CSPCL: Category Semantic Prior Contrastive Learning for Deformable DETR-Based Prohibited Item Detectors

Mingyuan Li[1] **Tong Jia**[1,2*] **Hao Wang**[1] **Bowen Ma**[1]
Hui Lu[1] **Shiyi Guo**[1] **Da Cai**[1] **Dongyue Chen**[1]

[1] College of Information Science and Engineering, Northeastern University, China
[2] State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University
542027743@qq.com, jiatong@ise.neu.edu.cn, ddsywh@yeah.net,
2010285@stu.neu.edu.cn, 2603813543@qq.com, guoshiyi@ise.neu.edu.cn,
2210329@stu.neu.edu.cn, chendongyue@ise.neu.edu.cn

Abstract

Prohibited item detection based on X-ray images is one of the most effective security inspection methods. However, the foreground-background feature coupling caused by the overlapping phenomenon specific to X-ray images makes general detectors designed for natural images perform poorly. To address this issue, we propose a Category Semantic Prior Contrastive Learning (CSPCL) mechanism, which aligns the class prototypes perceived by the classifier with the content queries to correct and supplement the missing semantic information responsible for classification, thereby enhancing the model sensitivity to foreground features. To achieve this alignment, we design a specific contrastive loss, CSP loss, which comprises the Intra-Class Truncated Attraction (ITA) loss and the Inter-Class Adaptive Repulsion (IAR) loss, and outperforms classic contrastive losses. Specifically, the ITA loss leverages class prototypes to attract intra-class content queries and preserves essential intra-class diversity via a gradient truncation function. The IAR loss employs class prototypes to adaptively repel inter-class content queries, with the repulsion strength scaled by prototype-prototype similarity, thereby improving inter-class discriminability, especially among similar categories. CSPCL is general and can be easily integrated into Deformable DETR-based models. Extensive experiments on the PIXray, OPIXray, PIDray, and CLCXray datasets demonstrate that CSPCL significantly enhances the performance of various state-of-the-art models without increasing inference complexity. The code is publicly available at `https://github.com/Limingyuan001/CSPCL`.

1 Introduction

Prohibited item detection based on X-ray images is an important security inspection measure, widely used in airports, postal services, government agencies, and border control areas. With the development of computer vision technologies, researchers have started using techniques such as image classification, object detection, and semantic segmentation to assist security personnel in examining packages imaged by security screening machines for prohibited items, thereby reducing the potential risks caused by work fatigue. However, as shown in Figure 1, X-ray images differ from natural light images in that they exhibit unique overlapping phenomena, which cause coupling of foreground and background features. This leads to the detection results of general object detection models being easily affected by background noise.

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

There are two mainstream approaches for improving the anti-overlapping detection performance of detectors in response to the overlapping phenomenon in X-ray images. Methods based on attention mechanisms, including DAM [1], RIA [2], and FCAM [3], focus on foreground information either channel-wise or spatially. Methods based on label assignment, including IAA [4], HSS [5], and LAreg[6], provide the model with more accurate candidate boxes during the training process, allowing it to concentrate on perceiving semantic information from overlapping foreground and background features. However, the aforementioned methods are all designed for CNN-based detectors and lack a decoder structure to leverage the reliable prior knowledge provided by queries to perceive specific foreground features [7; 8; 9]. Recently, researchers have found that clarifying the classification semantic information of content queries in Deformable DETR-based models can enhance the ability of the decoder to perceive foreground features in X-ray images and improve the anti-overlapping detection ability [10; 11]. AO-DETR [10] introduces the CSA label assignment strategy, which constrains content queries to detect specific categories of objects, but its portability is relatively limited. MMCL [11] proposes a plug-and-play contrastive learning mechanism that establishes sample pairs between content queries and categorizes them according to the number of classes. This method indiscriminately repels inter-class sample pairs while attracting intra-class sample pairs, thereby clarifying the semantic information of categories. However, it lacks category prior information, which may lead to content queries not aligning with the true distribution of feature manifold of categories.

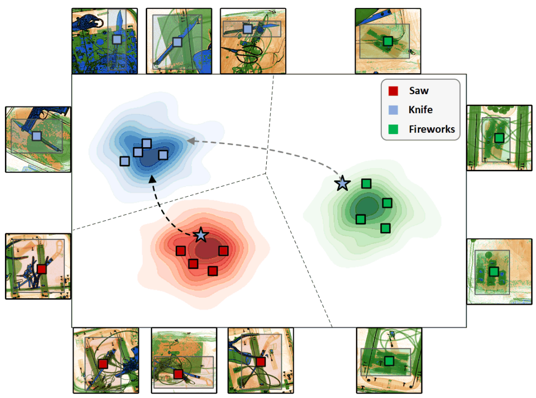


Figure 1: Feature manifold of saw (red), knife (blue), and fireworks (green). The squares and stars represent the feature representations of prohibited items and content queries, respectively.

As shown in Figure 1, the features of prohibited items such as knife, saw, and fireworks exhibit distinct differences, with varying degrees of differentiation between categories. Specifically, knife and saw have more similar color, texture, and contour information compared to knife and fireworks. Consequently, the representations of knife and saw are closer together in feature space than those of knife and fireworks. Based on this observation, we propose a novel Category Semantic Prior Contrastive Learning mechanism (CSPCL), which utilizes classifier weights as specific category prototypes to provide targeted correction and guidance to the content queries responsible for each category in the decoder layers, ensuring that the content queries align with the inherent characteristic distribution of the respective prohibited items. To achieve this multi-class multi-sample alignment, we design a contrastive loss, called the Category Semantic Prior (CSP) loss, which comprises the Intra-Class Truncated Attraction (ITA) loss and the Inter-Class Adaptive Repulsion (IAR) loss, and shows superior performance compared to classic contrastive losses [12; 13]. Specifically, we employ the ITA loss to attract content queries of the same category toward their respective category prototypes, providing the queries with category semantic prior information. It provides a gradient truncation function that stops the attraction when the similarity between the prototypes and intra-class content queries exceeds a certain threshold, thereby preventing the homogenization of intra-class content queries. Inversely, the IAR loss utilizes class prototypes to repel inter-class content queries, which has a repulsion factor adaptively adjusting the repulsion strength based on the similarity between prototype pairs, thereby facilitating inter-class content queries to learn discriminative identifying features of similar categories, and obtain adequate separability.

The main contributions of our work are as follows:

- We propose a plug-and-play CSPCL mechanism that uses contrastive learning to align content queries with the intrinsic feature distribution of the same category of prohibited items as perceived by the classifier, thereby enhancing the anti-overlapping detection capability of Deformable DETR-based models without increasing inference complexity.
- We propose the CSP loss tailored for the multi-class multi-sample alignment task, which more effectively aligns class prototypes with content queries than classic contrastive losses.
- The CSPCL mechanism demonstrates strong generalization across various Deformable DETR variants, such as RT-DETR [14], DINO [9], and AO-DETR [10], improving detection

accuracy on two prohibited item datasets, including PIXray [15] and OPIXray [2]. Furthermore, we validate CSPCL on the large-scale prohibited item detection datasets PIDray [1] and CLCXray [6], further advancing the state-of-the-art standards.

2 Related Work

DETR-like Detectors. DEtection TRansformer (DETR) pioneered the removal of the Non-Maximum Suppression (NMS) post-processing step, streamlining the pipeline of object detectors. However, it still has two major drawbacks: slow training convergence and the ambiguous meaning of queries [7]. To address the issue of slow convergence, Deformable DETR [16] introduces deformable attention, which only attends to a small set of key sampling points around one selected reliable reference point, significantly improving the training convergence speed. Stale-DINO [17] uses the IoU score as ground truth to supervise the predicted classification scores of positive examples, thereby improving the stability of matching. Group-DETR [18] proposes a group-wise one-to-many assignment that conducts one-to-one assignment within each group of object queries, resulting in one ground-truth object being assigned to multiple predictions. H-DETR [19] introduces an auxiliary one-to-many matching branch during training. Co-DETR [20] integrates multiple traditional one-to-many matching strategies. To clarify the meaning of queries, Conditional DETR [21] introduces 2D reference points in the query part as reliable spatial information hints. Anchor-DETR [22] and DAB-DETR [8] introduce 4D reference boxes to further provide potential position and size information of the objects. DN-DETR [23] proposes denoising training to help the network select high-quality queries, whose spatial positions are closer to the ground truth, to predict bounding boxes. DINO [9] further proposes contrastive denoising training, teaching the network how to reject queries that are far from the ground truth.

Contrastive Learning. The core idea of contrastive learning is to train more discriminative feature representations by pulling together the representations of similar samples and pushing apart the representations of dissimilar samples. Overall, contrastive learning can be divided into unsupervised contrastive learning and supervised contrastive learning. Unsupervised contrastive learning methods can be further divided into instance-wise contrastive learning and cluster-based contrastive learning. The former is represented by methods such as InstDisc [24] and SimCLR [25], where, for any given instance, its data-augmented version is treated as an intra-class sample, while other instances are considered as inter-class samples. Typically, the InfoNCE [13] loss is used for gradient updates. The latter, represented by methods such as CC [26] and TCL [27], uses pseudo-labels generated by clustering algorithms as the basis and then employs a supervised contrastive learning framework for training. Supervised contrastive learning methods have three paradigms of loss function. The first is based on the LMMN [28], which includes Triplet loss and N-pair loss. The second is based on the Softmax function, such as AM-Softmax [29], Circle loss [30], and SupCon loss [31]. The third paradigm is based on Cross-Entropy loss, including methods like EBM [32], C2AM [33], and MMCL [11]. As far as we know, the choice among these three paradigms often depends on the specific task and the reliability of the labels, with no definitive superiority among them. In this paper, we propose a targeted contrastive loss function, CSP loss, for the task of guiding content queries with category semantic priors.

Prohibited Item Detector. Since the introduction of the first large-scale pseudo-color X-ray prohibited item detection dataset by SIXray [34], research on prohibited item object detection in X-ray images has become increasingly diverse and manifold. Prohibited item detectors are typically based on general object detectors and are improved to address the overlap phenomena in X-ray images [35; 36; 37; 4; 38; 39; 40]. SIXray introduced the CHR mechanism to iteratively strip away overlapping background information. OPIXray [2] proposed the DOAM module to perceive the material and edge information of foreground objects. GADet [4] and Xdet [5] proposed the IAA and HSS label assignment strategies to mitigate the foreground-background class imbalance issue caused by overlapping phenomena. FDTNet [41] and FAPID [42] transform features into the frequency domain and then use self-attention mechanisms or convolutions to extract the texture and edge information of foreground objects. AO-DETR [10] proposed the first DETR-like model in the field of prohibited item detection, utilizing the CSA strategy to train category-specific content queries that are responsible for detecting specific types of contraband, thereby enhancing the ability to detect overlapping objects. However, the portability of CSA is limited. MMCL [11] proposed a plug-and-play contrastive learning strategy that helps clarify the category-specific semantic information of

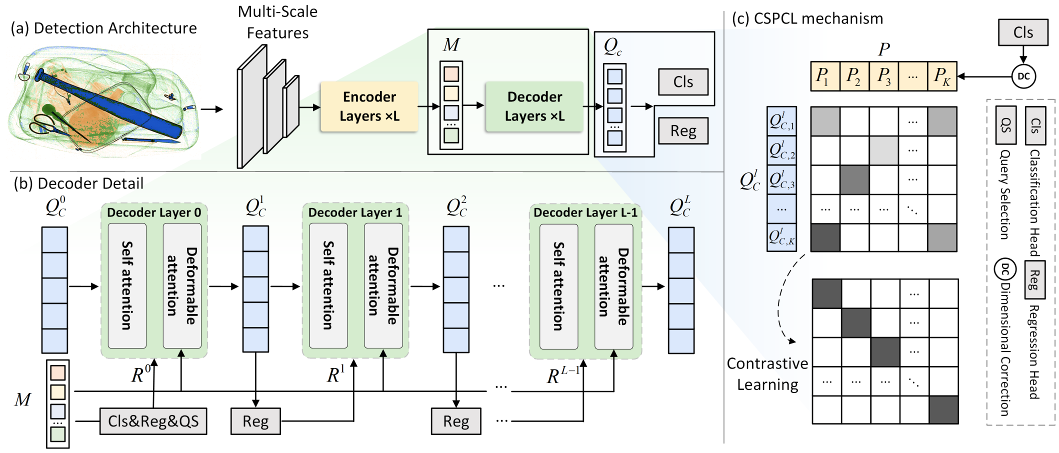


Figure 2: Illustrating the pipeline of CSPCL plugged into one Deformable DETR [16] variant DINO [9]. (a) The overall architecture of DINO. (b) The process of updating content queries in each layer of the decoder. (c) The CSPCL mechanism uses contrastive loss to leverage classifier weights as class prototypes, thereby supplementing and refining the semantic information responsible for classification in the content queries of the decoder at layer l.

content queries in Deformable DETR-based models. However, it lacks prior knowledge of the feature distribution for each category, and the indiscriminate repulsion of inter-class content queries will lead to the misalignment between the feature distributions of queries and objects within the same category. This paper presents improvements to address this issue.

3 Proposed Method

3.1 Explore the significance of feature alignment between objects and content queries

In this part, we examine the role of content queries in the decoder of Deformable DETR-based models and elucidate the importance of aligning object features with content queries. As shown in Figure 2(a), using DINO as a representative model of the Deformable DETR series, given an input image, the backbone of the model first extracts multi-scale features, which are then integrated through multiple layers of encoder layers to obtain global information, commonly referred to as the encoder memory $\mathbf{M} \in \mathbb{R}^{N\times C}$ [7], which is passed to the decoder. In the decoder, as depicted in Figure 2(b), the decoder first uses the $\mathbf{M}$ to predict a set of candidate results, including classification results $\mathbf{C} \in \mathbb{R}^{N\times K}$ and localization results $\mathbf{R} \in \mathbb{R}^{N\times 4}$, where K is the number of categories in the dataset. Then, through the query selection mechanism, the top N_{pred} most reliable predictions $\mathbf{R}^0 \in \mathbb{R}^{N_{pred}\times 4}$ are filtered based on their classification confidence.

Overall, the decoder refines and updates the randomly initialized content queries $\mathbf{Q}_c^0$ through L layers of decoder layers, using the encoder memory $\mathbf{M}$ and spatial prior information $\mathbf{R}^0$. This iterative process yields more accurate classification and localization results. The decoder layer at the l-th layer $\mathcal{D}^l$, along with the corresponding prediction head, can be expressed by the following formulas:

$$\{\mathbf{Q}_c^{l+1}, \mathbf{R}^{l+1}, \mathbf{C}^{l+1}\} = \mathcal{D}^l(\mathbf{Q}_c^l, \mathbf{R}^l, \mathbf{M}; \theta^l), \quad \mathbf{R}^{l+1} = \sigma(\sigma^{-1}(\mathbf{R}^l) + \Delta\mathbf{R}^{l+1}), \tag{1}$$

$$\{\Delta\mathbf{R}^{l+1}, \mathbf{C}^{l+1}\} = \mathcal{H}_{R,C}^l(\mathbf{Q}_c^{l+1}; \theta_{\mathcal{H}}^l), \tag{2}$$

where $l \in \{x \mid x \in \mathbb{Z}, 0 \leq x < L\}$ is the decoder block index, and L, topically set to 6, denotes the total number of decoder blocks. The θ, $\sigma(\cdot)$, $\sigma^{-1}(\cdot)$, and $\mathcal{H}_{R,C}$ denote learnable parameters, sigmoid, inverse sigmoid [16], and prediction heads of regression and classification [7], respectively. Eq. (2) shows that content queries almost directly determine the prediction results at each layer of the decoder. To analyze the update process of content queries more deeply, we further decompose the decoder in Eq. (1) into self-attention mechanisms and deformable attention mechanisms for separate explanations.

The self-attention $\mathcal{D}_s^l$ first performs element-wise addition to fuse the spatial information $\mathbf{Q}_p^l$, which is obtained by positional embedding of reference boxes $\mathbf{R}^l$ [9], with the content query $\mathbf{Q}_c^l$, resulting in the query of self-attention $\mathbf{Q}_s^l$ and the key of self-attention $\mathbf{K}_s^l$. This establishes a global mapping, and by extracting and aggregating features in $\mathbf{V}_s^l$ that have strong correlations with the spatial information, the output $\mathbf{Q}_{s,c}^l$ is obtained. The l-th self-attention layer can be denoted as follows:

$$\mathbf{Q}_{s,c}^l = \mathcal{D}_s^l(\mathbf{Q}_s^l, \mathbf{K}_s^l, \mathbf{V}_s^l; \theta_s^l), \quad \mathbf{Q}_s^l = \mathbf{K}_s^l = \mathbf{Q}_c^l + \mathbf{Q}_p^l, \quad \mathbf{V}_s^l = \mathbf{Q}_c^l, \tag{3}$$

where $\mathbf{K}$ and $\mathbf{V}$ are the key and value [43], respectively.

Deformable attention $\mathcal{D}_d^l$ uses reference boxes $\mathbf{R}^l$ to assist the deformable queries $\mathbf{Q}_d^l$, element-wise addition result of self-attention output $\mathbf{Q}_{s,c}^l$ and spatial embeddings $\mathbf{Q}_p^l$, and extract feature information from the encoder memory $\mathbf{M}$ to update the content queries $\mathbf{Q}_c^{l+1}$. The l-th deformable attention layer can be denoted as follows:

$$\mathbf{Q}_c^{l+1} = \mathbf{Q}_{d,c}^{l+1} = \mathcal{D}_d^l(\mathbf{Q}_d^l, \mathbf{R}^l, \mathbf{M}; \theta_d^l), \quad \mathbf{Q}_d^l = \mathbf{Q}_{s,c}^l + \mathbf{Q}_p^l. \tag{4}$$

From the above process, it can be seen that content queries $\mathbf{Q}_c^l$ are continuously integrated with positional information from $\mathbf{Q}_p^l$ through self-attention and deformable attention mechanisms. As the number of iterations in the decoder layer increases, the inherent category semantics in the original content query are gradually updated and replaced by the $\mathbf{Q}_p^l$ that contains location information such as contours, edges, and shapes. This results in the loss of the content query's role in providing classification-related semantic information, such as color and texture. Therefore, if we can supplement and correct the category semantics in the content queries, we can enhance the model's sensitivity to the corresponding foreground object features.

3.2 Category semantic prior contrastive learning (CSPCL)

We propose the CSPCL mechanism, as illustrated in Figure 2(c), which supplements and corrects the category semantic information of content queries by aligning them with the class prototypes perceived by the classifier weights. In the preprocessing stage, we group the content queries $\mathbf{Q}_c^l \in \mathbb{R}^{N_{pred} \times M}$ from the l-th layer based on the number of categories K, resulting in K groups of category-specific content queries $\mathbf{Q}_{c,k}^l \in \mathbb{R}^{n \times M}$, where $k \in \{x \mid x \in \mathbb{Z}, 0 < x \leq K\}$, and n is the quotient of the number of content queries N_{pred} divided by the number of categories K. Next, the classifier weights $\mathbf{W} \in \mathbb{R}^{K \times M}$ are dimensionally expanded and adjusted to align with the dimensions of $\mathbf{Q}_c^l$, resulting in prototypes $\mathbf{P} \in \mathbb{R}^{N_{pred} \times M}$. In this way, the prototype, $\mathbf{P}_k \in \mathbb{R}^{n \times M}$, which is responsible for category k, shares the same dimensions as $\mathbf{Q}_{c,k}^l$. The prototypes $\mathbf{P}_k$ and the content queries $\mathbf{Q}_{c,k}^l$ consist of n sample pairs for each category k. For this multi-class multi-sample alignment task, we specifically designed the CSP loss consisting of the Intra-Class Truncated Attraction (ITA) loss and the Inter-Class Adaptive Repulsion (IAR) loss. The IAR loss employs category prototypes to repel inter-class content queries, while the ITA loss attracts content queries towards the corresponding intra-class prototypes, thereby enhancing inter-class discriminability—especially among similar categories—while preserving essential intra-class diversity. The formula for CSP loss is as follows:

$$\mathcal{L}_{CSP}(\mathbf{P}, \mathbf{Q}, K) = \alpha \mathcal{L}_{ITA}(\mathbf{P}, \mathbf{Q}, K) + \beta \mathcal{L}_{IAR}(\mathbf{P}, \mathbf{Q}, K), \tag{5}$$

wherein α and β are hyperparameters used to balance two losses.

ITA loss. To attract category-specific content queries to the feature manifold of intra-class prohibited items, a simple and effective method is to compute the cross-entropy loss using the cosine similarity between category prototypes and intra-class content queries. However, even prohibited items of the same category still exhibit feature variations under different backgrounds and poses, meaning that intra-class content queries need to maintain a certain degree of variability to align with these differences, thereby improving the detection accuracy. Therefore, we specifically designed a gradient truncation mechanism to ensure that content queries retain enough variance. The overall formula for the ITA loss is as follows:

$$\mathcal{L}_{ITA}(\mathbf{P}, \mathbf{Q}, K) = \frac{-1}{Kn(n-1)} \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{n} \log\Big(\mathcal{T}\big(\mathrm{sim}(\mathbf{p}_i^k, \mathbf{q}_j^k), \gamma\big)\Big), \mathcal{T}(x, \gamma) = \begin{cases} 1-\gamma, & 1-\gamma < x \leq 1 \\ x, & 0 < x < 1-\gamma \\ 0, & -1 \leq x < 0 \end{cases} \tag{6}$$

where $\mathrm{sim}(\mathbf{p}_i^k, \mathbf{q}_j^k)$ is the cosine similarity between the i-th prototype of the k-th class and j-th content query of the k-th class. $\mathcal{T}(x, \gamma)$ is the gradient truncation function. As shown in Figure 3, when the input value x, i.e. $\mathrm{sim}(\mathbf{p}_i^k, \mathbf{q}_j^k)$, exceeds the threshold $1-\gamma$, the returned ITA loss value is constant $1-\gamma$, and its derivative becomes 0. At this point, the prototype stops attracting the intra-class content queries, allowing the content queries to retain the necessary variability and margin among queries. Therefore, by adjusting the value of γ, we can control the variance (diversity) of similarities of intra-class content queries.

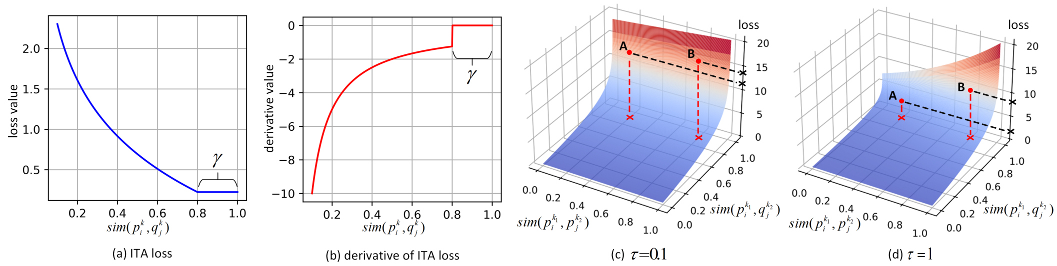


Figure 3: (a) and (b) are the curves of the ITA loss and its derivative, respectively. (c) and (d) are the 3D surface plots of our IAR loss with different τ. The samples A and B have the same $\mathrm{sim}(\mathbf{p}_i^{k_1}, \mathbf{q}_j^{k_2})$, but $\mathrm{sim}(\mathbf{p}_i^{k_1}, \mathbf{p}_j^{k_2})$ of A is smaller that of B. When $\tau = 0.1$, the IAR loss values of A and B are almost the same, but when $\tau = 1$, The loss value of A is much smaller than that of B.

IAR loss. Although the ITA attracts content queries toward the intra-class prototypes, some stubborn samples still drift away from the sample center, and may even appear in the feature manifold of other categories. A simple solution is to use the prototype to repel inter-class content queries indiscriminately. However, the differences between the prototypes of different categories vary naturally. As shown in Figure 1, the feature manifolds of knife and saw have less margin than those of knife and firework, which means the former is naturally more coupled than the latter, requiring stronger repulsion by the prototype to create enough margin for clarifying the semantic information of content queries for better detection performance [44; 45; 10; 11]. Therefore, we propose the IAR loss as follows:

$$\mathcal{L}_{IAR}(\mathbf{P}, \mathbf{Q}, K) = \frac{-1}{K(K-1)n^2} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}\left[k_1 \neq k_2\right] \mathcal{R}(k_1, k_2) \log(1 - \mathrm{sim}(\mathbf{p}_i^{k_1}, \mathbf{q}_j^{k_2})), \tag{7}$$

$$\mathcal{R}(k_1, k_2) = e^{1-\tau\cdot\left(1-\mathrm{sim}(\mathbf{p}_i^{k_1}, \mathbf{p}_j^{k_2})\right)}, \tag{8}$$

where $\mathbb{1}[k_1 \neq k_2] \in \{0, 1\}$ is an indicator function. It equals 1 if $k_1 \neq k_2$, and 0 in the other case. Additionally, $\mathcal{R}(k_1, k_2)$ is the repulsion factor, used to adjust the strength of the repulsive effect exerted by the category prototype on content queries from different categories. The higher the similarity between the prototype of the k_1-th category $\mathbf{p}_i^{k_1}$ and the prototype of the k_2-th category $\mathbf{p}_j^{k_2}$, the stronger the repulsion exerted by the prototype $\mathbf{p}_i^{k_1}$ on the content query $\mathbf{q}_j^{k_2}$. τ is the temperature coefficient, which controls the sensitivity of the IAR loss to inter-class prototype similarity. As shown in Figure 3, the larger the τ, the more sensitive the IAR loss becomes to differences between prototypes. This means that the prototype only repels content queries belonging to similar categories while minimizing the influence on content queries from unrelated categories, which helps align the features between prototypes and queries. When $\tau = 0$, $\mathcal{R}(k_1, k_2)$ becomes a constant, and the prototypes indiscriminately repel content queries from all categories. Therefore, inter-class content queries can learn discriminative identifying features for similar categories, through our IAR loss with a satisfied τ.

3.3 Plug CSPCL into target decoder layers

The CSPCL mechanism can be inserted into the decoder of Deformable DETR-like models in a plug-and-play manner without increasing model complexity, as described in Algorithm 1. Given the number of classes K in the dataset, the content queries $\mathbf{Q}_c$, localization results set R and classification results set C of all decoder layers set L, the class prototypes $\mathbf{P}$, and the set of target layers $\hat{L}$ where our CSPCL mechanism will be inserted. For each decoder layer l, the model's classification and localization detection results are matched with the ground truth through the Hungarian matching mechanism [7] to establish a one-to-one correspondence. Then, we compute the built-in loss $\mathcal{L}_{BASE}$ of the model for the current layer l based on the matched prediction result and ground truth pairs.

Additionally, we determine whether the current layer belongs to the target layers set $\hat{L}$. If it does, we use the content queries from the current layer $\mathbf{Q}_c^l$, prototypes $\mathbf{P}$, and the number of categories K to compute the CSP loss according to Eq. (5). This loss is then added to the model's inherent loss to update the current layer's loss value $\mathcal{L}^l$. Finally, we use the sum of the losses from each layer, denoted as $\mathcal{L}$, to update both the model parameters and the content queries of all layers.

3.4 Effect of the CSPCL mechanism

To analyze the effectiveness of the CSPCL mechanism for content queries, we visualize their t-SNE [46] dimensionality reduction distribution results with classifier weights. As shown in Figure 4(a) and Figure 4(c), content queries of all categories of the original DINO are scattered and evenly distributed in the feature space, while also exhibiting significant differences from the category prototypes perceived by the classifier. This suggests that during the model training process, the content queries in the decoder fail to capture the semantic information responsible for classification, which is consistent with the conclusion drawn in Section 3.1. In contrast, Figure 4(b) and Figure 4(d) demonstrate that, under the CSPCL mechanism, intra-class content queries cluster together while maintaining necessary distances, whereas inter-class content queries repel each other. Additionally, the content queries of the same category show a clear correlation with the category prototypes perceived by the model, i.e., the classifier weights. This indicates that our method successfully aligns the content queries with the inherent feature distribution of contraband, thereby addressing the deficiency in Deformable DETR-like models.

Algorithm 1 Plug CSPCL into the Decoder.

Require:
  Constant K; Set L, $\hat{L}$, $\mathbf{Q}_c$, $\mathbf{P}$, $\mathbf{R}$, $\mathbf{C}$, $\mathbf{G}$;
Ensure:
  Initialize the total loss $\mathcal{L}$ to 0;
  for $\forall$ decoder layer index $l \in L$ **do**
    $\{\mathbf{R}_i^l, \mathbf{C}_i^l; \mathbf{G}_i\} \leftarrow \mathcal{H}^l(\mathbf{R}^l, \mathbf{C}^l; \mathbf{G})$;
    $\mathcal{L}^l \leftarrow \mathcal{L}_{BASE}(\{\mathbf{R}_i^l, \mathbf{C}_i^l; \mathbf{G}_i\})$;
    if $l \in \hat{L}$ **then**
      $\mathcal{L}^l \leftarrow \mathcal{L}^l + \mathcal{L}_{CSP}(\mathbf{P}, \mathbf{Q}_c^l, K)$;
    end if
    $\mathcal{L} \leftarrow \mathcal{L} + \mathcal{L}^l$;
  end for
  update model parameters and $\mathbf{Q}_c$ to minimize $\mathcal{L}$;
  return $\mathbf{Q}_c$;

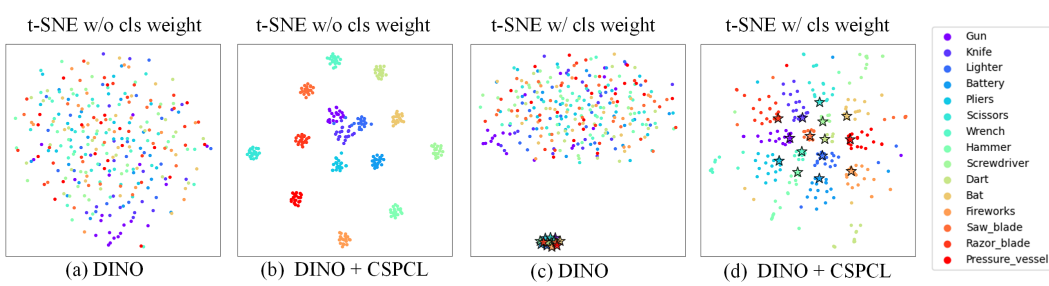
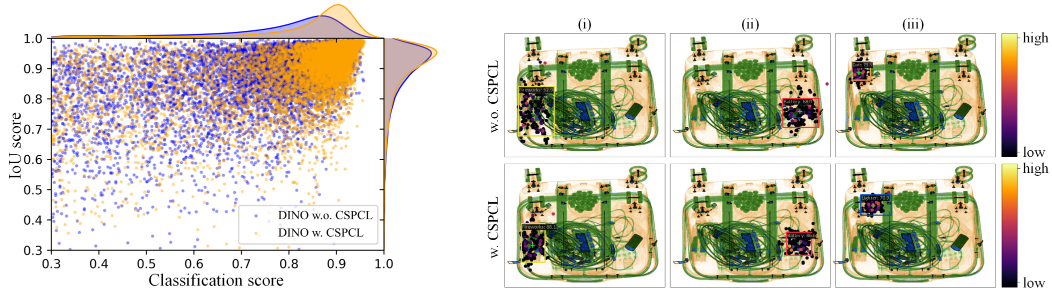


Figure 4: The t-SNE visualization results of the content queries (dots) and classifier weights (stars) from the first decoder layer. (a) and (b) show the visualization of content queries for DINO and DINO with the CSPCL mechanism, respectively. (c) and (d) show the visualization results of the classifier weights and content queries simultaneously reduced for the two models, DINO and DINO with the CSPCL mechanism, respectively.

To analyze the effectiveness of the CSPCL mechanism for detection results, we visualize the IoU scores and classification scores of prediction results of DINO on the PIXray dataset, as shown in Figure 5a. We draw the scatter plot of prediction results whose classification scores and IoU scores are higher than 0.3, along with the Kernel Density Estimation (KDE) curves. Blue and orange represent the results of DINO and DINO with the CSPCL mechanism, respectively. The orange points are more concentrated and distributed further to the right compared to the blue points, indicating that under the CSPCL mechanism, the content queries capture more classification semantic prior information without losing localization features, improving the accuracy of classification results. This demonstrates that our CSPCL mechanism can help the model more effectively perceive and extract foreground information of specific classes from the overlapping features in X-ray images.



(a) (b)

Figure 5: (a) The scatter plot and Kernel Density Estimation (KDE) joint distribution plot of prediction results of the final decoder layer. Blue and Orange represent the results of DINO and DINO with the CSPCL mechanism, respectively. (b) Visualization of deformable attention sampling points, reference points, and prediction results for corresponding group content queries in the last decoder layer. (i)–(iii) represent different groups. Each sampling point is shown as a filled circle, with color indicating its attention weight, and the reference point is marked by a green cross.

4 Experiments

4.1 Experimental setup

Datasets and evaluation metrics. We conduct extensive experiments on four large-scale publicly available X-ray prohibited item datasets, OPIXray [2], PIXray [15], CLCXray [6], and PIDray [1], which are widely used in the field of prohibited item detection. The PIXray, CLCXray, and PIDray datasets adopt the COCO [47] evaluation metric, AP, which measures the detector's average precision across various IoU thresholds, providing a comprehensive assessment of overall detection performance. The OPIXray dataset adopts the VOC [48] evaluation metric, mAP, which is calculated as the mean average precision across all categories at an IoU threshold of 0.5.

Implementation details. For fair, all training and testing are conducted on the same computer platform, equipped with an NVIDIA GeForce RTX 4090 GPU, an Intel Core i9-13900K CPU, 64GB of memory, running Windows 10, and using PyTorch 1.13.1. Furthermore, we utilize pre-trained ResNet-50 models from the official MMDetection [49] repository as the backbone, and all models are trained by the SGD optimizer with a learning rate of 0.01, a momentum of 0.9, and a weight decay of 0.1. In addition, all models are trained for 12 epochs, following the original training strategy, with an image size of 320×320 unless otherwise stated.

4.2 Generalization

Contrastive loss. To verify the generalization of the CSPCL mechanism for contrastive loss, we replace the proposed CSP loss with classic contrastive losses, such as N-pair loss and InfoNCE loss, to align content queries and prototypes of the same category, and evaluate them on DINO. As shown in Table 1, these losses still improve DINO's AP score on the PIXray dataset by 1.7% and 1.1%, respectively, but are less effective than the 3.2% improvement achieved by CSP loss. From these results, we draw two conclusions: (1) The strategy of aligning content queries with category prototypes in the CSPCL mechanism is generally effective for prohibited item detection in X-ray images; (2) Thanks to gradient truncation and the adaptive repulsion mechanism, CSP loss is better suited for the multi-class multi-sample aligning task compared to the aforementioned classic contrastive losses. More theoretical comparison analysis with other contrastive losses can be found in Appendix B.

Table 1: Comparing classic contrastive losses with our CSP loss under CSPCL mechanism.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
DINO	64.3	86.5	71.0	19.3	48.9	73.9
DNIO + N-pair	66.0 (+1.7)	88.1	72.2	23.7	51.3	75.4
DINO + InfoNCE	65.4 (+1.1)	87.4	72.3	22.1	51.1	74.6
DINO + CSP (our)	**67.5 (+3.2)**	**88.7**	**74.6**	**23.3**	**52.3**	**76.8**

Models and datasets. To verify the generalization of the CSPCL mechanism for models and datasets, we apply it to multiple representative Deformable DETR variants, including the basic Deformable DETR, the powerful DINO, RT-DETR, and the contraband-specialized AO-DETR on two prohibited

Table 2: Generalization analysis of CSPCL mechanism on PIXray and OPIXray.

Method	CSPCL	FPS	PARAMs	GFLOPs	PIXray			OPIXray					
					AP	AP_{50}	AP_{75}	mAP	FO	ST	SC	UT	MU
Deformable DETR	✗	60	52.14M	13.47	36.6	66.4	37.1	65.6	68.1	30.1	88.9	66.0	75.0
	✓	60	52.14M	13.47	41.8 (+5.2)	72.3	44.5	69.0 (+3.4)	78.4	38.8	86.0	62.8	79.1
RT-DETR	✗	64	42.81M	17.07	61.4	84.0	68.5	69.2	75.6	30.5	88.4	66.7	84.9
	✓	64	42.81M	17.07	61.8 (+0.4)	84.3	68.7	70.1 (+0.9)	76.0	34.4	88.6	67.4	84.3
DINO	✗	54	58.38M	26.89	64.3	86.5	71.0	77.0	81.7	56.3	90.0	71.4	85.5
	✓	54	58.38M	26.89	67.5 (+3.2)	88.7	74.6	77.9 (+0.9)	82.8	56.0	89.9	74.2	86.7
AO-DETR	✗	54	58.38M	26.89	65.6	86.1	72.0	77.7	82.8	57.6	89.8	71.2	86.9
	✓	54	58.38M	26.89	66.4 (+0.8)	86.8	73.6	78.5 (+0.8)	84.3	57.7	90.2	73.2	87.1

Table 3: Comparison with SOTA models on PIDray.

Method	Backbone	Image Size	AP				PARAMs	FLOPs
			easy	hard	hidden	overall		
Faster R-CNN [50]	ResNeXt-101-32x4d	1333×800	70.0	64.5	49.7	61.4	60.2M	300.5G
Mask R-CNN[51]	ResNeXt-101-32x4d	1333×800	77.7	70.0	53.7	67.0	101.9M	514.7G
FSAF [52]	ResNeXt-101-64x4d	1333×800	71.0	61.9	50.7	61.2	94.0M	461.6G
FCOS [53]	ResNeXt-101-64x4d	1333×800	72.9	63.4	51.1	62.5	89.6M	457.0G
ATSS [54]	ResNet-101	1333×800	71.0	64.5	55.4	63.6	51.1M	295.5G
VFNet [55]	ResNet-101	-	70.3	62.8	48.6	60.6	32.74M	-
TOOD [56]	ResNet-101	-	68.5	63.8	48.9	60.4	32.04M	-
UniFormer [57]	UniFormer-B	1333×800	70.4	63.9	40.7	58.3	69.0M	399.0G
Deformable DETR [16]	ResNet-50	1333×800	72.1	65.7	46.2	61.3	41.0M	210.2G
DN-DETR [23]	ResNet-50	1333×800	76.8	71.0	55.6	67.8	47.0M	860.0G
SDANet [1]	ResNet-101	500×500	71.2	64.2	49.5	61.6	-	-
ForkNet [39]	ResNet-101	500×500	75.0	66.9	58.6	66.8	-	-
AO-DETR [10]	Swin-L	512×512	81.8	72.6	55.7	70.0	229.0M	276.5G
$\mathcal{C}$-AO-DETR (ours)	Swin-L	512×512	**82.2**	**73.3**	**56.9**	**70.8**	229.0M	276.5G

item datasets PIXray and OPIXray. As shown in Table 2, the CSPCL mechanism improves the detection accuracy of Deformable DETR, RT-DETR, DINO, and AO-DETR on the PIXray dataset by 5.2%, 0.4%, 3.2%, and 0.8% AP, respectively. Notably, CSPCL does not impose additional burdens on model complexity such as GFLOPs and PARAMs, nor does it reduce the model inference speed (FPS). Similarly, CSPCL enhances the detection accuracy of the four models on the OPIXray dataset by 3.4%, 0.9%, 0.9%, and 0.8% mAP, respectively. These experimental results demonstrate that the CSPCL mechanism can effectively improve the Deformable DETR-based models' ability to handle overlapping detection tasks by aligning class prototypes and content queries.

Comparison with SOTA Models. To demonstrate the enhancement effect of CSPCL on the prohibited item detection model, we chose AO-DETR as the baseline and train it with CSPCL under the condition of 512×512 image size, 14 epochs, 240 queries to challenge other SOTA models, including general detectors and prohibited item detectors, on the largest known dataset for the prohibited item detection dataset PIDray [1]. As shown in Table 3, $\mathcal{C}$-AO-DETR (AO-DETR + CSPCL) obtains the highest AP over easy, hard, and hidden sets, while outperforms in terms of Image Size and FLOPs. The validation on the fine-grained dataset for cutters and liquid containers, CLCXray [6], is provided in Appendix A.

4.3 Ablation experiments

To obtain the optimal hyperparameters of CSP loss, we conduct extensive ablation experiments about integrating CSPCL into DINO on the PIXray dataset. As shown in Table 4a, we first independently insert ITA loss at the 0-th decoder layer to adjust the truncation factor γ, with the optimal truncation factor $\gamma^* = 5 \times 10^{-3}$, achieving 65.4% AP on the PIXray dataset. Similarly, as shown in Table 4b, the independent insertion of IAR loss yields the optimal temperature factor $\tau^* = 0.3$, achieving 65.6% AP. Furthermore, Table 4c shows that when both ITA loss and IAR loss are used simultaneously, i.e., the full version of CSP loss, the model's AP increases by 2.4%, which is higher than the individual use of ITA loss and IAR loss. This demonstrates the compatibility of the two loss functions. Subsequently, we adjust the target layer set where the CSPCL is inserted. As shown in Table 4d, the model achieves 67.2% AP by aligning the content queries of all decoder layers with the category prototypes better than aligning any single layer. Notably, when aligning only one layer, shallower layers, such as layer 0, tend to perform better. This could be because deeper layers, such as layer 5, involve content queries that more directly contribute to detection results, and indiscriminate alignment might disrupt the spatial features perceived by the model. Finally, as shown in Table 4e and Table 4f, we adjust the parameters α and β in Eq. (5) to control the contribution ratio of the two losses. The optimal values for α^* and β^* are 1 and 0.5, respectively. Overall, the CSPCL mechanism with the optimal hyperparameters can help DINO achieve the highest AP score of 67.5% on PIXray.

Table 4: Ablation study about the CSPCL mechanism on DINO [9] on the PIXray [15] dataset. L means all decoder layers. The superscript '$*$' represents the optimal value of the hyper-parameter.

γ	AP	AP_{50}	AP_{75}
5×10^{-2}	64.9 (+0.6)	86.1	72.6
$\mathbf{5\times10^{-3}}$	**65.4 (+1.1)**	**86.9**	**72.8**
5×10^{-4}	65.2 (+0.9)	86.6	72.6
5×10^{-5}	64.8 (+0.5)	86.3	72.4

(a) Ablation study of ITA.

τ	AP	AP_{50}	AP_{75}
0.1	64.6 (+0.3)	85.5	72.2
0.3	**65.6 (+1.3)**	**86.8**	**73.3**
1	65.3 (+1.0)	86.1	73.1
3	64.5 (+0.2)	86.1	71.7

(b) Ablation study of IAR.

ITA	IAR	AP	AP_{50}	AP_{75}
✗	✗	64.3	86.5	71.0
✓	✗	65.4 (+0.9)	86.9	72.8
✗	✓	65.6 (+1.1)	86.8	73.3
✓	✓	**66.7 (+2.4)**	**87.5**	**74.4**

(c) Ablation study of ITA and IAR.

$\hat{L}$	AP	AP_{50}	AP_{75}
0	66.7 (+2.4)	87.5	74.4
1	66.0 (+1.7)	87.0	73.5
5	60.6 (-3.7)	82.9	67.0
L	**67.2 (+2.9)**	**88.4**	**74.4**

(d) Ablation study of the target layer set.

α	AP	AP_{50}	AP_{75}
5	66.6 (+2.3)	88.1	73.7
1	**67.2 (+2.9)**	**88.4**	**74.4**
0.5	66.7 (+2.4)	87.5	74.4
0.1	65.1 (+0.8)	86.8	72.3

(e) $\beta = 1$, $\gamma^* = 5\times10^{-3}$, $\tau^* = 0.3$

β	AP	AP_{50}	AP_{75}
5	66.3 (+2.0)	87.6	73.3
1	67.2 (+2.9)	88.4	74.4
0.5	**67.5 (+3.2)**	**88.7**	**74.6**
0.1	66.7 (+2.4)	87.9	73.8

(f) $\alpha^* = 1$, $\gamma^* = 5\times10^{-3}$, $\tau^* = 0.3$

4.4 Visualization

To investigate the effect of the CSPCL mechanism on how content queries perceive and detect foreground prohibited items, we visualize the sampling points, reference points, and detection results in the last decoder layer for both DINO and "DINO+CSPCL" models, as shown in Fig. 5b. Specifically, columns (i) and (ii) demonstrate that, compared to the original DINO model, the sampling points responsible for feature extraction in the "DINO+CSPCL" are more concentrated on the contraband objects themselves, and the confidence and localization accuracy of the detection results are higher. In column (iii), under the condition of severe background overlap, DINO mistakenly identifies a "metal shaft" in the background as a "dart", while the "DINO+CSPCL" focuses on the less conspicuous "lighter". These results indicate that the CSPCL mechanism helps the model focus on foreground information and enhances its ability to the anti-overlapping detection task.

5 Conclusion

In this paper, for the prohibited item detection task based on X-ray images, we propose a plug-and-play CSPCL mechanism that aligns the classifier weights and content queries in Deformable DETR-based models. This mechanism supplements and clarifies the semantic information responsible for classification within the content queries, thereby improving the model's ability to perceive foreground information in overlapping features. Our proposed CSP loss outperforms classic contrastive losses and is better suited for the multi-class multi-sample alignment task. The CSPCL mechanism has strong generalizability, and it can enhance the anti-overlapping detection ability of various models on four datasets by classic contrastive losses, all without increasing model complexity.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant U22A2063, 62173083, 62276186, and 62206043; the China Postdoctoral Science Foundation under No.2023M730517 and 2024T170114; the Liaoning Provincial "Selecting the Best Candidates by Opening Competition Mechanism" Science and Technology Program under Grant 2023JH1/10400045; the Fundamental Research Funds for the Central Universities under Grant N2424022; the Major Program of National Natural Science Foundation of China (71790614) and the 111 Project (B16009); and the scholarship of China Scholarship Council (202506080096).

References

[1] B. Wang, L. Zhang, L. Wen, X. Liu, and Y. Wu, "Towards real-world prohibited item detection: A large-scale x-ray benchmark," in *International Conference on Computer Vision (ICCV)*, 2021, pp. 5412–5421.

[2] Y. Wei, R. Tao, Z. Wu, Y. Ma, L. Zhang, and X. Liu, "Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module," in *ACM International Conference on Multimedia (ACM MM)*, 2020, pp. 138–146.

[3] M. Li, T. Jia, H. Lu, Y. Li, B. Ma, H. Wang, and D. Chen, "Multi-scale dense detector for prohibited items detection in x-ray images," in *2024 7th International Symposium on Autonomous Systems (ISAS)*. IEEE, 2024, pp. 1–6.

[4] M. Li, B. Ma, H. Wang, D. Chen, and T. Jia, "Gadet: A geometry-aware x-ray prohibited items detector," *IEEE Sensors Journal (JSEN)*, vol. 24, no. 2, pp. 1665–1678, 2024.

[5] A. Chang, Y. Zhang, S. Zhang, L. Zhong, and L. Zhang, "Detecting prohibited objects with physical size constraint from cluttered x-ray baggage images," *Knowledge-Based Systems (KBS)*, vol. 237, p. 107916, 2022.

[6] C. Zhao, L. Zhu, S. Dou, W. Deng, and L. Wang, "Detecting overlapped objects in x-ray security imagery by a label-aware mechanism," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 17, pp. 998–1009, 2022.

[7] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 213–229.

[8] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "Dab-detr: Dynamic anchor boxes are better queries for detr," *arXiv preprint arXiv:2201.12329*, 2022.

[9] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," in *International Conference on Learning Representations (ICLR)*, 2023.

[10] M. Li, T. Jia, H. Wang, B. Ma, H. Lu, S. Lin, D. Cai, and D. Chen, "Ao-detr: Anti-overlapping detr for x-ray prohibited items detection," *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2024.

[11] M. Li, T. Jia, H. Lu, B. Ma, H. Wang, and D. Chen, "Mmcl: Boosting deformable detr-based detectors with multi-class min-margin contrastive learning for superior prohibited item detection," *arXiv preprint arXiv:2406.03176*, 2024.

[12] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 29, 2016.

[13] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[14] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," *arXiv preprint arXiv:2304.08069*, 2023.

[15] B. Ma, T. Jia, M. Su, X. Jia, D. Chen, and Y. Zhang, "Automated segmentation of prohibited items in x-ray baggage images using dense de-overlap attention snake," *IEEE Transactions on Multimedia (TMM)*, 2022.

[16] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable {detr}: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations (ICLR)*, 2021.

[17] S. Liu, T. Ren, J. Chen, Z. Zeng, H. Zhang, F. Li, H. Li, J. Huang, H. Su, J. Zhu *et al.*, "Detection transformer with stable matching," in *International Conference on Computer Vision (ICCV)*, 2023, pp. 6491–6500.

[18] Q. Chen, X. Chen, J. Wang, S. Zhang, K. Yao, H. Feng, J. Han, E. Ding, G. Zeng, and J. Wang, "Group detr: Fast detr training with group-wise one-to-many assignment," in *International Conference on Computer Vision (ICCV)*, 2023, pp. 6633–6642.

[19] D. Jia, Y. Yuan, H. He, X. Wu, H. Yu, W. Lin, L. Sun, C. Zhang, and H. Hu, "Detrs with hybrid matching," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19 702–19 712.

[20] Z. Zong, G. Song, and Y. Liu, "Detrs with collaborative hybrid assignments training," in *International Conference on Computer Vision (ICCV)*, 2023, pp. 6748–6758.

[21] D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, and J. Wang, "Conditional detr for fast training convergence," in *International Conference on Computer Vision (ICCV)*, 2021, pp. 3651–3660.

[22] Y. Wang, X. Zhang, T. Yang, and J. Sun, "Anchor detr: Query design for transformer-based detector," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 36, no. 3, 2022, pp. 2567–2575.

[23] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "Dn-detr: Accelerate detr training by introducing query denoising," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 619–13 627.

[24] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3733–3742.

[25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (ICML)*, 2020, pp. 1597–1607.

[26] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 35, no. 10, 2021, pp. 8547–8555.

[27] Y. Li, M. Yang, D. Peng, T. Li, J. Huang, and X. Peng, "Twin contrastive learning for online clustering," *International Journal of Computer Vision (IJCV)*, vol. 130, no. 9, pp. 2205–2221, 2022.

[28] K. Q. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 18, 2005.

[29] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters (SPL)*, vol. 25, no. 7, pp. 926–930, 2018.

[30] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6398–6407.

[31] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 18 661–18 673, 2020.

[32] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 539–546 vol. 1.

[33] J. Xie, J. Xiang, J. Chen, X. Hou, X. Zhao, and L. Shen, "C2am: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation," in *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022, pp. 989–998.

[34] C. Miao, L. Xie, F. Wan, C. Su, H. Liu, J. Jiao, and Q. Ye, "Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2119–2128.

[35] R. Tao, H. Wang, Y. Guo, H. Chen, L. Zhang, X. Liu, Y. Wei, and Y. Zhao, "Dual-view x-ray detection: Can ai detect prohibited items from dual-view x-ray images like humans?" in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025.

[36] B. Ma, T. Jia, M. Li, S. Wu, H. Wang, and D. Chen, "Towards dual-view x-ray baggage inspection: A large-scale benchmark and adaptive hierarchical cross refinement for prohibited item discovery," *IEEE Transactions on Information Forensics and Security (TIFS)*, pp. 1–1, 2024.

[37] R. Tao, T. Wang, Z. Wu, C. Liu, A. Liu, and X. Liu, "Few-shot x-ray prohibited item detection: A benchmark and weak-feature enhancement network," in *ACM International Conference on Multimedia (ACM MM)*, 2022, pp. 2012–2020.

[38] R. Tao, H. Li, T. Wang, Y. Wei, Y. Ding, B. Jin, H. Zhi, X. Liu, and A. Liu, "Exploring endogenous shift for cross-domain detection: A large-scale benchmark and perturbation suppression network," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 21 157–21 167.

[39] T. Jia, B. Ma, H. Wang, M. Li, S. Lin, and D. Chen, "Forknet: Overlapping image disentanglement for accurate prohibited item detection," *IEEE Transactions on Instrumentation and Measurement (TIM)*, vol. 73, pp. 1–12, 2024.

[40] R. Tao, Y. Wei, X. Jiang, H. Li, H. Qin, J. Wang, Y. Ma, L. Zhang, and X. Liu, "Towards real-world x-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 923–10 932.

[41] Z. Zhu, Y. Zhu, H. Wang, N. Wang, J. Ye, and X. Ling, "Fdtnet: Enhancing frequency-aware representation for prohibited object detection from x-ray images via dual-stream transformers," *Engineering Applications*

of Artificial Intelligence (EAAI), vol. 133, p. 108076, 2024.

[42] H. Liao, B. Huang, and H. Gao, "Feature-aware prohibited items detection for x-ray images," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 1040–1044.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[44] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[45] S. M. Kakade, K. Sridharan, and A. Tewari, "On the complexity of linear prediction: Risk bounds, margin bounds, and regularization," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 21, 2008.

[46] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research (JMLR)*, vol. 9, no. 86, pp. 2579–2605, 2008.

[47] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 740–755.

[48] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision (IJCV)*, vol. 88, pp. 303–338, 2010.

[49] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[50] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 39, no. 06, pp. 1137–1149, 2017.

[51] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *International Conference on Computer Vision (ICCV)*, 2017, pp. 2961–2969.

[52] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 840–849.

[53] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 9627–9636.

[54] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9759–9768.

[55] H. Zhang, Y. Wang, F. Dayoub, and N. Sunderhauf, "Varifocalnet: An iou-aware dense object detector," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8514–8523.

[56] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "Tood: Task-aligned one-stage object detection," in *International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2021, pp. 3490–3499.

[57] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "Uniformer: Unifying convolution and self-attention for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPMAI)*, vol. 45, no. 10, pp. 12 581–12 600, 2023.

[58] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[59] H. Zhang, H. Chang, B. Ma, N. Wang, and X. Chen, "Dynamic r-cnn: Towards high quality object detection via dynamic training," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 260–275.

[60] Y. Xu, Q. Zhang, Q. Su, and Y. Lu, "Pixdet: Prohibited item detection in x-ray image based on whole-process feature fusion and local–global semantic dependency interaction," *IEEE Transactions on Instrumentation and Measurement (TIM)*, vol. 72, pp. 1–17, 2023.

[61] W. Li, Z. Fan, J. Huo, and Y. Gao, "Modeling inter-class and intra-class constraints in novel class discovery," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 3449–3458.

[62] H. Chen, B. Lagadec, and F. Bremond, "Ice: Inter-instance contrastive encoding for unsupervised person re-identification," in *International Conference on Computer Vision (ICCV)*, 2021, pp. 14 960–14 969.

Appendix

A Comparison with SOTA Models of CLCXray

We further conduct experiments to verify the superiority of $\mathcal{C}$-AO-DETR on the fine-grained dataset for cutters and liquid containers, CLCXray [6]. As shown in Tab. 5, $\mathcal{C}$-AO-DETR obtains the highest AP of 63.1% among general objectors, including models like Cascade R-CNN, Dynamic R-CNN, and VarifocalNet, as well as prohibited item detectors, such as DOAM, LAreg, and ForkNet. It is worth mentioning that, although our model uses Swin-L as the backbone, resulting in a relatively high number of parameters, the smaller input image size of 512×512 reduces the model's computational complexity, making it slightly lower than the previously state-of-the-art dual-backbone prohibited item detector, ForkNet.

Table 5: Comparison with state-of-the-art general detectors on CLCXray.

Method	Backbone	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$	PARAMs	FLOPs
Cascade R-CNN [58]	ResNet-50	58.4	71.4	68.4	6.8	31.5	63.8	77.06M	1789.9G
Dynamic R-CNN [59]	ResNet-50	56.7	70.9	66.9	2.7	28.5	62.8	41.75M	176.1G
VarifocalNet [55]	ResNet-50	55.3	68.3	64.6	3.4	29.4	60.9	32.74M	159.7G
TOOD [56]	ResNet-50	57.9	71.1	66.5	6.7	34.8	63.4	32.01M	165.9G
DOAM [2]	ResNet-50	54.3	68.5	63.5	31.2	27.3	59.9	-	-
LAreg [6]	ResNet-50	58.5	70.9	67.7	12.6	30.5	63.8	-	-
LAcls [6]	ResNet-50	59.3	71.8	68.2	23.0	32.4	64.5	-	-
PIXDet [60]	CFB	60.4	72.4	69.4	25.3	31.7	66.0	-	-
ForkNet [39]	R50&R18	61.0	73.5	69.4	31.1	36.3	63.4	39.81M	224.6G
ForkNet [39]	R101&R18	62.0	74.0	71.4	35.6	**41.9**	67.7	55.95M	278.5G
$\mathcal{C}$-AO-DETR(ours)	Swin-L	**63.1**	**74.6**	**72.6**	**49.4**	36.7	**68.9**	229.0M	276.5G

B Comparison with other contrastive losses.

We take the last two representative contrastive losses as examples to analyze the advantage of our CSP loss.

IIC loss [61]. The inter-class Repulsion function of the IIC loss is based on Kullback-Leibler divergence, as follows:

$$L_{sKLD} = -\frac{1}{2}\left(D_{KL}(p_i^l \| p_j^u) + D_{KL}(p_j^u \| p_i^l)\right), \tag{9}$$

where p^l and p^u compose negative sample pairs, while i and j are the sample index number.

Weakness: This method is designed for comparing only two classes of samples. In the multi-class, multi-sample space, it applies the same strength of inter-class repulsion across all categories, and does not ensure that inter-class content queries of similar categories maintain sufficient differences to learn discriminative identifying features.

The intra-class Attraction function of IIC loss is as follows:

$$L = \frac{1}{N}\sum_{i=1}^{N} D_{KL}(p_l \| \hat{p_l}) + D_{KL}(\hat{p_l} \| p_l), \tag{10}$$

where $\hat{p_l}$ is the probability distribution of augmented data, which forms positive sample pairs with p_l. This constraint helps the model learn consistent feature representations of intra-class.

Weakness: It doesn't have the gradient truncation mechanism like our ITA loss (Eq. (6)) to prevent excessive homogenization of intra-class content queries.

ICE loss [62]: The intra-class Attraction function of the ICE loss:

$$L_{hins} = E\left[-\log\left(\frac{\exp\left(\langle f_i, m_k^i\rangle/\tau_{h_{ins}}\right)}{\sum_{j=1}^{J+1}\exp\left(\langle f_i, m_j\rangle/\tau_{h_{ins}}\right)}\right)\right], \tag{11}$$

where m_k^i represents the hardest positive instance for anchor f_i, and $\tau_{h_{ins}}$ is the temperature hyper-parameter. This method is essentially the InfoNCE loss [13] with a special construction method of positive sample pairs, and it only considers the furthest intra-class sample from the anchor/center.

Weakness: It will also lead the intra-class content queries to lose the necessary differences, as it lacks a gradient truncation mechanism like our ITA loss, which prevents intra-class homogenization.

The intra-class Attraction function of the ICE loss:

$$Ls_{ins} = D_{KL}(P \| Q), \tag{12}$$

where P and Q represent the distributions of inter-instance similarities before and after augmentation. It is based on KLD loss and its defects are the same as previously mentioned IIC loss.

Overall, compared with other existing contrastive losses, our CSP loss, through the intra-class gradient truncation and inter-class adaptive repulsion mechanisms, ensures that intra-class content queries do not become homogenized, preserving the necessary feature diversity within the class, while also ensuring that inter-class content queries align with the class prototypes' feature distribution, learning discriminative identifying features for similar categories.

C The significance of the alignment.

Aligning class prototypes with content queries essentially utilizes class prototypes to attract the content queries to prevent them from drifting into irrelevant regions in the feature space specific to the corresponding category, thereby correcting and supplementing the missing semantic information. Specifically, this alignment serves two purposes as follows:

1. Correcting the misguiding information caused by overlapping background features: After being trained on X-ray images with overlapping phenomena, the content queries can become ambiguous and misled by irrelevant background information. Aligning the content queries with class prototypes enables the queries to focus on the features specific to their corresponding categories, thereby correcting the noisy information in the queries caused by the overlapping background in the training phase.
2. Supplying the missing semantic information caused by the integration of positional encoding information in the decoder phase: As shown in Section 3.1, in the training phase, more and more positional encoding information is integrated into the content queries, which leads to the missing and catastrophic forgetting of the original class semantic information. The class prototypes are composed of classifier weights, which have learned the reliable inherent class semantic priors. Therefore, the alignment can ensure that the content queries do not deviate too much from the category semantic priors, supplying the missing and forgetting category semantic information.

As shown in Figure 4(c), the content queries of the original DINO are scattered randomly in the feature space and are far from the class prototypes (indicating low correlation between them). In contrast, as shown in Figure 4(d), after the alignment, the intra-class queries are clustered and distributed around their corresponding class prototypes (classifier weights). This also demonstrates that, through the alignment effect of CSPCL, the content queries are supplemented with effective category semantic information.

D Limitations

The CSPCL mechanism performs exceptionally well on X-ray image datasets with overlapping instances, typically helping to improve the model's AP (Average Precision) by over 1%. However, on natural light image datasets such as COCO [47], where there is no feature coupling phenomenon, CSPCL only contributes to an AP improvement of about 0.1%. We are currently delving into the reasons behind this and exploring ways to further refine the CSPCL strategy and CSP loss to extend its application to general object detection datasets.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have discussed them on page 1 and page 2.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed about that in the supplementary material.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We do not have theoretical contributions in this work, where our contributions are validated with experiments.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the hyper-parameters and network organizations are provided in the main paper and the appendix. Additionally, the code is included in the supplementary materials

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code is included in the supplementary materials, and it will be open sourced upon acceptance of the paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The experimental setup and implement details are provided in Sec. 4.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We follow existing related works for the setting of error bars.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide information on the compute memory and model efficiency in Sec. 4.1 and Tab. 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed that and claim we conform that Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: Our work is only for academic research purpose.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not have such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited them in the references.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We use and cite existing datasets in this work. Other assets including code/model will be released after submitting.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.