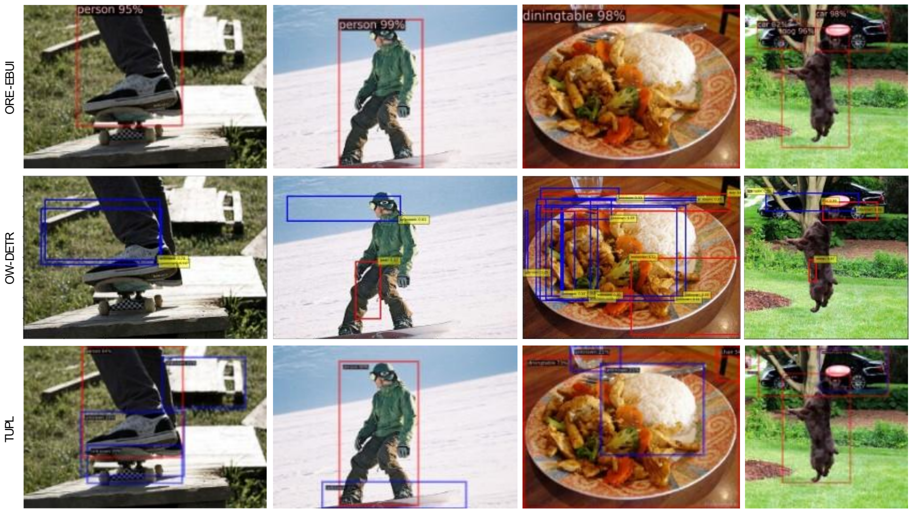

Fig. 7. Visualization of our TUPL comparison with ORE-EBUI and OW-DETR. "unknown" in the figure represents unknown class object in current stage.

Limitations. While our proposed TUPL method exhibits superior object detection capabilities across both known and unknown classes compared to existing approaches relying on objectness scores (as evident in **Fig. 7**), it faces a limitation characterized by relatively low confidence levels when identifying unknown class objects, typically around 20%. This limitation stems from the single pseudo-label mechanism, which assigns a single label to all unknown class objects, potentially compromising the purity of the pseudo-labels. Future research will concentrate on addressing this issue by exploring the potential of providing more specific pseudo-tags for individual unknown class objects, which we believe will enhance the model's performance in this regard.

5 Conclusions

This paper initially analyzes recent research in open-world object detection (OWOD) and identifies a trend wherein models using objectness scores for unknown pseudo-labeling show biases toward known classes. In contrast to images, which mainly convey superficial information, textual data often carries higher-level semantic meaning with practical implications. To overcome this limitation, we integrate text into the pseudo-labeling process for unknown class within the OWOD context, employing a novel text-driven strategy with random debiasing. This approach effectively reduces bias toward known classes and improves the model's performance. Moreover, we improve the model's capability to identify all foreground objects by integrating a one-to-many label matching technique and an ROI feature enhancement strategy. The results of our proposed TUPL, as demonstrated on benchmark datasets, surpass those of existing methods, significantly advancing detection capabilities for unknown class compared to current pseudo-labeling techniques. We anticipate that our research will substantially contribute to the field of object detection in open-world environments.

References

1. Joseph, K., Khan, S., Khan, F.S., Balasubramanian, V.N., 2021. To wards open world object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5830 5840.
2. Pershouse D, Dayoub F, Miller D, et al. Addressing the Challenges of Open-World Object Detection[J]. arXiv preprint arXiv:2303.14930, 2023.
3. Wu Y, Zhao X, Ma Y, et al. Two-branch objectness-centric open world detection[C]//Proceedings of the 3rd International Workshop on Human-Centric Multimedia Analysis. 2022: 35-40.
4. Yu J, Ma L, Li Z, et al. Open-world object detection via discriminative class prototype learning[J]. arXiv preprint arXiv:2302.11757, 2023.
5. Zohar O, Wang K C, Yeung S. Prob: Probabilistic objectness for open world object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 11444-11453.
6. Dong N, Zhang Y, Ding M, et al. Open world detr: Transformer based open world object detection[J]. arXiv preprint arXiv:2212.02969, 2022.
7. Gupta A, Narayan S, Joseph K J, et al. Ow-detr: Open-world detection transformer[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 9235-9244.
8. Ma S, Wang Y, Wei Y, et al. Cat: Localization and identification cascade detection transformer for open-world object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 19681-19690.
9. Wang Y, Yue Z, Hua X S, et al. Random boxes are open-world object detectors[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 6233-6243.
10. Wu Z, Lu Y, Chen X, et al. Uc-owod: Unknown-classified open world object detection[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 193-210.

11. Yang, S., Sun, P., Jiang, Y., Xia, X., Zhang, R., Yuan, Z., Wang, C., Luo, P., Xu, M., 2022. Objects in semantic topology, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=d5SCUJ5t1k.
12. Zhao X, Ma Y, Wang D, et al. Revisiting open world object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023.
13. Zhang, W., Cheng, T., Wang, X., Chen, S., Zhang, Q., Liu, W., 2022. Featurized query r-cnn. arXiv preprint arXiv:2206.06258.
14. Chen Q, Chen X, Wang J, et al. Group detr: Fast detr training with group-wise one-to-many assignment[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 6633-6642.
15. Jia D, Yuan Y, He H, et al. Detrs with hybrid matching[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 19702-19712.
16. Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge[J]. International journal of computer vision, 2010, 88: 303-338.
17. Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.
18. Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1137-1149.
19. Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition[J]. International journal of computer vision, 2013, 104: 154-171.
20. Ma Y, Li H, Zhang Z, et al. Annealing-based label-transfer learning for open world object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 11454-11463.
21. Jaiswal A, Wu Y, Natarajan P, et al. Class-agnostic object detection[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 919-928.
22. Kim D, Lin T Y, Angelova A, et al. Learning open-world object proposals without learning to classify[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 5453-5460.
23. Gonçalves G R, Sena J, Schwartz W R, et al. Pixel-level Class-Agnostic Object Detection using Texture Quantization[C]//2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). IEEE, 2022, 1: 31-36.
24. Saito K, Hu P, Darrell T, et al. Learning to detect every thing in an open world[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 268-284.
25. Huang H, Geiger A, Zhang D. Good: Exploring geometric cues for detecting objects in an open world[J]. arXiv preprint arXiv:2212.11720, 2022.
26. Maaz M, Rasheed H, Khan S, et al. Class-agnostic object detection with multi-modal transformer[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 512-531.
27. Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR, 2021: 8748-8763.
28. Rao Y, Zhao W, Chen G, et al. Denseclip: Language-guided dense prediction with context-aware prompting[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 18082-18091.
29. Zhao S, Zhang Z, Schulter S, et al. Exploiting unlabeled data with vision and language models for object detection[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 159-175.

30. Wei T, Chen D, Zhou W, et al. Hairclip: Design your hair by text and reference image[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 18072-18081.
31. Patashnik O, Wu Z, Shechtman E, et al. Styleclip: Text-driven manipulation of stylegan imagery[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 2085-2094.
32. Xu H, Ghosh G, Huang P Y, et al. Videoclip: Contrastive pre-training for zero-shot video-text understanding[J]. arXiv preprint arXiv:2109.14084, 2021.
33. He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.
34. Kim J, Choi J, Choi H J, et al. Shepherding slots to objects: Towards stable and robust object-centric learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 19198-19207.
35. Li J, Li D, Xiong C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//International conference on machine learning. PMLR, 2022: 12888-12900.
36. Ge Z, Liu S, Li Z, et al. Ota: Optimal transport assignment for object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 303-312.
37. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
38. Zhu X, Su W, Lu L, et al. Deformable detr: Deformable transformers for end-to-end object detection[J]. arXiv preprint arXiv:2010.04159, 2020.
39. Liang X, Song P. Excavating roi attention for underwater object detection[C]//2022 IEEE International Conference on Image Processing (ICIP). IEEE, 2022: 2651-2655.