electronics

MDPI

Article

Text-Guided Unknown Pseudo-Labeling for Open-World Object Detection

Xuefei Wang and Dong Xu *

School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China; xuefei_wang@shu.edu.cn
* Correspondence: dxu@shu.edu.cn

Abstract: Open-world object detection (OWOD) focuses on training models with partially known class labels, enabling the detection of objects from known classes while concurrently identifying objects from unknown classes. Current models often perform suboptimally in generating pseudo-labels for unknown objects based on objectness scores due to inherent biases towards known classes. To address this issue, we propose a cross-modal learning model named **T**ext-**G**uided Unknown Pseudo-Labeling for **O**pen-world **O**bject **D**etection(**TGOOD**) building on the Featurized Query R-CNN (FQR-CNN) framework. Specifically, we introduce a module called Similarity-Random-Similarity (**SRS**) to guide the model in detecting unknown objects during training. Additionally, we replace the one-to-one label assignment strategy in FQR-CNN with a one-to-many (**OTM**) label assignment strategy to provide more supervisory information during training. Moreover, we propose the ROI features Refinement Module (**RRM**) to enhance the discriminability of all objects. Experimental evaluations on the PASCAL VOC, MS-COCO, and COCO-O benchmarks demonstrate TGOOD's superior open-world detection capability.

Keywords: open-world object detection; cross-modal learning; pseudo-labeling

1. Introduction

Deep learning has made significant advancements in object detection. However, conventional object detection methods operate under a closed-set assumption, where all target classes seen during testing must also be present during training. In contrast, real-world scenarios involve a diverse array of objects and the continuous emergence of new categories, complicating object detection tasks. While effective in closed-set frameworks, traditional methods often struggle to handle open-world scenarios. Recently, researchers have made substantial progress in developing methods capable of detecting unseen object classes.

The first open-world object detection (OWOD) method, ORE, introduced by Joseph [1], addresses the critical challenge of enabling models to recognize previously unseen object categories as the "unknown" class while continuing to detect known categories. Recent advancements in OWOD can be classified into two main approaches. The first focuses on learning feature-level distributions for both known and unknown object categories during training [2–5]. The second 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 often requires defining a threshold to classify predictions as the unknown class, while the latter relies on pseudo-labeling unknown class objects to guide model learning during training. However, this method often depends on limited statistics from known classes, such as objectness scores, which can introduce bias. Specifically, as illustrated in Figure 1a, during pseudo-labeling, proposals that include parts of known-class objects are prone to being mislabeled as the unknown class. This mislabeling undermines the model's ability to differentiate between known and unknown classes, thereby compromising overall detection performance.

check for updates

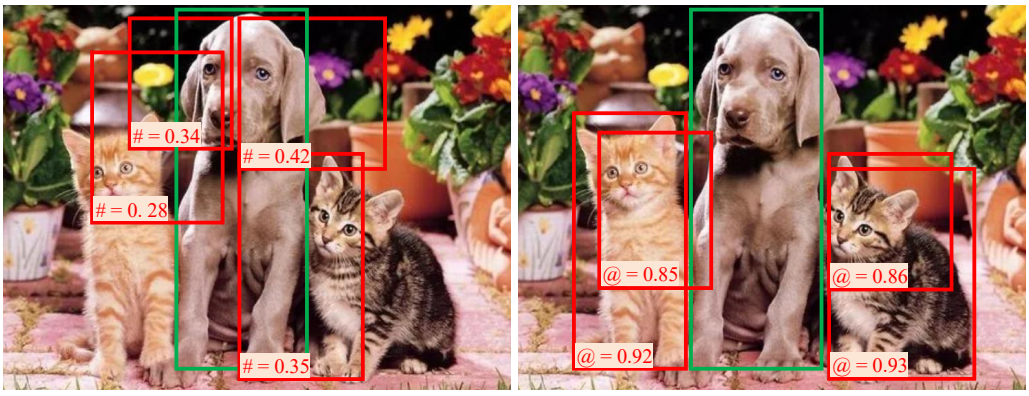
Citation: Wang, X.; Xu, D. Text-Guided Unknown Pseudo-Labeling for Open-World Object Detection. *Electronics* **2024**, *13*, 4528. https://doi.org/10.3390/electronics13224528

Received: 26 October 2024
Revised: 16 November 2024
Accepted: 16 November 2024
Published: 18 November 2024

Our primary objective is to reduce the bias towards known classes in current pseudo-labeling methods. Leveraging advancements in cross-modal tasks, we can seamlessly integrate information from diverse modalities, including images, text, audio, and other forms of data. We propose that while images provide detailed visual information through shapes, textures, and colors, language provides abstract semantic information, offering more comprehensive guidance. Building on the FQR-CNN framework [13], we have developed a simple yet effective pseudo-labeling strategy for OWOD, termed TGOOD (**T**ext-**G**uided Unknown Pseudo-Labeling for **O**pen-World **O**bject **D**etection). During training, we apply a novel module called Similarity-Random-Similarity (SRS) to label candidate boxes as unknown when they do not match any ground truth (gt). Additionally, we adopt a one-to-many (OTM) matching strategy during training, combined with Non-Maximum Suppression (NMS) during inference, as a replacement for the original one-to-one algorithm used in FQR-CNN. This modification resolves the problem of insufficient supervision [14,15]. High-quality feature representation of foreground objects is critical for the training of object locators and classifiers. To enhance the model's ability to recognize objects across all categories, we introduce the ROI features Refinement Module (RRM).

Our main contributions are summarized as follows:

- **TGOOD:** We propose an OWOD detector, TGOOD, built upon the FQR-CNN framework [13], which incorporates the SRS module, OTM technique, and RRM module. TGOOD leverages the strengths of both Faster R-CNN-style and DETR-style OWOD detectors, leading to a streamlined and efficient detection approach.
- **Benchmark Performance:** Extensive evaluations on OWOD benchmarks, including PASCAL VOC [16] and MS-COCO [17], demonstrate TGOOD's ability to effectively adapt to open-world environments. It maintains high performance on known classes while exhibiting strong capabilities in detecting unknown objects.
- **Cross-Domain Generalization:** Comparative analysis on COCO-O [18], which includes images from diverse real-world domains, reveals that TGOOD not only excels in generalizing across different classes but also performs exceptionally well across diverse domains, underscoring its robustness and superiority.



(a) (b)

Figure 1. Comparison of pseudo-labeling methods using the objectness score (**a**) and TGOOD (**b**). Green boxes represent ground-truth objects of known classes, while red boxes denote candidate boxes that the model identifies as potentially containing an unknown object. The symbol # indicates the objectness score, and @ denotes the similarity score of ROI features with the embedding of the word "cat". The method using the objectness score tends to misclassify boxes containing fractional-known-class objects as unknown. In contrast, TGOOD provides more accurate labels of unknown-class objects guided by text.

2. Related Work

2.1. Open-World Object Detection

Joseph pioneered the task of open-world object detection, which involves a model's ability to both recognize previously unknown objects and progressively acquire the capability to detect new known objects. They introduced ORE [1] as a solution to this complex challenge, building upon Faster R-CNN [19]. A key difficulty in this task is the accurate identification of unknown-class objects due to the lack of labels. The use of pseudo-labels has gained substantial attention and is now a commonly employed technique [1,6–10,12]. Several studies, including ORE [1], OW-DETR [7], RandBox [9], UC-OWOD [10], and SA [11], leverage objectness scores of candidate proposals to generate pseudo-labels for unknown instances. Additionally, other works, such as Open World DETR [12], CAT [8], and RE-OWOD [12], utilize supplementary proposal generation techniques (e.g., selective search [20]) to improve the selection of potential candidates for unknown classes. These well-developed pseudo-label methods for unknowns have demonstrated considerable effectiveness in real world applications. In contrast, there are other methods that do not use pseudo-labels. Like OW-RCNN [2], 2B-OCD [3], OCPL [4], PROB [5] and Ann [21], they try to distinguish between known- and unknown-class objects in the feature space. Our primary focus is on OWOD algorithms that leverage pseudo-labeling, with the goal of developing a simple yet robust approach for generating pseudo-labels.

2.2. Class-Agnostic Object Detection

In open object detection tasks, it is crucial for models to learn to detect unknown objects. The class-agnostic object detection task aims to enhance the capability of object detection models to identify objects without considering their specific class. This paradigm depends on a finite set of known-class training datasets to develop a detector that can recognize all foreground objects within an image, regardless of class distinctions. WACV [22] highlights that in certain real-world scenarios, accurately determining the presence and precise location of objects is more critical than classifying them into specific categories, thus introducing the challenge of class-agnostic object detection. Methods such as OLN [23], SIBGRAPI [24], LDET [25], and GOOD [26] focus on improving the model's detection capabilities at the image level. MAVL [27] noted that prior methods often lack supervision from easily interpretable semantic signals. To address this, they utilized a multi-modality visual transformer trained on aligned image–text data, leading to improved performance in detecting unknown objects. Being inspired by this, we propose to use semantic information as a guiding signal to assist the model in detecting objects of unknown classes during training.

2.3. Pre-Trained Visual–Language Models

In recent years, pre-trained vision–language models have garnered significant attention in both computer vision and natural language processing. These models, trained on extensive text and image datasets, acquire comprehensive semantic and visual representations, serving as powerful feature extractors for a wide range of downstream tasks. For example, CLIP [28] utilizes contrastive learning on a dataset of 400 million image/text pairs sourced from the internet. By aligning the features of images and their corresponding text in the feature space, CLIP produces highly robust image and text encoders, demonstrating exceptional performance across various downstream tasks, such as semantic segmentation [29], object detection [30], image editing [31], image generation [32], and video comprehension [33]. Contemporary computer vision research primarily employs CLIP-based methodologies, where text features from the CLIP text encoder are used as substitutes for traditional classifiers. However, in our approach, we leverage CLIP to align visual and text features within the feature space. We use embedded text information as high-level semantic guidance, allowing the model to be directed without bias. Consequently, we enhance the model's detection capabilities across all objects in an image, including both known and novel classes.

3. Method

3.1. Preliminary: Formulation for OWOD

OWOD consists of a series of subtasks, represented as $T = \{T_1, T_2, T_3, \ldots\}$. The corresponding training data can be divided into $D = \{D_1, D_2, D_3 \ldots\}$. The set of categories for all annotated objects in D is $C = \{C_1, C_2, C_3 \ldots\}$, where C_i contains all annotated categories in D_i, and it satisfies $C_i \cap C_j = \emptyset, i \neq j$. To maintain consistency with the existing literature, in this study, the sizes of sets T, D, and C are all set to 4 to simulate a realistic open environment. As the training data is sequentially fed, the model learns each subtask step by step. Assuming the initial model is represented as $Model_0$, after completing the training for task T_1 on D_1, the model will be updated to $Model_1$, and so on, following this pattern. During the T_i phase, the categories in C_i are referred to as the currently known classes, which are the classes included in the annotated data during model training. 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 classes, which will be gradually learned in the subsequent incremental learning process. Assume there are KN annotated objects in an image from D_i, with each object's label containing category information and bounding box coordinates, denoted as $\{Y_1, Y_2, \ldots, Y_{KN}\}$, where $Y_{kn} = [c_{kn}, x_{kn1}, y_{kn1}, x_{kn2}, y_{kn2}]$ and $kn \in \{1, 2, \ldots, KN\}$. Here, $c_{kn} \in C_i$ represents the category to which the object belongs, and $x_{kn1}, y_{kn1}, x_{kn2}, y_{kn2}$ indicate the positional coordinates of the object in the image. After completing the learning phase of each subtask T_i, we evaluate the current model $Model_i$ on a test dataset that contains all categories (both known and unknown). We are constantly working to enhance the model's ability to generalize, allowing it to accurately detect objects from unknown categories while maintaining exceptional detection performance for known categories in diverse and dynamic real-world detection environments.

3.2. Overall Architecture

TGOOD is implemented using the FQR-CNN framework due to its advantageous trade-off between accuracy and speed, achieved by incorporating the query mechanism from DETR [34] into the R-CNN-like detector. The overall training process follows the standard paradigm of existing OWOD methods like [1]. After each subtask, the model is fine-tuned on a small dataset containing a few samples from previously known classes.

As illustrated in Figure 2, the main training process of TGOOD is divided into the following five steps: (1) Feature extraction: for a given image and its description, the backbone network, QGN [13], and ROI Align [35] modules are used to extract queries (Q) and region of interest (ROI) features from the image. Nouns are extracted from the description, and the text embeddings (E) are obtained by the CLIP text encoder. (2) Feature interaction: the ROI features from step (1) are firstly enhanced through the RRM module, allowing the query features to better capture the discriminative information of the foreground objects. By the Query-based RCNN Head [13], we successfully derive the query features (Q_{img}), which have undergone interaction with the augmented ROI features (R_{aug}). (3) One-to-many label assignment: in this phase, at least one candidate query is assigned to each gt, ensuring sufficient supervision signals. (4) Text-guided pseudo-labeling: after label assignment for known-class objects, candidate queries with high similarity to novel text embeddings are selected as pseudo-labels for unknown objects in an unbiased manner. (5) Loss calculation: the corresponding box loss and classification loss are computed for known classes, while only the classification loss is calculated for unknown classes.

During the inference stage, text input is not required, nor is it necessary to set separate threshold boundaries for detecting unknown objects. The model treats unknown-class objects as a distinct category, labeled "unknown", and detects them in the same manner as known-class objects.

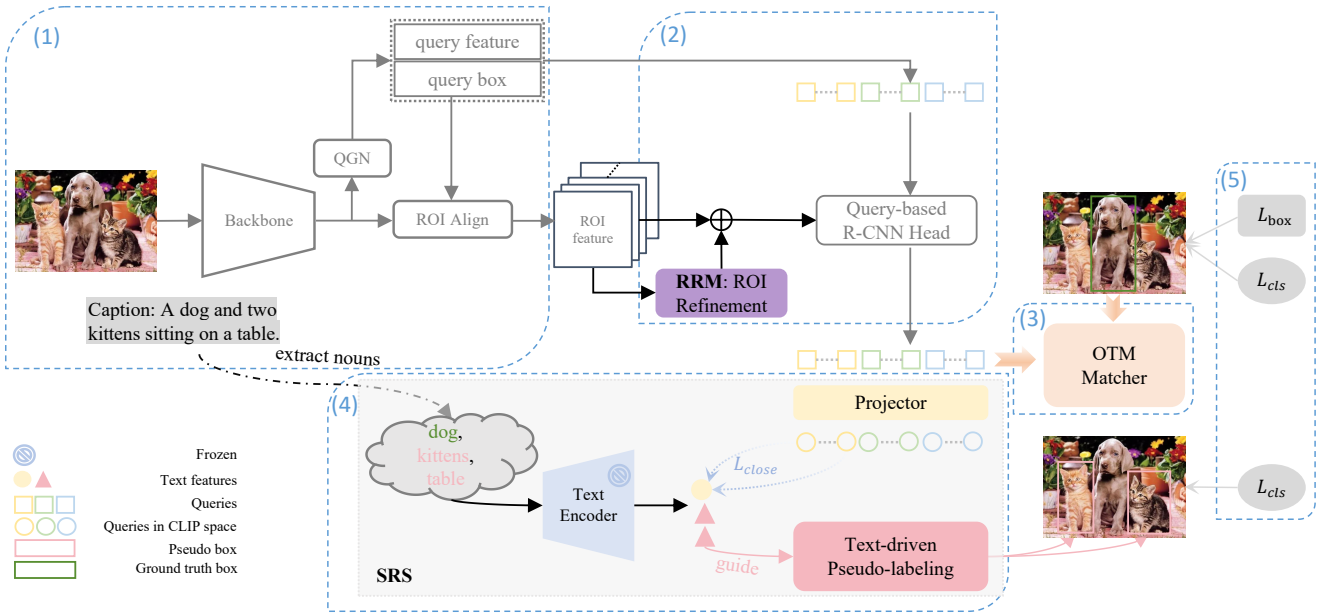


Figure 2. The architecture of our proposed TGOOD. The symbol ⊕ represents an addition operation. TGOOD consists of five main steps: (1) image/text feature extraction and queries generation; (2) the interaction between queries and enhanced ROI features; (3) one-to-many label assignment; (4) text-guided pseudo-labeling; (5) loss calculation.

3.2.1. SRS: Text-Guided Pseudo-Label Generation Strategy

As shown in Figure 2, part (4), following the label assignment process in part (3), the SRS module is applied to obtain candidate queries for unknown classes from those that do not match any known class gt. Figure 3 illustrates the flowchart of the SRS module. Two pseudo-label generation strategies are explored: a standard version and an enhanced version.

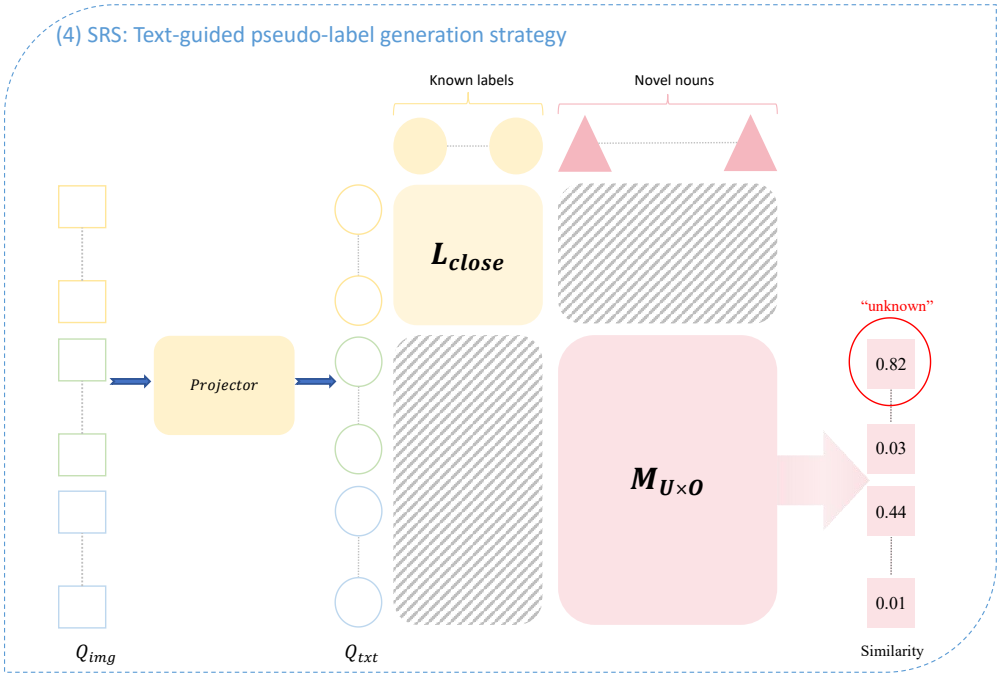


Figure 3. Diagram of text-guided pseudo-label generation strategy.

Firstly, the query features $Q_{img} = \{q_{img_1}, q_{img_2}, \ldots, q_{img_N}\}$, with N representing the number of queries, are mapped to the CLIP text feature space for dimensional alignment. A feature projector $Proj : \mathbb{R}^{d_{img}} \rightarrow \mathbb{R}^{d_{txt}}$, consisting of linear layers and an activation layer, is designed to perform this mapping, as shown in Equation (1). The dimensions of the query features before and after mapping are denoted by d_{img} and d_{txt}, respectively.

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

Then, the queries projected into the CLIP text feature space are divided into two groups: (1) the queries matching the known-class targets (the yellow hollow circles in Figure 3), denoted as Q^m_{txt}, with a quantity of M; (2) the remaining queries, denoted as Q^u_{txt}, with a quantity of U, where $M + U = N$. For Q^m_{txt}, a loss function combining cosine similarity and Euclidean distance is applied to minimize the distance between the query and the corresponding text embedding in the feature space. As shown in Equation (2), e_i represents the text embedding feature corresponding to the known class. The first term of L_{close} represents the Euclidean distance, while the second term represents the cosine distance.

$$L_{close} = \sum_{i=1}^{M} \{ \frac{1}{d_{txt}} \| q_{txt_i} - e_i \|_2 + [1 - sim(q_{txt_i}, e_i)] \} \tag{2}$$

For Q^u_{txt}, we use it to generate pseudo candidates for unknown classes. After removing information related to known-class objects from the image caption, the remaining nouns are likely to represent unknown-class objects. These nouns are referred to as novel texts, and the corresponding text embedding features are expressed as $E_{novel} = \{e_1, e_2, \ldots, e_O\}$, where O represents the number of novel texts. It is hypothesized that if the features of candidate queries show similarity above a certain threshold to any novel text feature, they may correspond to potential candidate boxes for unknown objects. Therefore, the similarity matrix between Q^u_{txt} and E_{novel} is computed, resulting in a matrix $Matrix \in \mathbb{R}^{U \times O}$. The maximum value along the last dimension of the matrix is then used to represent the similarity score $Similarity \in \mathbb{R}^U$.

1. The standard version:
   In the standard version, the pseudo-labeling process is conducted according to the rule outlined in Equation (3), where bg means background, and l_{q_i} denotes the label of the i-th candidate query, which will be regarded as an unknown class if its similarity score exceeds the threshold δ_1 and ranks highly among all similarity scores. $Similarity_{topK} \in \mathbb{R}^K$ contains the topK scores from $Similarity \in \mathbb{R}^U$. Empirically, the value of K was set to 5.
   $$l_{q_i} = \begin{cases} unknown, & Similarity_{topK}[i] > \delta_1 \\ bg, & Similarity_{topK}[i] \leq \delta_1 \end{cases}, i \in 1, 2, \ldots, K \tag{3}$$
2. The enhanced version:
   Since the text encoder used in this study is trained on 400 million image/text pairs, while the dataset employed in our experiments consists of only 80 common categories, the CLIP text features and unknown text features may exhibit some similarity. This could introduce a degree of interference in the labeling process, potentially leading to the generation of false labels. To mitigate this, we propose an enhanced pseudo-labeling strategy combined with random de-biasing. The overall framework of this strategy aligns with the standard version, but the decision-making process based on the similarity score is divided into three steps: (1) initially applying a small threshold δ_1 to filter out most candidate queries containing bg class; (2) randomly selecting r candidate queries from the remaining pool; and (3), finally, using a larger threshold δ_2 to filter out low-quality candidate queries and generate the final unknown-class pseudo-labels.

3.2.2. OTM: One-to-Many Label Assignment

The FQR-CNN framework we utilize adopts a decoding structure similar to DETR. It employs the Hungarian algorithm [36] for label matching, which follows a one-to-one matching mode. While this retains the end-to-end detection advantages by eliminating post-processing operations like NMS, it often lacks sufficient supervisory information. To address this issue, we implement a one-to-many matching strategy using the optimal transport algorithm [37]. During inference, NMS is applied as a post-processing step. Since the number of candidate boxes in the FQR-CNN framework is considerably smaller than that in traditional R-CNN-based frameworks, the computational cost of applying NMS remains minimal.

Following OTA [38], we employ the concept of optimal transport to achieve one-to-many label assignment. For each image, let S represents the labels (both known and bg classes) as suppliers, where each supplier s_i can provide k_i labels. The N candidate query boxes act as demanders, with each requiring a label. The value of k_i for the known class is dynamically determined based on the overlap between the target box and all candidate query boxes, as in OTA [38]. The total number of labels provided by all known classes is $\sum k_i$, while the bg class provides $N - \sum k_i$ labels. The element c_{ij} in the cost matrix represents the cost of assigning one label (one of the k_i labels) from supplier s_i to the j-th query, which is calculated as shown in Equation (4), where gt represents the target label and α is a moderator. If it is a known class, c_{ij} is composed of the classifier's predicted loss and the locator's predicted loss; for the bg class, c_{ij} only includes the classifier's predicted loss.

$$c_{ij} = \begin{cases} L_{cls}(p_j, gt_i) + \alpha \cdot L_{box}(p_j, gt_i), & if\ gt\ is\ known\ class \\ L_{cls}(p_j, gt_i), & if\ gt\ is\ bg \end{cases} \tag{4}$$

The goal of the optimal transport algorithm is to find an optimal assignment plan $\pi^* \in \mathbb{R}^{S \times N}$ that minimizes the overall matching cost. The symbolic expression is provided in Equation (5).

$$\begin{aligned} &\min_{\pi} \sum_{i=1}^{S} \sum_{j=1}^{N} c_{ij} \cdot \pi_{ij}, \\ s.t. &\sum_{i=1}^{S} \pi_{ij} = d_j, \\ &\sum_{j=1}^{N} \pi_{ij} = s_i, \\ &\sum_{i=1}^{S} s_i = \sum_{j=1}^{N} d_{ij}, \\ &\pi_{ij} \geq 0 \end{aligned} \tag{5}$$

The classical solution for finding π^* involves multiple iterations using the Sinkhorn–Knopp algorithm [39]. However, to enhance computational efficiency, we did not employ this iterative algorithm to find the optimal π^*. Instead, after determining k_i and c_{ij}, the labels of the known class are assigned to the k_i candidate queries with the lowest matching cost. If a candidate query is matched with multiple gt labels, it is assigned to the gt with the lower matching cost.

Suppose an image contains two known target objects, which can provide $k_1 = 2$ and $k_2 = 1$ labels, respectively, and the number of candidate queries is 5. This means the bg class provides $5 - 2 - 1 = 2$ labels. The pairs formed by the one-to-one matching method and the one-to-many matching method are illustrated in Figure 4. The one-to-one matching algorithm only assigns Q_1 to GT_1. In the one-to-many matching mode, 2, 1, and 2 candidate queries are assigned to GT_1, GT_2, and the bg class, respectively. For GT_1, both Q_1 and Q_4 are

candidate queries with a high matching degree, thereby providing a more comprehensive supervisory signal.

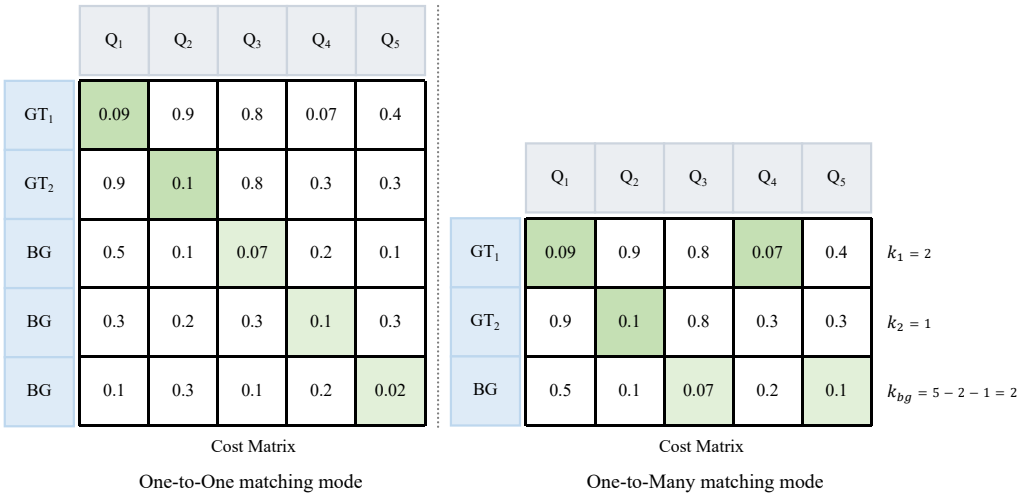


Figure 4. One−to−one vs. one−to−many matching mechanism. GT: ground truth, BG: background, Q: query.

3.2.3. RRM: ROI Features Refinement Module

To enable the query to better capture information related to foreground objects, allowing the classifier and locator to make more accurate decisions, we introduce an ROI feature refinement module based on the attention mechanism. The rationale is that a single ROI feature conveys limited information. By taking a more holistic view and incorporating relevant nearby information, the representational capacity of the ROI features can be enhanced. Specifically, an improved self-attention mechanism is applied to the ROI features before their interaction with the queries. While the attention mechanism effectively enhances the relevant information near the ROI features, directly applying attention to these high-dimensional features can result in excessive computational complexity. To address this, we employ a pooling operation to retain key information while reducing the dimensionality of the ROI features.

Figure 5 illustrates the overall process of foreground feature enhancement using an attention mechanism. As shown in Equation (6), the original ROI features, ROI_{ori}, are processed through max pooling and flattened into a one-dimensional vector, which serves as the input Q for the attention module. Meanwhile, ROI_{ori} are average pooled and flattened into a one-dimensional vector, and used as the inputs K and V for the attention module. Due to the inevitable introduction of noise when absorbing information from others, Ref. [40] used the Weights Normalized Convolutional kernel to reduce noise around the object. Inspired by this, we propose applying an average pooling operation to the attention matrix before performing the softmax operation (Equation (6)) to retain only the most salient information, thus minimizing noise. After passing through the attention module, the results are restored to the original dimensionality of ROI_{ori} using rearrangement and upsampling operations. Finally, they are subsequently fused with ROI_{ori} via residual connections to get the refinement ROI features ROI_{aug}.

$$\begin{aligned} Q &= Flatten(MaxPool(ROI_{ori})), \\ K, V &= Flatten(AvgPool(ROI_{ori})), \\ ROI_{aug} &= ROI_{ori} + Upsample(Reshape(Attention(Q, K, V))) \end{aligned} \tag{6}$$

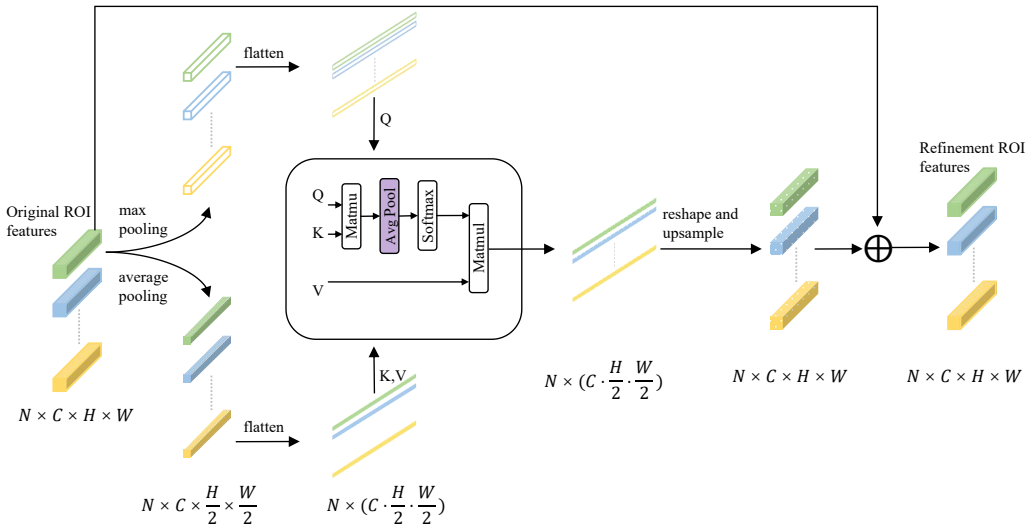


Figure 5. Feature enhancement module based on attention mechanism.

4. Experiments and Results

4.1. Datasets and Metrics

4.1.1. Datasets

The datasets used in this study include MS-COCO [17], Pascal VOC [16], and COCO-O [18]. The current mainstream OWOD dataset configurations include OWOD SPLIT, proposed in ORE [1], and MS-COCO SPLIT, proposed in OW-DETR [7]. These two configurations address open-world object detection with category increments but overlook the fact that many real-world scenarios involve significant distribution shifts—referred to as domain increments. To account for this, we propose a third dataset configuration, COCO-O SPLIT, which evaluates the model's domain generalization capability under these conditions. The following sections introduce these three dataset configurations in detail:

OWOD SPLIT: In this setup, the MS-COCO [17] and Pascal VOC [16] datasets were first combined, with all objects belonging to categories in Pascal VOC designated as known categories for task 1. The remaining 60 categories in MS-COCO [17] were then divided into three groups, with each group containing 20 categories of objects, which were designated as known categories for tasks 2, 3, and 4. Since the images in Pascal VOC do not include description information, meanwhile, BLIP [41], a large model pre-trained on a vast amount of data, demonstrates strong performance across multiple visual–language tasks. So, we used BLIP to randomly generate a caption for each image in Pascal VOC.

MS-COCO SPLIT: Superclass data leakage occurs in OWOD SPLIT, with task 1 predominantly featuring classes from the vehicle and animal categories, while task 2 introduces related subclasses like truck, elephant, bear, zebra, and giraffe. The MS-COCO SPLIT partitioning method was proposed by Gupta et al. [7] to address the issue of superclass information leakage that may occur in the OWOD SPLIT configuration. It uses only the MS-COCO [17] dataset and organizes the known categories for each task based on a superclass partitioning strategy, with nearly 20 classes in each task.

COCO-O SPLIT: COCO-O SPLIT utilizes distribution shift data from COCO-O [18] to assess the model's performance in diverse domains. COCO-O, proposed by [18], is a test dataset derived from the MS-COCO validation set and includes six natural distribution shifts: weather, painting, handmade, cartoon, tattoo, and sketch. These six types of data are regarded as unknown domain data derived from the original images in MS-COCO.

4.1.2. Metrics

The evaluation metrics we used align with the mainstream evaluation criteria of existing methods [1]. For the detection performance of known classes, mean Average Precision (mAP) is used, including metrics for previously known classes (pre mAP), the

current task's known classes (cur mAP), and all known classes up to the current task (both mAP). To assess the detection ability of unknown-class objects, the Unknown Recall (UR) is the key metric. Additionally, Absolute Open Set Error (A-OSE) and the Wilderness Impact (WI) factor, proposed by [1], are used to evaluate the model's overall ability to detect both known- and unknown-class objects. A-OSE represents the absolute number of errors where the model mistakenly classifies unknown objects as known classes, while WI quantifies how the model's ability to detect unknown-class objects impacts its detection performance on known classes.

4.2. Implementation Details

The implementation is carried out using Python 3.7, with model training and inference performed in the PyTorch deep learning framework on two GeForce RTX 3090 GPUs. The backbone network is ResNet-50, initialized with ImageNet [42] pre-trained weights. The number of queries, N, is set to 100 with a feature dimension of 256. During training, the number of pseudo-candidate boxes randomly selected for unknown classes, r, is set to 5. Two similarity thresholds δ_1 and δ_2 are 0.5 and 0.8, respectively. In the inference stage, the threshold of NMS is set to 0.6 by default, aligning with the standard practice reported in [38].

4.3. Main Results

This section presents a detailed comparative analysis of TGOOD against various OWOD algorithms. To maintain fairness in comparison, we follow prior work by removing the energy evaluation module from ORE, denoting the modified version as ORE-EBUI. For OWOD SPLIT dataset, we categorize those models into two groups: those that utilize pseudo-labels and those that do not. Tables 1 and 2 summarize the comparison results for these two categories. Table 3 presents the performance comparison between TGOOD and existing models on the MS-COCO SPLIT dataset. Table 4 shows the performance comparison of TGOOD with classic OWOD methods relying on objectness scores to select pseudo-labels on the COCO-O SPLIT. A downward arrow (↓) indicates that lower values signify better model performance, while an upward arrow (↑) indicates that higher values are preferred. Best results are marked in red, with the next results indicated in blue.

Table 1 compares TGOOD with existing pseudo-label-based OWOD methods under the OWOD SPLIT.

Table 1. Performance comparison of TGOOD vs. methods with pseudo-labels (OWOD SPLIT).

Task IDs	**Task 1**				**Task 2**						**Task 3**						**Task 4**			
	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	0.0621	10,459	56.00	4.9	0.0282	10,445	52.70	26.00	39.40	2.9	0.0211	7990	38.20	12.70	29.70	3.9	29.60	12.40	25.30
OW-DETR	0.0571	10,240	59.20	7.5	0.0278	8441	53.60	33.50	42.90	6.2	0.0156	6803	38.30	15.80	30.80	5.7	31.40	17.10	27.80
SA	0.0417	4889	56.20	1.9	0.0213	2546	53.39	26.49	39.94	0.8	0.0146	2120	38.04	12.81	29.63	0.1	30.11	13.31	25.91
UC-OWOD	0.0136	9294	50.66	2.4	0.0117	5602	33.13	30.54	31.84	3.4	0.0073	3801	28.80	16.34	24.65	8.7	25.57	15.88	23.14
RandBox	0.0240	4498	61.80	10.6	0.0078	1880	-	-	45.30	6.3	0.0054	1452	-	-	39.40	7.8	-	-	35.40
RE-OWOD	0.0449	-	59.70	9.1	0.0330	-	54.11	37.26	45.64	9.9	0.0241	-	43.06	24.64	37.59	11.4	37.99	28.66	35.66
CAT	0.0581	7070	59.90	21.8	0.0263	5902	54.00	33.60	43.80	18.6	0.0177	5189	42.10	19.80	34.70	23.9	35.10	17.10	30.60
TGOOD (ours)	0.0620	4995	60.31	23.1	0.0253	2598	51.92	34.31	43.12	18.7	0.0157	1978	41.06	23.10	35.07	22.1	35.06	19.02	31.05

Among all the comparative methods, ORE-EBUI, OW-DETR, SA, UC-OWOD, and RandBox all employ pseudo-label strategies based on objectness scores. A common deficiency of these methods is their weak detection capability for unknown-class objects, as indicated by their low UR values. Particularly, RandBox, despite achieving an advantage in detection accuracy for known classes with its diffusion model-based detection framework, only achieved recall rates of 10.6%, 6.3%, and 7.8% for unknown classes in tasks 1, 2, and 3, respectively. We attribute the insufficiency of these models in detecting unknown classes primarily to the mechanism of pseudo-label generation. Since the learning of objectness scores is based on labeled known-class objects, models tend to misjudge can-

didates containing parts of known classes as unknown-class objects when selecting labels based on objectness scores, thereby diminishing the model's ability to distinguish between known- and unknown-class objects and resulting in an inability to efficiently identify true unknown-class objects. In contrast, TGOOD, through its SRS module, generates unbiased pseudo-labels for unknown classes under textual guidance, significantly enhancing the model's detection capability for unknown-class objects, achieving recall rates of 23.1%, 18.7%, and 22.1% for unknown classes in tasks 1, 2, and 3, respectively.

Additionally, RE-OWOD and CAT, which employ additional pseudo-label generation strategies such as selective search [20], also demonstrate decent detection performance. Compared to these methods, TGOOD shows comparable or superior performance. Although RE-OWOD's detection accuracy for known classes is slightly higher than that of TGOOD in tasks 2, 3, and 4, its recall rate for unknown-class objects is almost half that of TGOOD. Overall, TGOOD has achieved a better balance between the detection performance of known and unknown classes among all comparative methods, especially excelling in the recall rate of unknown classes. However, there is still room for improvement in the model's performance on the WI and A-OSE metrics, indicating that there is further research potential in enhancing the model's ability to distinguish between known- and unknown-class objects, which will become one of the focal points of future research.

Table 2 presents a performance comparison between TGOOD and existing OWOD methods that do not use pseudo-labels under the OWOD SPLIT. Among those methods, OCPL aims to reduce the overlap between known- and unknown-class distributions in the feature space to construct more discriminative feature representations; 2B-OCD employs an object-centered calibrator to identify candidate boxes with scores above a certain threshold in the *bg* as unknown classes; PROB utilizes class-agnostic Gaussian distributions to model object features; Ann adopts a label transfer learning paradigm to decouple features of known- and unknown-class objects. The essence of these methods is to enhance the separability of known- and unknown-class objects in the feature space. Compared to the classic objectness score-based pseudo-label methods, approaches that do not rely on pseudo-labels indeed achieve higher recall rates for unknown classes. This indirectly confirms the bias problem towards known classes in objectness score-based pseudo-label methods. However, the proposed TGOOD in this paper, with the synergistic effect of multiple modules, has achieved optimal or near-optimal results in both the detection accuracy of known classes and the recall rate of unknown classes in all task phases.

Table 2. Performance comparison of TGOOD vs. methods without pseudo-labels (OWOD SPLIT).

Task IDs	**Task 1**				**Task 2**						**Task 3**						**Task 4**			
	WI	**AOSE**	**mAP (↑)**	**UR**	**WI**	**AOSE**	**mAP (↑)**			**UR**	**WI**	**AOSE**	**mAP (↑)**			**UR**	**mAP (↑)**			
	(↓)	**(↓)**	**Cur**	**(↑)**	**(↓)**	**(↓)**	**Pre**	**Cur**	**Both**	**(↑)**	**(↓)**	**(↓)**	**Pre**	**Cur**	**Both**	**(↑)**	**Pre**	**Cur**	**Both**
OCPL	0.0423	5670	56.64	8.3	0.0220	5690	50.65	27.54	39.10	7.7	0.0162	5166	38.63	14.74	30.67	11.9	30.75	14.42	26.67
2B-OCD	0.0480	-	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
PROB	0.0569	5195	59.50	19.4	0.0344	6452	55.70	32.20	44.00	17.4	0.0151	2641	43.00	22.20	36.00	19.6	35.70	18.90	31.50
Ann	0.0604	8332	56.67	12.8	0.0269	9454	51.96	29.13	40.55	5.0	0.0157	6635	40.82	14.56	32.07	9.8	31.68	13.09	27.03
TGOOD (ours)	0.0620	4995	60.31	23.1	0.0253	2598	51.92	34.31	43.12	18.7	0.0157	1978	41.06	23.10	35.07	22.1	35.06	19.02	31.05

Table 3 compares TGOOD with existing OWOD methods under the MS-COCO SPLIT. The experimental results demonstrate that TGOOD also significantly enhances the model's detection performance on the more challenging MS-COCO SPLIT dataset. Specifically, TGOOD excels in detecting unknown classes, achieving recall rates of 29.4%, 29.0%, and 35.1% for tasks 1, 2, and 3, respectively, surpassing all comparative methods. However, we observed that in task 1, although TGOOD showed substantial improvement in the detection accuracy of known classes compared to ORE-EBUI, its performance was still behind methods employing the Deformable DETR framework [43], such as OW-DETR, PROB, and CAT. We attribute this gap to the basic framework used. The FQR-CNN framework adopted by TGOOD is more lightweight compared to Deformable DETR.

However, as the tasks progress, TGOOD demonstrated superior performance in both known-class detection accuracy and unknown-class recall in tasks 2, 3, and 4. Notably, a significant issue common to comparative methods is that the model's detection capability for current known classes is markedly lower than for previously known classes, with the cur-mAP value significantly lower than the pre-mAP value. TGOOD effectively reduced this discrepancy and achieved the optimal value in overall detection accuracy (both mAP). We believe this is due to our proposed RRM module, which enables the model to learn more discriminative foreground object features, thereby enhancing the model's generalization and robustness across different tasks.

Table 3. Performance comparison of TGOOD with open-environment object detection methods (MS-COCO SPLIT).

Task IDs	**Task 1**				**Task 2**						**Task 3**						**Task 4**			
	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	-	-	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
OW-DETR	0.0458	19,815	71.50	5.7	0.0499	19,749	62.80	27.50	43.80	6.2	0.0248	9233	45.20	24.90	38.50	6.9	38.20	28.10	33.10
CAT	0.0234	2126	70.65	24.5	0.0330	4441	65.83	35.54	50.68	22.2	0.0208	3545	51.09	32.82	45.00	25.0	45.48	34.90	42.84
PROB	0.0196	1915	73.85	17.3	0.0307	3400	66.15	36.19	50.42	22.1	0.0170	1552	47.72	30.27	41.91	24.5	42.80	31.72	40.03
TGOOD (ours)	0.0443	2490	63.81	29.4	0.0244	1367	54.23	48.87	51.55	29.0	0.0174	1405	49.96	43.00	47.64	35.1	47.46	45.02	46.85

Table 4 compares the detection performance of TGOOD with traditional methods, ORE-EBUI and OW-DETR, which use objectness scores for pseudo-labeling, on the COCO-O SPLIT dataset. The average performance across the six domain datasets shows that TGOOD's ability to detect known classes is comparable to or slightly better than ORE-EBUI and OW-DETR. Notably, TGOOD maintains superior unknown-class recognition capabilities across unseen domains, outperforming the comparison methods. This is attributed to the effective guidance provided by text information containing abstract semantics. The general nature of semantic information in the text helps TGOOD sustain high generalization performance even with detection data from various domain fields.

Table 4. Performance comparison of TGOOD with classic OWOD methods (COCO-O SPLIT).

Task IDs	**Task 1**				**Task 2**						**Task 3**						**Task 4**			
	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	0.1115	1518	20.77	15.3	0.0482	1525	19.19	14.89	17.04	11.8	0.0252	1066	14.93	7.47	12.44	10.4	12.42	8.85	11.53
OW-DETR	0.1245	5590	4.68	28.9	0.0625	4076	4.13	0.55	2.34	28.8	0.0307	2243	1.72	3.54	2.32	23.9	0.05	1.07	0.30
TGOOD (ours)	0.1380	678	19.07	52.8	0.0518	401	14.42	19.28	16.85	48.5	0.0274	221	14.87	9.66	13.13	55.0	14.16	12.73	13.80

4.4. Ablation Study

Extensive ablation experiments were conducted to validate the effectiveness of TGOOD. Unless stated otherwise, these experiments were performed on task 1 using the OWOD SPLIT dataset setting.

4.4.1. Components of TGOOD

To evaluate the effectiveness of each module in TGOOD, we conducted ablation experiments for each module individually. The results are presented in Table 5. The baseline, shown in the first row, is the FQR-CNN model without any enhancements. This baseline highlights that the basic FQR-CNN detection model lacks the capability to detect objects of unknown classes.

"Obj1" refers to selecting a candidate query with the highest objectness score as the unknown-class object, similar to the pseudo-labeling method used in ORE-EBUI. The results in the second row demonstrate that "Obj1" allows our basic detection model to exhibit some capability in detecting unknown-class objects. However, it is insufficient

for effectively distinguishing between known and unknown categories, as indicated by the relatively high A-OSE score. "Obj5" involves selecting the top five candidate queries with the highest objectness scores as unknown-class objects, akin to the pseudo-labeling method used in OW-DETR. The third row of results shows that increasing the number of pseudo-labeled candidates enhances the model's ability to identify unknown-class objects, with the UR value improving from 10.0% to 13.7%. Nevertheless, this approach significantly degrades the accuracy of known-class object detection, with the mAP decreasing from 56.42% to 55.87%.

Table 5. Ablation experiments of TGOOD components.

Line ID	Baseline	Obj1	Obj5	SRS	OTM	RRM	WI (↓)	AOSE (↓)	mAP (↑)	UR (↑)
1	✓						0.0718	79,516	57.06	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	5143	59.96	22.8
6	✓			✓	✓	✓	0.0620	4995	60.31	23.1

"SRS" refers to the pseudo-label generation module proposed in this paper. The results in the fourth row demonstrate that this module significantly enhances the recall rate for unknown-class objects, with a UR rate 2.28 times higher than that of "Obj1". Additionally, "SRS" effectively mitigates the issue of mistakenly identifying unknown-class objects as known-class objects, reducing the A-OSE score from 77,927 to 15,169 compared to "Obj1". "OTM" stands for replacing the one-to-one label matching algorithm with a one-to-many label matching algorithm. As illustrated in the fifth row, assigning multiple candidate queries to each *gt* during training strengthens the supervisory signal, thereby improving the model's performance in detecting known-class objects. Furthermore, the integrated "RRM" module enhances the prominent features of foreground objects. As shown in the sixth row, this module further improves the model's ability to detect all foreground objects, including both known- and unknown-class objects. In Figure 6, we can intuitively perceive the influence of TGOOD's various components on the model's performance in terms of mAP and UR metrics.

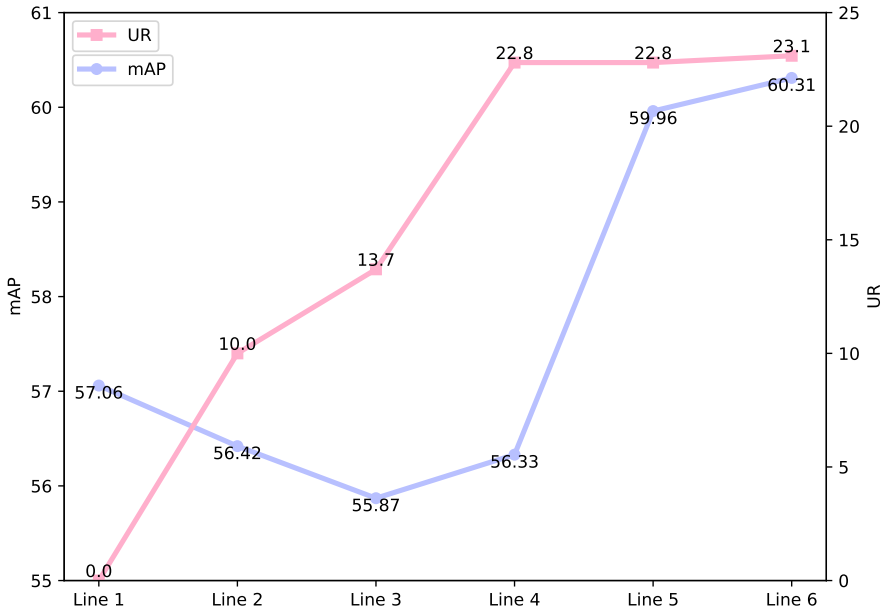


Figure 6. mAP and UR metrics under different configurations of TGOOD components.

4.4.2. The Versatility of SRS

For the pseudo-label generation module SRS proposed in this paper, we evaluated the effectiveness of its various substructures through comparative experiments. Figure 7 provides an overview of these results: Item **A** demonstrates the standard version of the text-guided pseudo-label generation strategy. It achieves a UR value of 19.4%, significantly outperforming existing pseudo-label methods that rely on objectness scores. Item **B** shows that incorporating the random de-biasing scheme further improves the UR value by an additional 3.5%. This indicates that random selection effectively reduces the model's bias towards known-class objects. Item **C** illustrates the enhanced version of the text-guided pseudo-label generation strategy. Applying the secondary filtering with a higher similarity threshold results in higher-quality pseudo-labels. This enhancement improves the model's detection performance for both known and unknown classes.

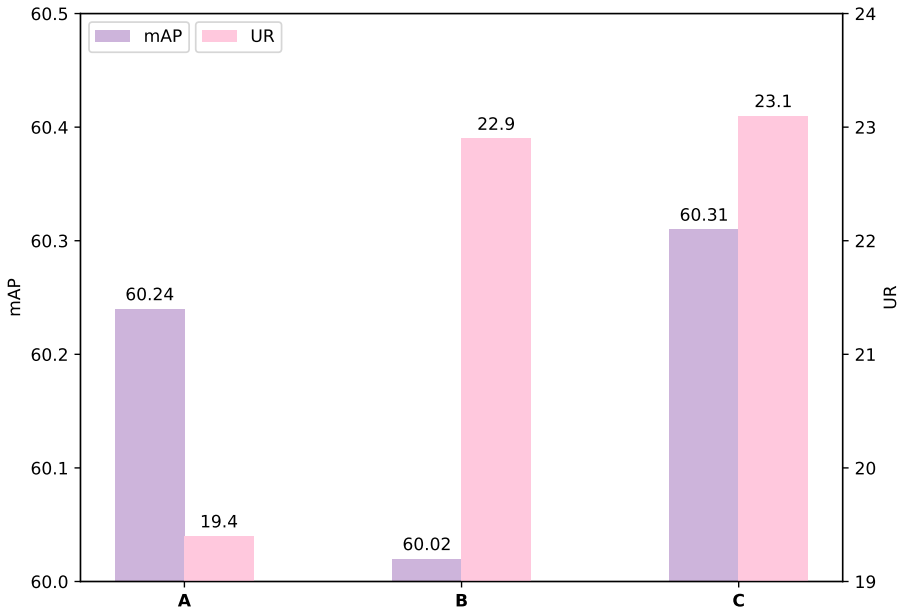


Figure 7. Ablation of substructures in SRS. (**A**) denotes the standard version of the SRS, (**B**) signifies the enhanced version of the SRS, excluding step (3), and (**C**) indicates the enhanced version of the SRS.

Additionally, since the basic target detection framework used in this paper differs from traditional OWOD methods based on objectness scores (e.g., ORE-EBUI), we tested the cross-framework effectiveness of the proposed SRS module by applying it to ORE-EBUI. During training, for candidate boxes that did not match any gt, those with extremely low objectness scores were excluded, and the fifty candidate boxes with the highest objectness scores were retained. This step aimed to reduce noise in the pseudo-labeling process. Subsequently, the SRS module was applied.

Table 6 presents the comparison results. "ORE-EBUI+SRS" denotes the integration of SRS with ORE-EBUI, while "ORE-EBUI+Obj5" represents the replacement of "Obj1" with "Obj5" in the original ORE-EBUI. The results indicate that simply increasing the number of pseudo-labels for unknown classes negatively impacts the detection performance for known classes, with "Obj5" causing a drop in known-class detection accuracy from 56.00% to 18.31%. In contrast, the SRS module not only maintains but also enhances the detection performance for known-class objects, and improves the recall rate for unknown-class objects from 4.9% to 7.6%. Furthermore, the SRS module significantly improves the WI and A-OSE indicators, which assess the model's comprehensive detection capabilities in open environments.

Table 6. Framework-agnostic validation of SMS on ORE-EBUI.

Strategies	WI (↓)	AOSE (↓)	mAP (↑)	UR (↑)
ORE-EBUI	0.0621	10,459	56.00	4.9
ORE+Obj5	0.0480	17,345	18.31	7.1
ORE+SMS	0.0528	12,120	56.03	7.6

4.4.3. Different Methods for ROI Refinement

To assess the effectiveness of the RRM module proposed in this paper, we compared it with the "RoIAttn" module from [44] and two other commonly used feature enhancement strategies. The comparison results are presented in Table 7, where "TGOOD-RRM" refers to the TGOOD model without any enhancement module. One comparison method is an LSTM network. In this study, all ROI features are treated as a sequence, and the average of the encoded bidirectional LSTM sequences is taken to generate the enhanced ROI features. The results of this approach are shown in the "BiLstm" row. Additionally, we adopted a graph convolutional neural network (GCN) as another comparison method. In this method, the ROI features are treated as nodes in a graph, with edges formed based on the cosine similarity between the ROI features. To reduce computational complexity, the enhanced features are obtained after performing a single graph convolution update, and the results are shown in the "GCN" row.

Table 7. Comparative experiments of different ROI feature enhancement modules.

Strategies	WI (↓)	AOSE (↓)	mAP (↑)	UR (↑)
TGOOD-RRM	0.0637	5143	59.96	22.8
BiLstm	0.0637	5068	60.22	22.9
GCN	0.0634	5074	60.06	23.0
RoIAttn	0.0633	5161	59.89	23.1
TGOOD	0.0620	4995	60.31	23.1

The use of an enhancement module positively impacts the model's performance compared to not using any ROI feature enhancement module. However, our RRM module consistently demonstrates superior overall performance. The BiLSTM and GCN methods increase mAP by 0.26% and 0.1%, and UR by 0.1% and 0.2%, respectively. In contrast, the RRM module proposed in this paper achieves higher improvements, increasing mAP by 0.35% and UR by 0.2%. Notably, when detecting known-class objects, the "RoIAttn" module negatively affects the model's performance. This issue is likely due to inherent flaws in the "RoIAttn" approach, which involves two additional memory units for clustering operations. In an open environment, where unknown-class objects lack labels, this process is prone to noise interference. In contrast, the RRM module enhances foreground object features by leveraging the similarity between ROI features, thus avoiding this drawback.

4.4.4. Hyperparameter Analysis

In the SRS module proposed in this study, two key hyperparameters, δ_1 and δ_2, were evaluated through ablation experiments, with results shown in Figure 8. To optimize the model's ability to detect both known- and unknown-class objects, the final values of δ_1 and δ_2 were set to 0.5 and 0.8, respectively, as these values provided the best performance enhancement.

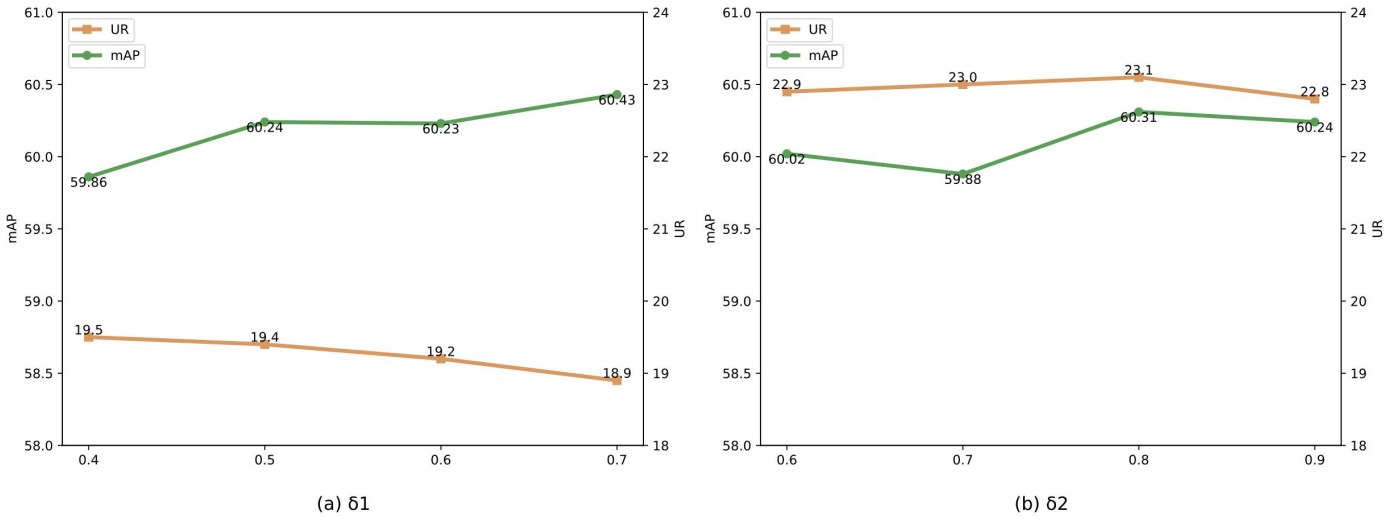


Figure 8. Ablation of hyperparameters in SRS.

4.5. Visualization

To provide a more intuitive demonstration of TGOOD's effectiveness, we present several test results in Figure 9. After training on task 1, the test results of ORE-EBUI, OW-DETR, and TGOOD are shown in the first, second, and third row, respectively.

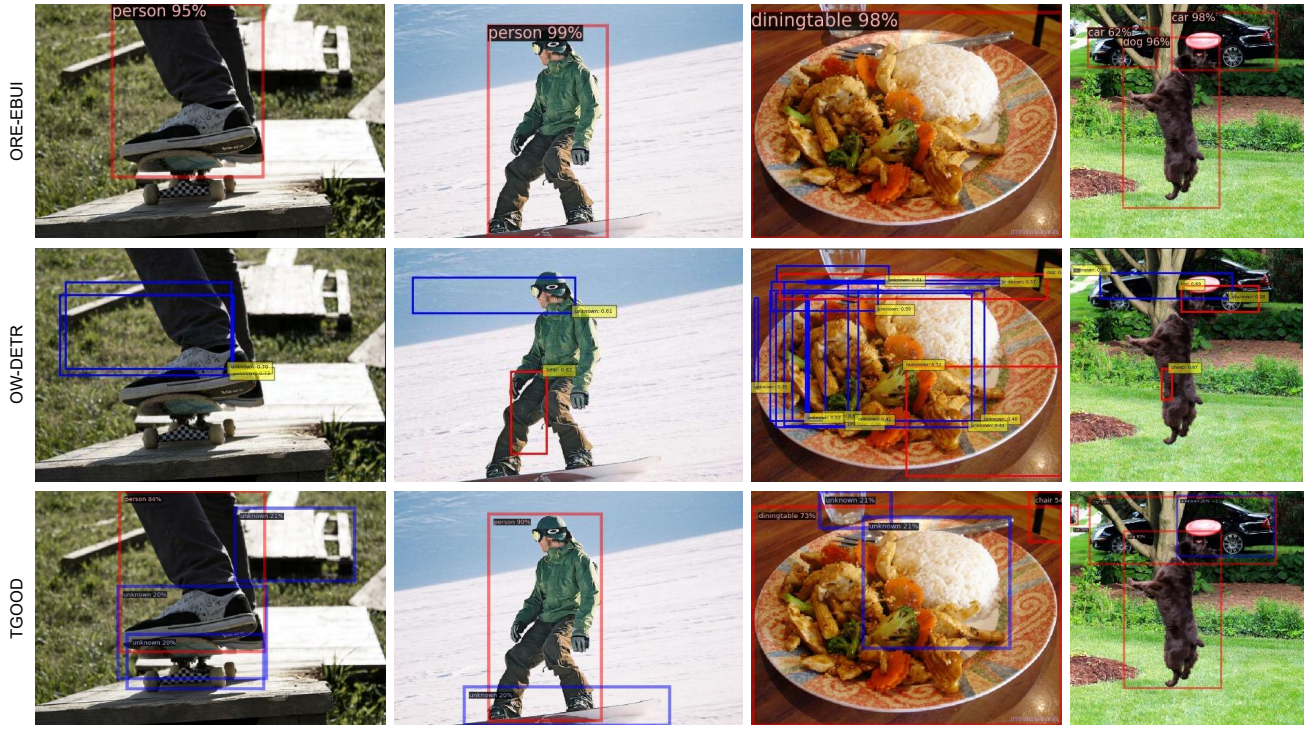


Figure 9. Visualization of TGOOD comparison with ORE-EBUI and OW-DETR. "unknown" in the figure represents unknown-class object in current stage. The red boxes indicates objects of known classes predicted by the model, while the blue boxes signifies objects of unknown classes predicted by the model.

Superior Performance: TGOOD demonstrates superior performance in detecting both known and unknown objects within images. First, TGOOD excels at accurately localizing and predicting the categories of known objects with high confidence, even in cases of severe occlusion. For example, TGOOD successfully identifies and classifies a chair leg in the top-right corner of the images in the third column and a small, heavily occluded car in

the top-left corner of the fourth column. In contrast, both ORE-EBUI and OW-DETR fail to detect these objects. Second, TGOOD shows a strong ability to detect unknown-class objects. For instance, in the first column, ORE-EBUI does not detect any unknown-class objects, while OW-DETR incorrectly identifies the *bg* as an unknown object, generating two nearly identical bounding boxes. TGOOD, however, correctly identifies unknown objects such as a skateboard, shoes, and a board in the image. Although OW-DETR detects more unknown objects in the third column (e.g., broccoli, carrot), it fails to delineate the boundaries of these objects accurately and misses known-class objects. Furthermore, it incorrectly labels two objects that do not belong to any known classes. In contrast, TGOOD accurately locates unknown objects (such as a cup and rice) and correctly identifies all known-class objects (dining table and chair).

Limitations: While the proposed method demonstrates superior object detection capabilities for both known and unknown categories compared to existing methods relying on objectness scores, it faces a limitation in the confidence level when identifying unknown-category objects. The confidence is generally low, around 20%. This limitation arises from the use of a single pseudo-label mechanism, which assigns the same label to all unknown-category objects. As a result, the purity of the pseudo-labels may be compromised. One of the key challenges for future research in open-environment object detection is how to generate high-purity pseudo-labels for unknown objects that lack labels.

5. Discussion and Conclusions

In this paper, we propose TGOOD, a cross-modal learning object detection method for open environments. The core idea behind TGOOD is to guide the model in generating high-quality pseudo-labels for unknown classes during training through the use of text containing high-level semantic information. This approach substantially reduces the model's bias towards known classes compared to traditional open-environment detection methods based on objectness scores (e.g., ORE [1] and OW-DETR [7]). Specifically, we propose three main improvement modules: (1) **SRS**, a pseudo-label generation module combining text guidance with random de-biasing to address the bias of existing strategies towards known classes; (2) **OTM**, a one-to-many label matching strategy that enriches supervisory signals during the learning process of the query-based object detection model; and (3) **RRM**, an ROI feature enhancement module that enhances the discriminability of foreground objects' ROI features using an advanced attention mechanism.

Comparative experiments were conducted on TGOOD using established evaluation benchmarks, such as OWOD SPLIT, MS-COCO SPLIT, and the COCO-O domain generalization benchmark newly proposed in this paper. These experiments assess the model's performance in detecting objects in open environments. These experiments evaluate the model's performance in detecting objects in open environments. While TGOOD demonstrates superior performance in detecting unknown-class objects, the prediction scores for these unknown classes remain lower than desired, as outlined in Section 4. Future research will focus on generating higher-purity pseudo-labels to further improve the model's detection capabilities for unknown-class objects in open environments.

Author Contributions: Conceptualization, X.W. and D.X.; methodology, X.W.; software, X.W.; validation, X.W. and D.X.; formal analysis, X.W. and D.X.; investigation, X.W.; resources, X.W.; data curation, X.W.; writing—original draft preparation, X.W. and D.X.; writing—review and editing, X.W. and D.X.; visualization, X.W.; supervision, D.X.; project administration, D.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

$\mathbb{R}$	real space
E	text feature
Q	query feature
ROI	region of interest
SRS	Similarity-Random-Similarity,text-guided pseudo-label generation strategy
OTM	one-to-many matching strategy
RRM	ROI features Refinement Module
TGOOD	Text-Guided Unknown Pseudo-Labeling for Open-world Object Detection
mAP	mean Average Precision
UR	Unknown Recall
A-OSE	Absolute Open Set Error
WI	Wilderness Impact
Obj1	selecting 1 candidate query with the highest objectness score as the unknown-class object
Obj5	selecting 5 candidate queries with the highest objectness scores as the unknown-class objects

References

1. Joseph, K.; Khan, S.; Khan, F.S.; Balasubramanian, V.N. Towards open world object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5830–5840.
2. Pershouse, D.; Dayoub, F.; Miller, D.; Sünderhauf, N. Addressing the Challenges of Open-World Object Detection. *arXiv* **2023**, arXiv:2303.14930.
3. Wu, Y.; Zhao, X.; Ma, Y.; Wang, D.; Liu, X. Two-branch objectness-centric open world detection. In Proceedings of the 3rd International Workshop on Human-Centric Multimedia Analysis, Lisboa, Portugal, 10 October 2022; pp. 35–40.
4. Yu, J.; Ma, L.; Li, Z.; Peng, Y.; Xie, S. Open-world object detection via discriminative class prototype learning. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 626–630.
5. Zohar, O.; Wang, K.C.; Yeung, S. Prob: Probabilistic objectness for open world object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 11444–11453.
6. Dong, N.; Zhang, Y.; Ding, M.; Lee, G.H. Open world DETR: Transformer based open world object detection. *arXiv* **2022**, arXiv:2212.02969.
7. Gupta, A.; Narayan, S.; Joseph, K.; Khan, S.; Khan, F.S.; Shah, M. OW-DETR: Open-world Detection Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
8. Ma, S.; Wang, Y.; Wei, Y.; Fan, J.; Li, T.H.; Liu, H.; Lv, F. Cat: Localization and identification cascade detection transformer for open-world object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19681–19690.
9. Wang, Y.; Yue, Z.; Hua, X.S.; Zhang, H. Random Boxes Are Open-world Object Detectors. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 6233–6243.
10. Wu, Z.; Lu, Y.; Chen, X.; Wu, Z.; Kang, L.; Yu, J. UC-OWOD: Unknown-classified open world object detection. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 193–210.
11. Yang, S.; Sun, P.; Jiang, Y.; Xia, X.; Zhang, R.; Yuan, Z.; Wang, C.; Luo, P.; Xu, M. Objects in Semantic Topology. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
12. Zhao, X.; Ma, Y.; Wang, D.; Shen, Y.; Qiao, Y.; Liu, X. Revisiting open world object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *34*, 3496–3509. [CrossRef]
13. Zhang, W.; Cheng, T.; Wang, X.; Chen, S.; Zhang, Q.; Liu, W. Featurized query r-cnn. *arXiv* **2022**, arXiv:2206.06258.
14. Chen, Q.; Chen, X.; Wang, J.; Zhang, S.; Yao, K.; Feng, H.; Han, J.; Ding, E.; Zeng, G.; Wang, J. Group detr: Fast detr training with group-wise one-to-many assignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 6633–6642.
15. Jia, D.; Yuan, Y.; He, H.; Wu, X.; Yu, H.; Lin, W.; Sun, L.; Zhang, C.; Hu, H. Detrs with hybrid matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19702–19712.
16. Everingham, M.; Gool, L.V.; Williams, C.K.I.; Winn, J.M.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
17. Lin, T.Y.; Maire, M.; Belongie, S.J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.

18. Mao, X.; Chen, Y.; Zhu, Y.; Chen, D.; Su, H.; Zhang, R.; Xue, H. COCO-O: A Benchmark for Object Detectors under Natural Distribution Shifts. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 6339–6350.
19. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
20. Uijlings, J.R.R.; van de Sande, K.E.A.; Gevers, T.; Smeulders, A.W.M. Selective Search for Object Recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154 – 171. [CrossRef]
21. Ma, Y.; Li, H.; Zhang, Z.; Guo, J.; Zhang, S.; Gong, R.; Liu, X. Annealing-Based Label-Transfer Learning for Open World Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 11454–11463.
22. Jaiswal, A.; Wu, Y.; Natarajan, P.; Natarajan, P. Class-agnostic object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 919–928.
23. Kim, D.; Lin, T.Y.; Angelova, A.; Kweon, I.S.; Kuo, W. Learning open-world object proposals without learning to classify. *IEEE Robot. Autom. Lett.* **2022**, *7*, 5453–5460. [CrossRef]
24. Gonçalves, G.R.; Sena, J.; Schwartz, W.R.; Caetano, C.A. Pixel-level Class-Agnostic Object Detection using Texture Quantization. In Proceedings of the 2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Natal, Brazil, 24–27 October 2022; Volume 1, pp. 31–36.
25. Saito, K.; Hu, P.; Darrell, T.; Saenko, K. Learning to detect every thing in an open world. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 268–284.
26. Huang, H.; Geiger, A.; Zhang, D. Good: Exploring geometric cues for detecting objects in an open world. *arXiv* **2022**, arXiv:2212.11720.
27. Maaz, M.; Rasheed, H.; Khan, S.; Khan, F.S.; Anwer, R.M.; Yang, M.H. Class-agnostic Object Detection with Multi-modal Transformer. In Proceedings of the 17th European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022.
28. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 8748–8763.
29. Rao, Y.; Zhao, W.; Chen, G.; Tang, Y.; Zhu, Z.; Huang, G.; Zhou, J.; Lu, J. Denseclip: Language-guided dense prediction with context-aware prompting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18082–18091.
30. Zhao, S.; Zhang, Z.; Schulter, S.; Zhao, L.; Vijay Kumar, B.; Stathopoulos, A.; Chandraker, M.; Metaxas, D.N. Exploiting unlabeled data with vision and language models for object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 159–175.
31. Wei, T.; Chen, D.; Zhou, W.; Liao, J.; Tan, Z.; Yuan, L.; Zhang, W.; Yu, N. Hairclip: Design your hair by text and reference image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18072–18081.
32. Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; Lischinski, D. Styleclip: Text-driven manipulation of stylegan imagery. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2085–2094.
33. Xu, H.; Ghosh, G.; Huang, P.Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; Feichtenhofer, C. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv* **2021**, arXiv:2109.14084.
34. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
35. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
36. Kuhn, H.W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83–97. [CrossRef]
37. Peyré, G.; Cuturi, M. Computational optimal transport: With applications to data science. *Mach. Learn.* **2019**, *11*, 355–607. [CrossRef]
38. Ge, Z.; Liu, S.; Li, Z.; Yoshie, O.; Sun, J. Ota: Optimal transport assignment for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 303–312.
39. Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In Proceedings of the 26th International Conference on Neural Information Processing Systems—Volume 2, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 2292–2300.
40. Kim, J.; Choi, J.; Choi, H.J.; Kim, S.J. Shepherding Slots to Objects: Towards Stable and Robust Object-Centric Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19198–19207.
41. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning, PMLR, Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.

42. Ridnik, T.; Ben-Baruch, E.; Noy, A.; Zelnik-Manor, L. Imagenet-21k pretraining for the masses. *arXiv* **2021**, arXiv:2104.10972.
43. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
44. Liang, X.; Song, P. Excavating roi attention for underwater object detection. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 2651–2655.