UN-DETR: Promoting Objectness Learning via Joint Supervision for Unknown Object Detection

Haomiao Liu*, Hao Xu*, Chuhuai Yue*, Bo Ma†,

Beijing Institute of Technology
{ndwxhmz, xuhao_cs, chuhuaiyue, bma000}@bit.edu.cn

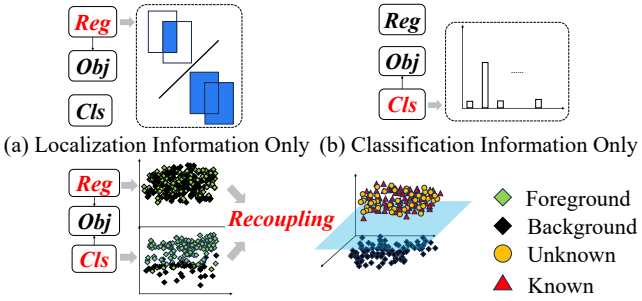
Abstract

Unknown Object Detection (UOD) aims to identify objects of unseen categories, differing from the traditional detection paradigm limited by the closed-world assumption. A key component of UOD is learning a generalized representation, i.e. objectness for both known and unknown categories to distinguish and localize objects from the background in a class-agnostic manner. However, previous methods obtain supervision signals for learning objectness in isolation from either localization or classification information, leading to poor performance for UOD. To address this issue, we propose a transformer-based UOD framework, UN-DETR. Based on this, we craft Instance Presence Score (IPS) to represent the probability of an object's presence. For the purpose of information complementarity, IPS employs a strategy of joint supervised learning, integrating attributes representing general objectness from the positional and the categorical latent space as supervision signals. To enhance IPS learning, we introduce a one-to-many assignment strategy to incorporate more supervision. Then, we propose Unbiased Query Selection to provide premium initial query vectors for the decoder. Additionally, we propose an IPS-guided post process strategy to filter redundant boxes and correct classification predictions for known and unknown objects. Finally, we pretrain the entire UN-DETR in an unsupervised manner, in order to obtain objectness prior. Our UN-DETR is comprehensively evaluated on multiple UOD and known detection benchmarks, demonstrating its effectiveness and achieving state-of-the-art performance.

Code — https://github.com/ndwxhmzz/UN-DETR

Introduction

Deep learning-based vision solutions have achieved remarkable success in the past (Krizhevsky, Sutskever, and Hinton 2012; He et al. 2016; Vaswani 2017), but their generalization performance in open scenarios still faces significant challenges. Within the closed-world assumption, conventional object detection frameworks (Ren et al. 2015; Redmon et al. 2016; Carion et al. 2020) are limited to detecting objects belonging to predefined categories present in the training set, thereby disregarding objects outside these predefined categories. This limitation leads to detrimental outcomes in real-world scenarios where high precision is essential. To advance from conventional detectors to open-world detectors, numerous outstanding works (Bendale and Boult 2016; Zheng et al. 2022) have emerged. Among them, Liang et al. (2023) clarified the definition of Unknown Object Detection (UOD).



(a) Localization Information Only (b) Classification Information Only

(c) Recoupled Positional and Categorical Lantent Space

Figure 1: Joint supervision for objectness learning

UOD can be viewed as a two-step problem: 1) Locating all objects, both known and unknown, in a class-agnostic manner; and 2) Assigning specific categories to these objects. The first step hinges on how to learn the generalized features of any object, i.e., objectness, to distinguish them from the background. Difficulty arises because, under the definition of UOD, unknown objects lack supervision, and their objectness is mostly generalized from known categories. New methods (Wu et al. 2022a; Liang et al. 2023; Zohar, Wang, and Yeung 2023) have been developed to learn objectness, but they still struggle with low recall and precision, often misclassifying unknown objects for background or vice versa.

Our core insight is to explain the deficiencies of previous methods by their lack of considering both known category and location information in objectness learning. Classification information represents both class-specific categorization and the class-agnostic probability of being foreground. UnSniffer (Liang et al. 2023), only utilizes position prediction (IoU between predictions and ground truths)(Figure

*These authors contributed equally.
†Corresponding author.

1(a)), may misclassify high-IoU boxes without instances as foreground. Class-agnostic positional information is essential for accurate object localization. PROB (Zohar, Wang, and Yeung 2023) ignores positional information and merges the objectness and classification head (Figure 1(b)), reducing localization accuracy. We hypothesize that extracting general objectness from complementary positional and categorical latent spaces may address the aforementioned issue.

To validate our theory, we developed UN-DETR, the first Transformer-based UOD framework, and introduced the Instance Presence Score (IPS), which integrates elements representing general objectness from positional and categorical latent spaces. Concretely, we introduce an IPS Predictor (IPP), alongside the original classification head and regression head, to directly output IPS. The optimization process is jointly supervised by signals from both spaces, ensuring their mutual complementarity. To enhance IPS learning, we propose a one-to-many assignment strategy that introduces more positive samples. Subsequently, we propose Unbiased Query Selection, to optimize the initialization of queries by replacing the original classification head with the learned IPP. Moreover, we propose an IPS-guided post-process strategy, filtering redundant boxes and further separating known from unknown objects. Finally, we pretrain the entire UN-DETR in an unsupervised manner using the region prior and the self-supervised encoder, to obtain objectness priors. Essentially, our approach improves the robustness and generalization of the detector, elevating UOD performance to a new level (Figure 1(c)).

To summarize, the contributions of our work are as follows:

- We reveal that a major flaw in previous UOD methods is the separate use of classification and localization information when learning objectness. To address this issue, we propose the very first Transformer-based UOD framework, UN-DETR.
- Our core design involves using a dedicated IPP to learn IPS under the joint supervision singals from complementary positional and categorical latent space. Moreover, IPS also participates in multiple stages of the UN-DETR, including query selection and post-processing.
- Extensive experiments on both UOD and known detection benchmarks clearly demonstrate that our approach surpasses previous methods, achieving state-of-the-art performance.

Related Work

Transformer-Based Detector

Since Detection Transformer (DETR) (Carion et al. 2020) pioneered the first fully end-to-end object detector, transformer-based detectors have gained significant attention for their outstanding performance and scalability. Deformable DETR (D-DETR) (Zhu et al. 2020) further improved this by introducing deformable attention, which efficiently samples key elements, reducing computational demands, speeding up convergence, and enhancing performance.

To enhance training efficiency, recent methods (Li et al. 2022) have optimized the one-to-one assignment in DETR, which pairs each ground-truth object with a single prediction. Group-DETR (Chen et al. 2023) implemented a group-wise one-to-many assignment, performing decoder self-attention within each group. Similarly, Co-DETR (Zong, Song, and Liu 2023) introduced a collaborative training scheme with multiple auxiliary heads using one-to-many label assignments.

Building on these advancements, we developed UN-DETR, the first transformer-based UOD method, based on D-DETR (Zhu et al. 2020). We apply one-to-many assignment specifically for IPS learning to facilitate generalized feature extraction from more positive sample queries.

Unknown Object Detection and Related Tasks

Recent years have seen the emergence of tasks aimed at detecting unknown objects. Open Set Detection (OSD) (Bendale and Boult 2016) requires identifying and excluding unknown samples, but issues with overconfidence affect accuracy. Open World Object Detection (OWOD) (Joseph et al. 2021) aims to detect both known and unknown objects, yet the absence of labels for unknowns prevents precise evaluation. Recently, Liang et al. (2023) further clarify the UOD evaluation protocol with both the precision and recall of unknown objects as metrics.

Early OSD methods (Bendale and Boult 2016; Liang, Li, and Srikant 2018), focus on distinguishing known and unknown objects. Techniques like maximum softmax probability (Hendrycks and Gimpel 2017), minimum Mahalanobis distance (Denouden et al. 2018), energy scores (Liu et al. 2020), and virtual outliers (Du et al. 2022a,b) have been used. However, these methods primarily enhance known object detection, reducing unknown object recall. In contrast, Our approach seeks unbiased detection of unknowns.

Joseph et.al (2021) introduce OWOD with the ORE detector, featuring RPN-based unknown pseudo-labeling and contrast clustering. OW-DETR (Gupta et al. 2022) and others (Yang et al. 2022; Zhao et al. 2023; Wu et al. 2022b; Gupta et al. 2022; Ma et al. 2023) explored various pseudo-labeling methods. Yet, pseudo-labeling often misclassifies non-objects as unknowns, reducing precision.

Recent efforts (Liang et al. 2023; Zohar, Wang, and Yeung 2023; Wu et al. 2022a) focus on objectness scores without pseudo-labeling, reducing false positives. (Wu et al. 2022a) extended ORE with a localization-based objectness head, improving recall. Similarly, (Liang et al. 2023) introduced a localization-based GOC score using only known samples for supervision, and a graph-based boxes decision scheme. On the other hand, (Zohar, Wang, and Yeung 2023) introduced a framework for classification-based objectness estimation that alternate between probability distribution estimation and objectness likelihood maximization of known objects in the embedded feature space, and ultimately estimating the objectness for different proposals.

In this paper, we propose a learnable objectness score, IPS, that integrates positional and categorical signals for robust objectness representation, avoiding pseudo-labeling pitfalls.

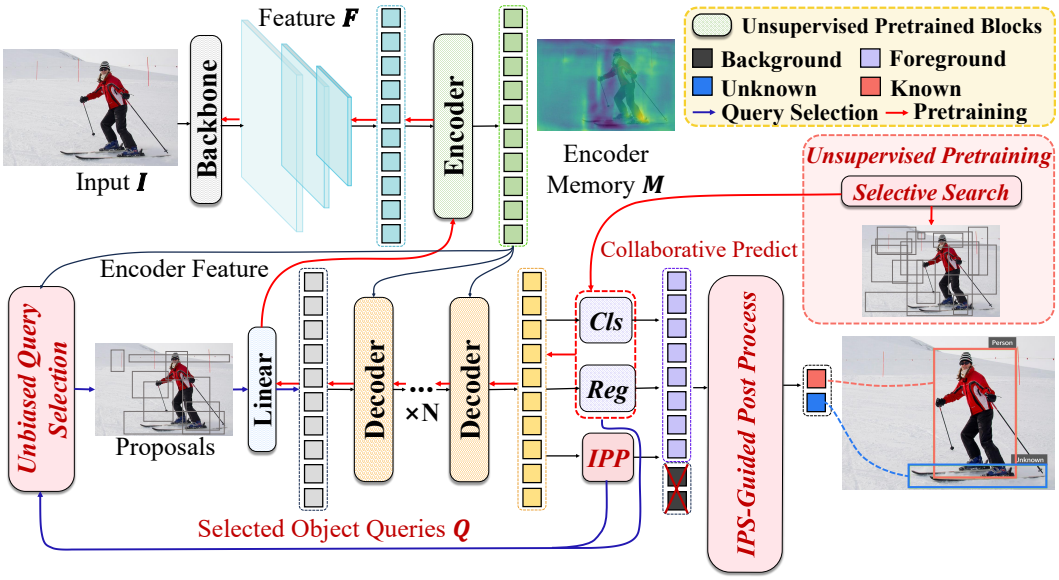


Figure 2: The overall architecture of UN-DETR

Preliminary

Two-Stage Deformable DETR Pipeline

In D-DETR Pipeline, an input image $\boldsymbol{I}$ is processed by a backbone network to extract features, which are fed into an encoder using attention mechanisms to produce an enhanced feature sequence. In the decoder, N_{query} object queries are updated through self-attention and cross-attention with the encoder's output, leading to refined queries $q \in R^D$. These are then processed by the bounding box regression head f_{bbox} and classification head f_{cls} for final predictions. A one-to-one bipartite matching using L_{match} ensures alignment with ground-truth (GT) labels for supervision. And two-stage D-DETR leverages region proposals generated by the encoder as initial object queries for further refinement in the decoder. The top-scoring region proposals are determined by applying f_{bbox} and f_{cls} to the encoder's output feature maps.

Unknown Object Detection

The task of unknown object detection represents an extension of conventional object detection frameworks. Referring to (Du et al. 2022b; Joseph et al. 2021), the problem of unknown detection is formulated as follows. Given dataset $D = \{\boldsymbol{I}, \boldsymbol{Y}\}$, where the N input images are denoted as $\boldsymbol{I} = \{\mathbf{I}_1, \ldots, \mathbf{I}_N\}$, with corresponding labels $\boldsymbol{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_N\}$. Each $\mathbf{Y}_i = \{y_1, \ldots, y_k\} (i \in [1, 2, \ldots, N])$ contains a set of objects with $y_k = [l_k, b_k]$, where l_k is the class label for bounding box b_k represented by x_k, y_k, w_k, h_k. The known class set is denoted as $\mathcal{K} = \{1, 2, \ldots, C\}$, and the unknown class is denoted as $\mathcal{U} = \{C+1\}$.

The model is trained on data labeled only with known-classes objects $\{(\mathbf{I}_n, \mathbf{Y}_n) \mid l_k \in \mathcal{K}\}_{n=1}^{N_{\text{train}}}$, but tested on data including both known and unknown objects $\{(\mathbf{I}_n, \mathbf{Y}_n) | l_k \in \mathcal{K} \cup \mathcal{U}\}_{n=1}^{N_{\text{test}}}$, where N_{train} is the image number in the training set, N_{test} for that in the test set, and $N = N_{\text{test}} + N_{\text{train}}$.

UN-DETR

Overall Architecture

The architecture of UN-DETR is depicted in Figure 2. The processing pipeline is as follows: An image $\boldsymbol{I}$ of dimensions $H \times W \times C$ is fed into the backbone to extract features $\boldsymbol{F}$. These features are then processed by the encoder to produce feature memory $\boldsymbol{M}$. The top K initial object queries are refined and filtered using $\boldsymbol{M}$, fed into the regression head, and a linear layer to produce M object queries $\boldsymbol{Q}$. Subsequently, N decoder layers transform $\boldsymbol{Q}$ into query embeddings $\boldsymbol{E}$, capturing the necessary spatial and semantic information for accurate Unknown Object Detection (UOD). These embeddings $\boldsymbol{E}$ are processed through three branches—classification, regression, and IPP—to predict potential instances. The predicted bounding boxes $\boldsymbol{B}$ undergo post-processing guided by IPS to yield the final prediction $\boldsymbol{P}$ of object instances. Prior to end-to-end training, the UN-DETR model undergoes unsupervised pretraining to establish objectness priors.

Instance Presence Score Predictor

Instance Presence Score As we discussed in the Introduction, extracting representation from complementary positional and categorical latent spaces favors objectness learning. Disregarding class-agnostic categorical latent space, UnSniffer (Liang et al. 2023) solely relies on position prediction (Intersection over Union, IoU, between the predictions and ground truth), may misclassify prediction boxes with high IoU scores but not containing any instances as foreground. Neglected by PROB (Zohar, Wang, and Yeung 2023), class-agnostic positional latent space directly impacts

the detector's ability to locate potential objects. PROB only integrates the objectness head within the classification head but does not adjust the regression head, impacting the localization accuracy of unknown objects, leading to partial or oversized object prediction boxes.

Furthermore, to validate the above conjecture, we extract representations from either positional or categorical latent spaces as objectness score and visualize the discriminability score of feature maps during inference, as shown in Figure 3. Solely considering categorical latent space, as shown in Figure 3(b), the model exhibited higher discriminability score between instances and the background but poor distinction among instances themselves. Conversely, when using only the positional latent space representations, as illustrated in Figure 3(c), the model demonstrated greater distinctiveness among different instances but lesser between instances and the background. These experiments suggest that the two latent spaces are complementary in learning objectness.

In object detection task, predictions require locating objects (regression) and identifying them (classification). This highlights the need to integrate both positional and categorical latent spaces, suggesting that their combined use in the UOD task is more effective than treating them separately. Therefore, we formulate IPS by leveraging attributes of objectness from them, enhancing the use of knowledge learned from known categories and improving generalization to unknown objects. This approach also enhances the distinction between foreground and background, increasing robustness in diverse real-world environments, such as those with varying appearance and scale, which aligns with the primary goal of the UOD task.

Similarly, to validate the discriminability of IPS for different instances, the distinction maps are visualized as shown in Figure 3(d). After fully utilizing the representations from both latent spaces, instances are clearly distinguished from one another as well as from the background. The comparison clearly demonstrates effectiveness and superiority of IPS.

Jointly Supervised IPP To accurately estimate the IPS, we design a specialized IPP alongside the classification head and regression head. Specifically, the IPP is a simple single-layer feed-forward neural network and it inputs the query embedding $e_i \in \boldsymbol{E}$ and computes the corresponding IPS $I(e_i)$. For IPP training, we propose a jointly supervised strategy. First, the query embedding e_i is fed into two heads $f_{cls}(e_i)$, $f_{bbox}(e_i)$ to obtain the representations e_i in lower dimensions e_{cls}, e_{bbox}, which represent categorical and positional information obeying the two potential spaces S_{cls}, S_{bbox}, respectively. To remove the class-related components in e_{cls}, e_{bbox}, we extract the embedding composed of components representing generic objectness e^o_{cls}, e^o_{bbox} from each of the two representations. For e^o_{bbox}, we compute generalized IoU (GIoU) after transforming it into a bounding box $\hat{b}_i$ with the matched GT b_i. GIoU is a metric based on spatial overlap that is independent of categories and thus performs better when confronted with unseen categories in the training set, and we formalize e^o_{bbox} in terms of GIOU as follows:

$$\hat{e}^o_{bbox,\sigma^{pos}} = \text{GIoU}(b_i, \hat{b}_{\sigma^{pos}}) \tag{1}$$

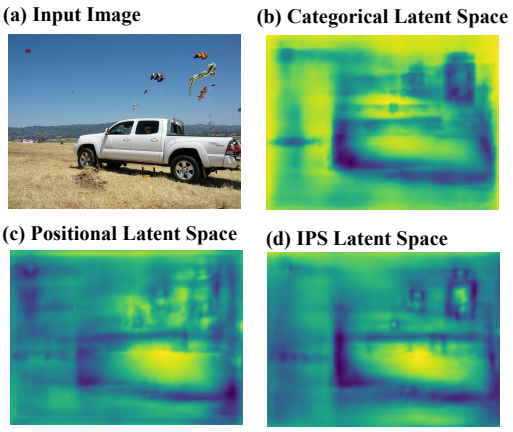


Figure 3: Visualizations of discriminability scores

where σ^{pos} is the index of the positive sample queries used to train IPP, as explained in the next subsection. For e^o_{cls} , we represent the generalized objectness by the sum of $\mathcal{K}$ (number of known categories) logits, since the $\mathcal{K}$ logits represent the confidence of all the categories appearing in the training set, and the sum represents the overall probability of the foreground. This avoids the model's tendency to favor a particular category or categories, and reflects the robustness of the model to different categories of objects in various environments. So e^o_{cls} is formalized as follows:

$$\hat{e}^o_{cls,\sigma^{pos}} = \hat{P}_{f\sigma^{pos}} \tag{2}$$

where P_f is the sum of $\mathcal{K}$ logits. To capitalize on the complementarity of e^o_{bbox}, e^o_{cls}, we set the objective probability $P_o(e_i) = \alpha \cdot e^o_{bbox} + \beta \cdot e^o_{cls}$, which serves as the supervised signal for IPP training. Therefore, the loss function of IPP is as follows:

$$L^{\text{H}}_{\text{IPS}} = \ell_1(P_o, I(e_i)_{\sigma^{pos}}), \text{if GIoU}(b_i, \hat{b}_{\sigma^{pos}}) > \tau \tag{3}$$

$$L^{\text{L}}_{\text{IPS}} = \ell_1(C, I(e_i)_{\sigma^{pos}}), \text{if GIoU}(b_i, \hat{b}_{\sigma^{pos}}) \leq \tau \tag{4}$$

where ℓ_1 denotes the L1 loss, C is the objective constant and τ is the GIOU threshold. Here we introduce C to increase the distinctiveness of IPS learning and maintain the stability of IPP training. The total IPS loss can be represented as:

$$L_{\text{IPS}} = L^{\text{H}}_{\text{IPS}} + L^{\text{L}}_{\text{IPS}} \tag{5}$$

Finally, the entire loss of UN-DETR can be represented as:

$$L_{\text{UN-DETR}} = \lambda_1 \cdot L_{\text{IPS}} + \lambda_2 \cdot L_{\text{cls}} + \lambda_3 \cdot L_{\text{bbox}} \tag{6}$$

where L_{cls} and L_{bbox} are consistent with the classification and regression loss of D-DETR, and $\lambda 1$, $\lambda 2$, and $\lambda 3$ are the weights of the loss, set to 3, 2, and 5, respectively.

One-to-Many Assignment In the Preliminary, we note that D-DETR employs one-to-one assignment to associate GT with potential objects. However, previous research has

highlighted several issues with this approach, such as inefficient training. As a result, various one-to-many assignment methods have been proposed. These methods primarily aim to enhance convergence speed and training stability. Nonetheless, in UOD, beyond the aforementioned challenges, the most significant challenge is the invisibility of labels for unknown objects during training. Consequently, the joint supervision process of IPP must rely solely on features from the positional and categorical latent spaces. This reliance introduces potential uncertainty in the supervisory information, hindering the model parameters from iterating towards a more optimal solution, thereby compromising training stability and adversely affecting model performance.

One-to-many assignment allows for more flexible use of all supervision even when some of it is noisy or uncertain. By allowing multiple predictions to capture the same GT, the model aggregates information from these different matches, learning more stable and comprehensive object features.

Therefore, we propose a simple one-to-many assignment strategy to provide more positive samples for jointly supervised IPP training. Specifically, we introduce one set of suboptimal queries besides one set of best-matching queries, since they also match known instances with high probability and are easily obtained by bilateral matching. The index for sub-optimal queries can be formalized as:

$$\hat{\sigma} = \arg \min_{\sigma \in \mathcal{S}_{\mathcal{N}} / \hat{\sigma^*}} \sum_i L_{\text{match}}(y_i, \hat{y}_{\sigma(i)}) \tag{7}$$

where $\hat{\sigma}$ is the index of the sub-optimal matching queries, which has the same length as the index $\hat{\sigma^*}$ of the best-matching queries. Then, the index of all positive sample queries for IPP training can be represented as:

$$\hat{\sigma^{pos}} = \hat{\sigma} \cup \hat{\sigma^*} \tag{8}$$

Note that we only use the sub-optimal queries obtained from the one-to-many assignment for the jointly supervised IPP, and not for the classification and regression head. This is because the supervision for the regression head and classification head is derived directly from the labels and is inherently accurate. Introducing suboptimal supervision may therefore negatively impact their training.

Unbiased Query Selection

To effectively select object queries relevant to the current input, the two-stage D-DETR employs an additional regression head along with a classification head to refine and filter appropriate proposals, initializing them as object queries. Specifically, this newly introduced classification head is trained using category labels represented as a tensor of all zeros, allowing the first dimension to indicate the probability that the input is a likely object. Consequently, during prediction, the top K proposals are selected based on the first dimension of the category prediction. This straightforward approach, however, disregards information from other dimensions predicted by the classification head, introducing a bias that causes this head to favor larger outputs in the first dimension. This bias leads to sparse gradient updates, as other classes contribute minimally, ultimately affecting the convergence and accuracy of this head, which is crucial to the overall detection task.

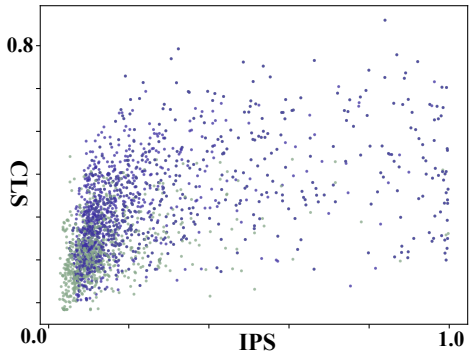


Figure 4: Visualization of classification scores and IPS for encoder features

The additional classification head introduced in two-stage D-DETR essentially treats query selection as an instance recognition task, similar to the target of IPP. To address the bias mentioned earlier, we follow this approach and introduce an additional IPP, naming our method Unbiased Query Selection (UQS). In UQS, we replace the original classification head with an additional IPP to filter proposals. The supervision information of IPP reflects its class-agnostic nature, as its predictions denote the probability if the query represents a foreground object, independent of any category-specific priors. This lack of reliance on class-specific information eliminates potential bias, allowing IPP to focus solely on the objectness of the query itself, rather than being influenced by class-based assumptions.

To analyze the effectiveness of UQS, we visualize the classification scores and IPS of the encoder features in Figure 4. In the scatterplot, blue and green dots represent encoder features from models trained with Unbiased Query Selection and vanilla query selection, respectively. Dots closer to the top right indicate higher-quality features, meaning a greater likelihood of representing foreground objects. Notably, the top right corner has more blue dots than green, indicating that queries filtered by UQS are of higher quality and more likely to contain instances. Introducing the extra IPP, which incorporates positional information, allows for verifying the spatial accuracy of queries, reducing false detections. This mitigates the class-specific bias, improving the quality of object queries forwarded to the decoder. Additionally, integrating bias elimination into the loss function further optimizes gradients, enhancing training stability.

IPS-Guided Post Process

In D-DETR, the top-K bounding boxes are directly output as detection results. However, in UOD, unknown objects may outnumber known ones. A fixed K constrains the recall rate and is unreliable to set manually. Additionally, the original D-DETR post-processing doesn't work due to the one-to-many assignment in IPP training and the need to further differentiate unknown from known categories. To solve this, we propose the IPS-Guided Post Process, which includes IPS-Guided Non-Maximum Suppression (NMS) to remove redundant proposals and a dual-criteria unknown distinguish

protocol.

IPS-Guided NMS Traditional NMS methods(Neubeck and Van Gool 2006) rank proposals based on classification confidence. However, when labels for unknown objects are missing, these methods may fail and even discard well-predicted boxes. Furthermore, (Jiang et al. 2018) highlight the inconsistency between localization and classification information in traditional NMS, where boxes with accurate localization may have low scores, or highly scored boxes may have poor localization. To eliminate as much of the background as possible and address the inconsistency, we propose IPS-Guided NMS. Considering the distance between center points of bounding boxes, we rank all boxes with IPS and calculate Distance IoU (DIoU) (Zheng et al. 2020) to measure overlap.

Dual-Criteria Unknown Distinguish Protocol Having identified all bounding boxes containing foreground objects, the next step is to classify the remaining proposals into known and unknown categories, after removing any redundant bounding boxes. Although there is no dedicated classification head for unknowns, utilizing IPS and classification confidence together still distinguishes between known and unknown. If both classification confidence and IPS are above set thresholds, the object is assigned to a known category. If classification confidence is low but IPS is high, the object is recognized but not confidently categorized, hence it's classified as unknown.

Unsupervised Pretraining with Objectness Priors

Self-supervised representation learning can reduce the amount of labeled data required by the model and improve its representation capability. We hope to improve the performance of UN-DETR by utilizing the related technology. To this end, following DETRreg (Bar et al. 2022), we pretrain the entire UN-DETR in an unsupervised manner to obtain objectness priors with both localization and classification. Specifically, we utilize an unsupervised region proposal generator, Selective Search (Uijlings et al. 2013), to match object localization boxes. Moreover, we adopt a self-supervised image encoder, SwAV (Caron et al. 2020), to align the object embeddings used for classification. Note that to avoid introducing additional data, we only use the training set for unsupervised pretraining.

Experiment

Following the UOD Benchmark, we utilize COCO-OOD, COCO-Mixed (Liang et al. 2023), and VOC (Everingham et al. 2010) as test sets and employ mAP, U-AP, U-F1, U-PRE, and U-REC as evaluation metrics, as detailed in the Appendix.[1]

Implementation Details

In training, we use ResNet50 as the UN-DETR backbone. Moreover, the entire UN-DETR is pretrained on the VOC training set in an unsupervised manner (Bar et al. 2022). We introduce only one additional set of suboptimal queries for joint supervision of IPP training. The weight parameters α and β are empirically set to 0.6 and 0.4, respectively. In Eq. 4, C is set to 0.5 and τ is set to 0.6.

[1]http://arxiv.org/abs/2412.10176

Methods	mAP	U-AP	U-F1	U-PRE	U-REC
MSP	47.0	21.3	31.4	27.9	35.9
Mahalanobis	44.7	12.9	27.1	30.9	24.1
Energy score	**47.4**	21.3	30.8	26.0	37.7
VOS	46.9	20.5	31.7	29.1	34.8
ORE	24.3	21.4	25.5	16.3	**78.2**
OW-DETR	42.0	3.3	5.6	3.0	38.0
PROB	36.0	4.3	17.5	11.7	35.2
UnSniffer	46.4	<u>45.4</u>	<u>47.9</u>	<u>43.3</u>	53.5
Ours	<u>47.2</u>	**47.0**	**54.9**	**54.5**	<u>55.3</u>

Table 1: Comparisons with other methods in the VOC-test and COCO-OOD datasets. The mAP is based on VOC-test, while the other metrics are from COCO-ODD. The best results are in **bold**, second best are <u>underlined</u>.

Results

Quantitative Analysis. Tables 1 and 2 present the results of our method UN-DETR, alongside 8 classic or recent state-of-the-art methods, on the UOD Benchmark. Notably, on the COCO-OOD dataset, our UN-DETR outperforms others in metrics except for U-REC. Particularly for U-F1 and U-PRE, our method surpasses the second-best result by 7.0% and 11.2%, respectively. ORE's U-REC outperforms our method but also recall many non-objects. This is evident as our method's U-AP, U-F1, and U-PRE are all approximately twice as good as ORE's. On the COCO-Mixed dataset, our method maintains a lead in U-AP and U-PRE, exceeding the second-best result by 4.1% and 0.5%, respectively. The aforementioned results demonstrate that our UN-DETR surpasses the previously leading method in unknown object detection, attributable to our proposed jointly supervised IPP training. In addition, experimental results on both the VOC-test and COCO-Mixed datasets show that UN-DETR performs comparably to existing methods on known detections, which demonstrates that it does not improve unknown detections by sacrificing known detections.

In general, our method outperforms other OSD methods as it is designed for detecting rather than excluding all unknown objects. Compared to pseudo-label-based OWOD detectors, our UN-DETR only introduces one additional set of query samples using one-to-many assignment strategy, which reduces the interference of negative samples and thus leads larger on U-PRE. Most importantly, benefiting from jointly supervised IPP training from both positional and categorical latent spaces, our approach exceeds other objectness-based approaches, as will be further demonstrated in the ablation study of Sec. 5.3.

Qualitative Analysis. Figure 5 visualizes the results of various methods. It is evident that our UN-DETR outperforms other methods both in localizing and identifying unknown

Methods	mAP	U-AP	U-F1	U-PRE	U-REC
MSP	36.4	5.5	16.9	19.0	15.3
Mahalanobis	35.1	5.1	14.9	20.7	11.6
Energy score	36.4	4.9	16.9	16.7	17.1
VOS	36.4	5.1	17.2	18.4	16.3
ORE	21.3	14.0	17.5	10.3	**59.2**
OW-DETR	**41.4**	0.7	2.5	1.4	16.1
PROB	40.1	9.4	26.2	17.0	56.7
UnSniffer*	35.9	14.8	**26.7**	19.3	40.9
Ours	34.0	**18.9**	24.7	**19.8**	32.8

Table 2: Comparisons with other methods in the COCO-Mixed datasets. * means that our replication results.

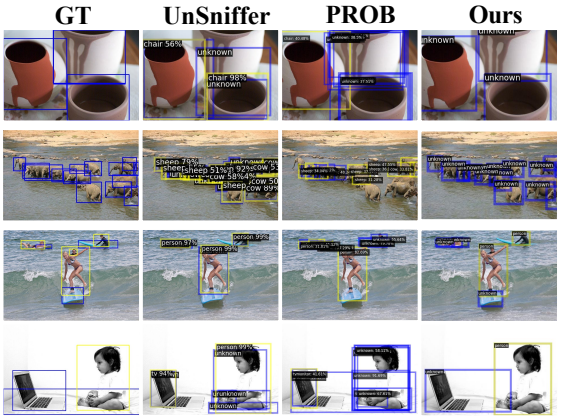


Figure 5: Example results on COCO-OOD (first two rows) and COCO-Mixed (last two rows) datasets. Detections are overlaid on known (yellow) and unknown (blue) objects.

objects. On the one hand, UnSniffer, which ignores category information, misidentifies some unknown objects, such as the elephant in the second row, and misses some obvious unknowns, such as the surfboard in the third row. On the other hand, PROB, which neglects positional information, has difficulty in accurately localizing the boundaries of objects, such as the water cup in the first row. On the contrary, our method accurately detects the most unknown objects, which benefits from the fact that we recouple generic objectness from both positional and categorical latent space. In addition, our method can accurately distinguish between known and unknown and exclude redundant boxes, which is attributed to our proposed IPS-guided post-processing. More visualization results are presented in the Appendix.

Ablation Study

To examine the contribution of each component in our method, we conduct adequate ablation experiments as presented in Table 3. For IPP training, individual supervision from either the latent space of category or position significantly degrades the performance of the model (rows 1 and 2). If only the classification information is considered, the detector may miss some unknown objects and dramatically reduce the recall, while if only the regression information is focused on, the detector may confuse unknowns with knowns and recall non-objects, thus impairing the precision. Our UN-DETR trade-offs both of the above to obtain excellent precision and recall simultaneously.

Row	Component	U-AP	U-F1	U-PRE	U-REC
1	IPP only cls	17.2	22.1	**57.3**	13.7
2	IPP only reg	40.9	23.9	14.7	**63.4**
3	one-to-one	46.0	51.3	51.4	51.1
4	one-to-three	46.0	53.7	53.5	53.9
5	UQS origin cls	39.7	51.1	56.7	46.4
6	UQS only reg	45.9	50.9	48.4	53.7
7	w/o IPS-NMS	46.7	52.3	50.1	56.8
8	w/o Unsupervised	45.8	44.5	35.0	60.1
9	All	**47.0**	**54.9**	54.5	55.3

Table 3: Ablation studies on COCO-OOD.

In the one-to-many assignment strategy, we introduce only one additional set of samples, which outperforms the original one-to-one matching (row 3) and prevents the performance decrease from introducing more sets due to the possible negative sample noise (row 4).

For UQS, we train an additional IPP to replace the original classification head for query filtering, and the experimental results show that IPP outperforms not only the original classification head (row 5), but also the IPP supervised only with regression information (row 6). This proves the superiority of our joint supervision and the effectiveness of UQS.

To demonstrate the effectiveness of our proposed post-processing, we replace our IPS-Guided NMS with the original NMS (row 7), and the experimental results show that all the metrics of UN-DETR decrease. When unsupervised pretraining is not used (row 8), the U-PRE of UN-DETR decreases significantly. It's because pretraining provides a prior on objectness for the model, allowing it to initially acquire a certain class-agnostic perceptual ability after pretraining, which is crucial for UOD.

Conclusion

We propose a novel transformer-based UOD method UN-DETR that outperforms existing state-of-the-art methods. We investigate the deficiencies of current methods in exploiting complementary classification and regression predictions, leading to unstable objectness learning. Therefore, the core insight of our approach is jointly supervised objectness IPS learning from both positional and categorical latent spaces. Then, we propose a one-to-many assignment strategy to provide more positive samples for IPS learning. Furthermore, IPS is employed for query selection and post-processing in UN-DETR due to its encoding class-agnostic categorization and localization information. Finally, we pretrain the entire UN-DETR in an unsupervised manner to obtain the objectness prior.

Acknowledgments

This work was supported by the Joint Funds of the National Natural Science Foundation of China (No. U2441206) and the National Natural Science Foundation of China (No. 62072042).

References

Bar, A.; Wang, X.; Kantorov, V.; Reed, C. J.; Herzig, R.; Chechik, G.; Rohrbach, A.; Darrell, T.; and Globerson, A. 2022. Detreg: Unsupervised pretraining with region priors for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14605–14615.

Bendale, A.; and Boult, T. E. 2016. Towards Open Set Deep Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1563–1572.

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

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33: 9912–9924.

Chen, Q.; Chen, X.; Wang, J.; Zhang, S.; Yao, K.; Feng, H.; Han, J.; Ding, E.; Zeng, G.; and Wang, J. 2023. Group detr: Fast detr training with group-wise one-to-many assignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6633–6642.

Denouden, T.; Salay, R.; Czarnecki, K.; Abdelzad, V.; Phan, B.; and Vernekar, S. 2018. Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. *arXiv preprint arXiv:1812.02765*.

Du, X.; Wang, X.; Gozum, G.; and Li, Y. 2022a. Unknown-Aware Object Detection: Learning What You Don't Know from Videos in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13678–13688.

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

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Gupta, A.; Narayan, S.; Joseph, K.; Khan, S.; Khan, F. S.; and Shah, M. 2022. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9235–9244.

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

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*.

Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; and Jiang, Y. 2018. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European conference on computer vision (ECCV)*, 784–799.

Joseph, K.; Khan, S.; Khan, F. S.; and Balasubramanian, V. N. 2021. Towards open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5830–5840.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L. M.; and Zhang, L. 2022. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13619–13627.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *International Conference on Learning Representations*.

Liang, W.; Xue, F.; Liu, Y.; Zhong, G.; and Ming, A. 2023. Unknown Sniffer for Object Detection: Don't Turn a Blind Eye to Unknown Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3230–3239.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33: 21464–21475.

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

Neubeck, A.; and Van Gool, L. 2006. Efficient non-maximum suppression. In *18th international conference on pattern recognition (ICPR'06)*, volume 3, 850–855. IEEE.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Uijlings, J. R.; Van De Sande, K. E.; Gevers, T.; and Smeulders, A. W. 2013. Selective search for object recognition. *International journal of computer vision*, 104: 154–171.

Vaswani, A. 2017. Attention is All You Need. *arXiv preprint arXiv:1706.03762*.

Wu, Y.; Zhao, X.; Ma, Y.; Wang, D.; and Liu, X. 2022a. Two-branch objectness-centric open world detection. In *Proceedings of the 3rd International Workshop on Human-Centric Multimedia Analysis*, 35–40.

Wu, Z.; Lu, Y.; Chen, X.; Wu, Z.; Kang, L.; and Yu, J. 2022b. UC-OWOD: Unknown-classified open world object detection. In *European Conference on Computer Vision*, 193–210. Springer.

Yang, S.; Sun, P.; Jiang, Y.; Xia, X.; Zhang, R.; Yuan, Z.; Wang, C.; Luo, P.; and Xu, M. 2022. Objects in Semantic Topology. In *International Conference on Learning Representations*.

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

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

Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; and Ren, D. 2020. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12993–13000.

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

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

Zong, Z.; Song, G.; and Liu, Y. 2023. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6748–6758.