DisCo DETR: Distance-Aware Multi-View Contrastive Learning for DETR Pre-Training

Chao Ouyang[1], Yuyang Bai[1], Jun Zhang[1]*, Tianlu Gao[1], Jun Hao[2], Lijun Kong[3], David Wenzhong Gao[1]

[1]School of Electrical Engineering and Automation, Wuhan University
[2]China Datang Technology Innovation Co., Ltd.
[3]China Yangtze Power Co., Ltd.
jun.zhang.ee@whu.edu.cn

Abstract

Recent self-supervised pre-training methods for object detection often rely on generic object proposals for localization and semantic feature learning for classification, but they yield limited improvements when applied to Detection Transformers (DETR) due to a lack of architectural alignment. Hence, we propose an elegant and versatile self-supervised framework tailored for DETR-like models called **Dis**tance-aware Multi-view **Co**ntrastive Learning (**DisCo DETR**). **DisCo DETR** enhances localization and semantic features through two core components. (i) **Distance-aware Multi-view Object Query Fusion** explicitly guides object queries to focus on spatially close objects across views, stabilizing training and improving localization accuracy. (ii) **Contrastive Learning for DETR** uses native bipartite matching to identify positive output pairs across views and pull them closer, enhancing semantic features discrimination with no extra matching. DisCo DETR can be seamlessly integrated into DETR-like models and achieves SOTA transfer performance on PASCAL VOC and COCO benchmarks across multiple variants.

Code — https://github.com/oooyc/DisCo_DETR

Introduction

DETR (Carion et al. 2020) reformulates object detection as a set prediction task using a Transformer encoder-decoder, removing hand-crafted components like Non-maximum suppression and providing a clean detection pipeline. However, it converges slowly and heavily relies on large-scale labeled data, limiting scalability in real-world scenarios.

To reduce this dependency, recent self-/unsupervised approaches tackle two key challenges corresponding to detection's core tasks: (1) object localization, often via heuristic proposals (Dai et al. 2021; Chen et al. 2023b; Bar et al. 2022; Jin et al. 2023; Li et al. 2023) or attention cues (Melas-Kyriazi et al. 2022; Metaxas et al. 2024; Wang et al. 2023a, 2022a, 2023b; Siméoni et al. 2023); and (2) learning semantic image feature for classification, using contrastive learning (Bouniot et al. 2023; Jäckl et al. 2024; Kumar et al. 2024; Li et al. 2023; Wang et al. 2021; Xie et al. 2021; Jin et al. 2023; Grill et al. 2020), clustering (Caron et al. 2020; Metaxas et al. 2024), or feature reconstruction (Chen et al. 2023b; Dai et al. 2021). While effective for CNN-based detectors, these methods often yield limited gains on DETR due to the lack of architectural alignment.

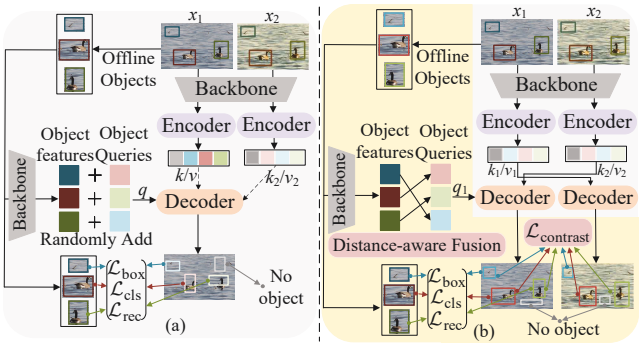


Figure 1: Comparison of self-supervised pre-training for DETR. (a) Previous methods. (b) DisCo DETR fuses cross-view features into spatially close queries and aligns matched outputs via contrastive learning.

Recent works attempt to address this by adapting self-supervised learning to DETR. Some methods rely on high-level supervision, such as DETReg (Bar et al. 2022) distilling knowledge from external modules, SeqCo-DETR (Jin et al. 2023) ensuring output sequence consistency, and AptDet (Metaxas et al. 2024) using self-training with pseudo-labels. These approaches, however, largely treat DETR as a black box. Others, like UP-DETR (Dai et al. 2021) and Siamese-DETR (Chen et al. 2023b), interact more directly with object queries but suffer from unstable supervision due to random injection of object features into queries. Consequently, existing methods (as shown in Figure 1 (a)) either overlook or fail to stably leverage DETR's core designs—learnable object queries and bipartite matching, showing limited fine-tuning improvements and highlighting the need for a more architecture-aligned solution.

In DETR, localization is driven by learnable object queries, which serve as positional embeddings that attend to spatially relevant regions via cross-attention and regress

*Corresponding Author. E-mail: jun.zhang.ee@whu.edu.cn.

offsets to the object. Inspired by this, we propose Distance-aware Multi-view Object Query Fusion (DMOQF) to improve query stability and localization. Given offline proposals (e.g., EdgeBoxes (Zitnick and Dollár 2014)) from different views, we use bipartite matching to find spatially close query-proposal pairs, then fuse proposal features into queries. Unlike previous methods (Chen et al. 2023b; Dai et al. 2021) that randomly inject features, our approach provides stable and distance-aware supervision. This explicitly guides queries to always focus on nearby objects, leading to smaller offsets and thus easier regression, which in turn stabilizes convergence and improves localization.

Beyond localization, we revisit DETR's bipartite matching for semantic feature learning. Prior methods often rely on coarse pseudo classification labels that merely indicate object presence, offering little semantic discrimination. In contrast, we observe that when applied independently to two augmented views, DETR's intrinsic bipartite matching naturally identifies queries matched to the same object. Unlike previous work (Jin et al. 2023; Li et al. 2023), which requires extra matching steps, this insight allows us to create positive samples for contrastive learning without any overhead. Our proposed Contrastive Learning on Decoder outputs (CLD) thus fully exploits DETR's intrinsic matching mechanism to inject semantic discrimination in a principled manner.

Our contributions can be summarized as:

- DMOQF: Fuses spatially close object features into queries, improving localization and stabilizing training.
- CLD: Leverages DETR's native bipartite matching to align semantic features across views, achieving stronger discrimination in a principled manner.

Moreover, DisCo DETR, as shown in Figure 1 (b), requires no architectural changes and can be easily integrated with various DETR variants, offering improved transferability and stronger detection performance.

Related Work

Unsupervised Pre-Training for Object Detectors

Recent unsupervised pre-training methods for object detection focus on generating object proposals and learning semantically discriminative features without manual annotations. For anchor generation, most approaches leverage self-supervised features (Metaxas et al. 2024; Wang et al. 2022a, 2023a,b; Siméoni et al. 2023; Melas-Kyriazi et al. 2022) (e.g., DINO (Caron et al. 2021)) to identify potential object regions. For instance, TokenCut (Wang et al. 2023b) segments objects by constructing a graph of image patches with self-supervised Transformer features and applying normalized cuts to cluster semantically coherent regions, while AptDet (Metaxas et al. 2024) generates region proposals through spectral clustering and K-means on feature maps from a self-supervised backbone. Other methods use heuristic anchors including (Dai et al. 2021; Bar et al. 2022; Chen et al. 2023b; Jin et al. 2023; Li et al. 2023)

To learn semantically discriminative features for classification, some works use contrastive learning, which often involves feature-matching computation (Wang et al. 2021; Xie et al. 2021; Li et al. 2023; Jin et al. 2023), aggressive color & cropping augmentations (Jäckl et al. 2024; Kumar et al. 2024), and other complex and computationally costly mechanisms (Bouniot et al. 2023). While AptDet (Metaxas et al. 2024) used cluster methods to group proposals into semantic categories to generate discriminative pseudo-labels. Other works (Dai et al. 2021; Chen et al. 2023b), focusing on DETR-like models, used feature reconstruction to preserve semantic features from the frozen pre-trained backbone.

Unsupervised Pre-Training for DETR

As initial DETR encountered challenges such as slow convergence and high computational costs, recent improvements to DETR have focused on three main areas: (1) reducing computational costs through deformable attention mechanisms (Zhu et al. 2020), hierarchical salience filtering (Hou et al. 2024), hybrid encoders like (Zhao et al. 2024b) and rethinking multi-scale features (Lin et al. 2023), (2) treating queries as anchors with positional priors to speed up convergence such as (Wang et al. 2022b; Meng et al. 2021; Liu et al. 2022; Zhang et al. 2022; Lin et al. 2023), and (3) implementing auxiliary training methods like denoising (Li et al. 2022; Zhang et al. 2022) and one-to-many label assignment (Chen et al. 2023a; Zong, Song, and Liu 2023; Jia et al. 2023; Zhao et al. 2024a; Huang et al. 2024) to enhance supervision for accelerating DETR's convergence.

Despite these improvements, DETR and its variants still require extensive training data. To address this, researchers have explored various self-supervised pre-training methods. UP-DETR (Dai et al. 2021) randomly cropped patches from each image, extracted patch features using a frozen backbone, and randomly added them to object queries, thus guiding DETR to detect areas corresponding to the patch features to accelerate convergence. Besides, UP-DETR also reconstructed patch features with outputs to inherit the frozen backbone's discriminative features of high-level semantic information. DETReg (Bar et al. 2022) subsequently uses Selective Search (Uijlings et al. 2013) to generate object proposals as annotations but overlooks classification without feature reconstruction. Siamese DETR (Chen et al. 2023b) used EdgeBox (Zitnick and Dollár 2014) to generate proposals and introduced multi-view cross-attention, which adds cross-view object features to object queries for learning view-invariant local features with cross-view feature reconstruction. SeqCo DETR (Jin et al. 2023) proposed sequence consistency as a pretext task, which could also provide discriminative features, and combined with a masking strategy.

Method

DETR Review

Given an input image $\boldsymbol{x} \in \mathbb{R}^{3\times H_0\times W_0}$, DETR first uses the backbone to encode the input image $h = \mathrm{Backbone}(x)$, then refines image features $h \in \mathbb{R}^{C\times H_1\times W_1}$ to get the global context $c \in \mathbb{R}^{C\times H_1\times W_1}$ via a Transformer encoder $c = \mathrm{Encoder}(h, \phi_p)$ with positional embedding $\phi_p \in \mathbb{R}^{C\times H_1\times W_1}$. The decoder is composed of self-attention layers and cross-attention layers. Firstly, N randomly initialized object queries $\phi_q \in \mathbb{R}^{N\times C}$ pass through the self-attention

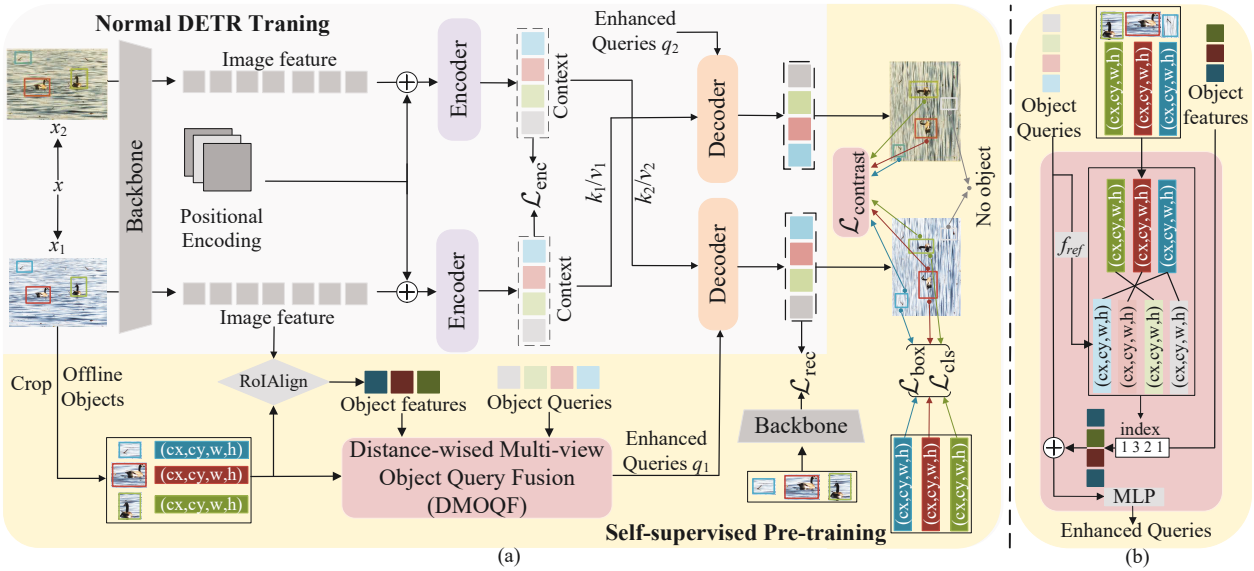


Figure 2: (a) shows the overall framework of DisCo DETR. (b) illustrates the DMOQF module, which fuses object queries with spatially close object features. CLD enforces consistency between the cross-view decoder outputs matched to the same objects by contrast loss ($\mathcal{L}_{contrast}$). Other losses ($\mathcal{L}_{rec}, \mathcal{L}_{loc}, \mathcal{L}_{enc}$) are the same as previous work (Chen et al. 2023b).

layers and a projection to obtain $q = f_q(\phi_q)$. The global context c serves as the k and v after projection $k = f_k(c)$ and $v = f_v(c)$ in the cross-attention layer where q will search similar k to obtain a weighted sum of v as $\hat{q} \in \mathbb{R}^{N\times C}$, which is formulated as below:

$$\hat{q} = \text{CrossAttn}(q, k, v) = \sum \text{Softmax}(\frac{qk^T}{\sqrt{d_k}})v \quad (1)$$

Then, two prediction heads are applied to the weighted sum $\hat{q}$. Firstly, $f_{box} : \mathbb{R}^C \to \mathbb{R}^4$, predicts the bounding boxes. Secondly, $f_{cat} : \mathbb{R}^C \to \mathbb{R}^L$, outputs the classification.

Distance-Aware Multi-View Object Query Fusion

Preparation Following Siamese DETR (Chen et al. 2023b), we generate paired views $\{x_1, x_2\}$ with corresponding n (usually $n = 10$) object proposals $\{b_1, b_2\}$ that represent the same objects. We then extract two distinct sets of features. Object features $\{z_1, z_2\}$ are obtained by applying RoIAlign (He et al. 2017) on the feature maps produced by the backbone from the full images. In parallel, crop features $\{p_1, p_2\}$ are generated by cropping patches from the original images $\{x_1, x_2\}$ and then feeding these patches through the same backbone. Finally, both feature sets are globally average-pooled to a dimension of $\mathbb{R}^{n\times C}$.

Query-Proposal Matching In contrast to prior works (Dai et al. 2021; Chen et al. 2023b) that use unstable random feature injection, our method provides stable guidance via a distance-aware, two-stage bipartite matching process. As illustrated in Figure 2, this mechanism explicitly pairs each query with a spatially close object proposal, ensuring consistent supervision based on spatial proximity.

1. Initial Matching: We first derive reference points $\phi_r \in \mathbb{R}^{N\times 2}$ from the object queries ϕ_q like (Zhu et al. 2020), representing their initial spatial positions in the image. Then, for each of the two augmented views, we use the Hungarian algorithm (Kuhn 1955) to find the global optimal bipartite matching between ϕ_r and centers of the n object proposals $\{b'_1, b'_2\} \in \mathbb{R}^{n\times 2}$. This matching minimizes the total ℓ_1 distance, pairing each proposal with its spatially close query.

$$\hat{\sigma'_1} = \arg\min_{\sigma'_1 \in \mathcal{G}_n} \sum_{i=1}^{n} |\phi_{r,\sigma'_1(i)} - b'_{1,i}|_1$$
$$\hat{\sigma'_2} = \arg\min_{\sigma'_2 \in \mathcal{G}_n} \sum_{i=1}^{n} |\phi_{r,\sigma'_2(i)} - b'_{2,i}|_1 \quad (2)$$

where σ'_1 is a one-to-one map from the n proposals to a subset of the N queries, and $\sigma'_1(i)$ is the index of the query assigned to the i-th proposal in view x_1. $\mathcal{G}_n$ represents the set of all such possible one-to-one maps.

2. Full Matching: To assign the remaining $N-n$ queries, we first form a new target set of $N - n$ centers by replicating each of the original n proposal centers $\frac{N-n}{n}$ times. The same bipartite matching process as the initial step is then applied to the unassigned queries and this replicated set. This process, combined with the initial assignments, yields the final many-to-one maps $\{\hat{\sigma_1}, \hat{\sigma_2}\}$ from all N queries to the original n proposals, ensuring guidance for every query.

Distance-Aware Query Enhancement With the stable query-proposal pairings established, we guide DETR's learning process. For a given view x_1, we enhance its queries using features from the other view x_2:

$$q_2 = f_q(\text{MLP}(z_2 + \phi_{q,\hat{\sigma_1}})) \quad (3)$$

where z_2 represents the object features from view x_2, and $\phi_{q,\hat{\sigma_1}}$ are queries matched to the same objects in view x_1.

These enhanced queries q_2 are then used to attend to the global context c_1 of the view x_1:

$$\hat{q}_1 = \text{CrossAttn}(q_2, f_k(c_1), f_v(c_1)) \quad (4)$$

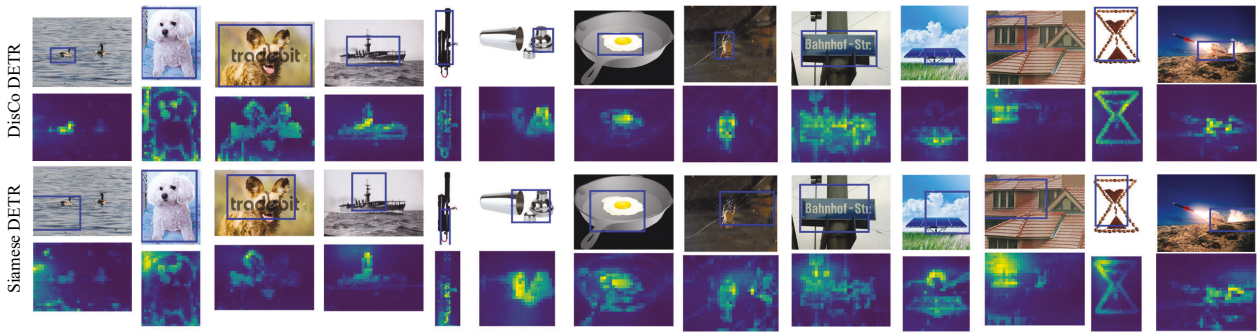


Figure 3: Visualization of detection results and attention maps from Conditional DETR on mini-ImageNet. The models were pre-trained using DisCo DETR and Siamese DETR under identical settings, and are shown here without any fine-tuning.

By consistently fusing features from spatially close object proposals, DMOQF guides queries to detect nearby objects, resulting in stable convergence. This could also simplify the regression task, as queries only need to predict small offsets to the objects, leading to improved localization accuracy.

Contrastive Learning for DETR

Beyond improved localization, our CLD module enhances the semantic discrimination of image features. It achieves this by fully exploiting DETR's intrinsic bipartite matching mechanism to generate positive pairs for contrastive learning, adding supervision without extra matching steps.

Positive Pair Mining Our key insight is that DETR's inference process naturally identifies positive pairs. For two views $\{x_1, x_2\}$, we perform the forward pass and bipartite matching for each view independently. This native matching process assigns output embeddings $\{\hat{q_1}, \hat{q_2}\}$ to proposals $\{b_1, b_2\}$. Crucially, any two outputs from different views that are matched to the same object in each view inherently form a positive pair. This approach repurposes DETR's native bipartite matching for pair mining, avoiding extra computation overhead like prior works (Jin et al. 2023; Li et al. 2023).

Contrastive Objective Once positive pairs are identified, we enforce their consistency using a contrastive objective inspired by SimSiam (Chen et al. 2023c). This involves processing the features with a projection head f_{proj} (3-layer MLP), followed by a prediction head f_{pred} (2-layer MLP). For a positive pair $\{\hat{q_1}, \hat{q_2}\}$, the process is:

$$t_1 = f_{\text{proj}}(\hat{q_1}) \qquad t_2 = f_{\text{proj}}(\hat{q_2}) \tag{5}$$

$$\hat{t_2} = f_{\text{pred}}(t_1) \qquad \hat{t_1} = f_{\text{pred}}(t_2) \tag{6}$$

The contrastive loss $\mathcal{L}_{\text{contrast}}$ then maximizes the agreement between the projection from one view and the prediction from the other, using a negative cosine similarity loss and a stop-gradient operation:

$$\mathcal{L}_{\text{contrast}} = \frac{1}{2}\mathcal{D}\big(\text{stop\_grad}(t_1), \hat{t_2}\big) + \frac{1}{2}\mathcal{D}\big(\text{stop\_grad}(t_2), \hat{t_1}\big) \tag{7}$$

Here, $\mathcal{D}(a, b)$ is the negative cosine similarity. By pulling features of the same object from different views closer in the embedding space, CLD effectively teaches the model to learn more semantically discriminative features.

Loss Function

Feature Reconstruction Loss Following (Dai et al. 2021), we also use the feature reconstruction loss $\mathcal{L}_{rec}$ to guide local feature refinement of the DETR decoder with the Frozen backbone's semantically discriminative features after bipartite matching as below:

$$\mathcal{L}_{rec} = \mathcal{D}(f_{rec}(\hat{q_2}), p_1) + \mathcal{D}(f_{rec}(\hat{q_1}), p_2) \tag{8}$$

where f_{rec} denotes for a MLP.

Localization Loss and Global Discrimination Loss Firstly, we use a prediction head f_{box} for box prediction:

$$\hat{b_1} = f_{box}(\hat{q_1}) \quad \hat{b_2} = f_{box}(\hat{q_2}) \tag{9}$$

After bipartite matching, calculate the multi-view symmetrical localization loss as:

$$\mathcal{L}_{loc} = l_{box}(\hat{b_2}, b_2) + l_{box}(\hat{b_1}, b_1) \tag{10}$$

where l_{box} is a combination of generalized IoU loss and l_1 loss as (Carion et al. 2020). Following (Chen et al. 2023b), we use the global discrimination loss $\mathcal{L}_{enc}$ symmetrically to improve the global feature discrimination of the encoder as:

$$\begin{aligned}\mathcal{L}_{enc} = \mathcal{D}[\text{MLP}(c_1), \text{detach}(c_2)] + \\ \mathcal{D}[\text{MLP}(c_2), \text{detach}(c_1)]\end{aligned} \tag{11}$$

where MLP consists of three layers (FC-BN-ReLU).

Overall Loss We define the overall loss function for DisCo-DETR as below:

$$\mathcal{L} = \lambda_0\mathcal{L}_{rec} + \lambda_1\mathcal{L}_{enc} + \lambda_2\mathcal{L}_{loc} + \lambda_3\mathcal{L}_{contrast} \tag{12}$$

where $\lambda_{0/1/2/3}$ are the loss weighting hyper-parameters.

Experiment

Implementation

Framework and DETR Variants DisCo DETR is a self-supervised pre-training framework that seamlessly integrates with DETR-like models. We validate its effectiveness on multiple DETR variants (using 300 queries unless specified) (Zhu et al. 2020; Meng et al. 2021; Zhao et al. 2024b; Liu et al. 2022; Zhang et al. 2022) compared to Siamese DETER (the past SOTA), DETReg, and baseline model without pre-training (denoted as $from\ scratch$).

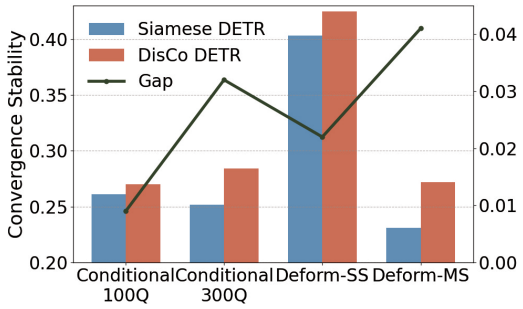
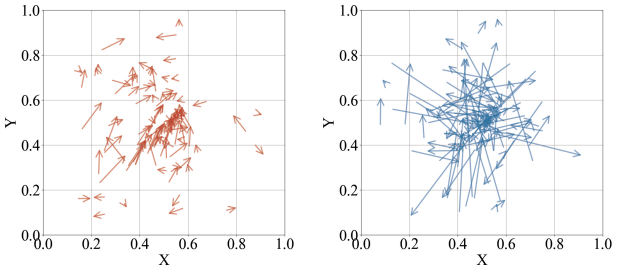


Figure 4: Comparison of ***CS*** between DETR variants pre-trained by DisCoDETR and SiameseDETR. Deform MS and SS stand for Deformable-DETR multi-scale and single-scale. 100Q and 300Q stand for 100 and 300 queries.



(a) DisCo DETR. (b) Siamese DETR.

Figure 5: We visualize arrows from object queries to matched objects of Deformable DETR in the last epoch of pre-training under identical settings.

Datasets Models are pre-trained on mini-ImageNet (Vinyals et al. 2016) and fine-tuned on PASCAL VOC `trainval07+12` (Everingham et al. 2010) and COCO `train2017` (Lin et al. 2014). Evaluation follows COCO-style metrics.

Pre-Training & Fine-Tuning Our pre-training stage uses EdgeBoxes (2014) for object proposals and mainly adopts a $40/60$ schedule (learning rate decay at 40 epochs, total 60 epochs) unless specified. For the backbone, we use ResNet50 initialized with SwAV and SSLD (Cui et al. 2021) for the RT-DETR model. For the subsequent fine-tuning, schedules vary by model (usually a 40/50 schedule unless specified). A fixed batch size of 64 is used across fine-tuning.

Visualization and Analysis

Detection Results with Attention Map Visualization of detection results (Figure 3) demonstrates that DisCo DETR indeed improves the localization ability of Conditional DETR. Moreover, DisCo DETR's attention maps are much cleaner and clearer, indicating that each object query does its job and knows whom it needs to detect based on its more precise positional information than Siamese DETR.

DMOQF for Stable Convergence Let $\phi_r^n(t) \in [0,1]^2$ denote the normalized reference point of the n-th object query

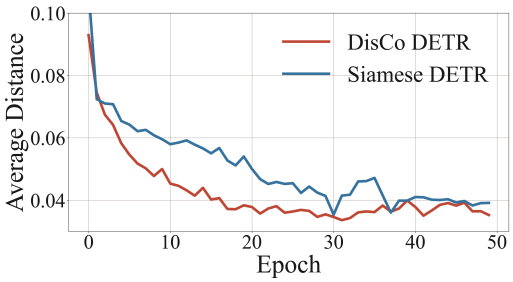


Figure 6: Comparison of the average distance RT-DETRs pre-trained by DisCoDETR and SiameseDETR under identical settings when fine-tuning on PASCAL VOC.

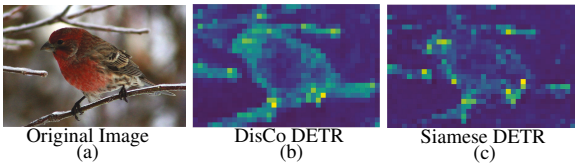


Figure 7: Attention heat map of encoder outputs.

at epoch t (I epochs in total). We define:

$$d_n^i = \|\phi_r^n(i+1) - \phi_r^n(i)\|_2, \quad S_n = \|\phi_r^n(I) - \phi_r^n(0)\|_2$$

Then introduce a metric to quantify **C**onvergence **S**tability:

$$\boldsymbol{CS} = \frac{S_n}{\sum_i^{I-1} d_n^i} \tag{13}$$

A lower ***CS*** indicates greater variability in queries' convergence direction, while a higher ***CS*** suggests more stable convergence. Figure 4 illustrates that DisCo DETR exhibits a higher average ***CS***, indicating more stable convergence.

DMOQF for Detecting Objects with Spatially Close Queries We first visualize arrows from queries to objects in pre-training. As shown in Figure 5, clearly Siamese DETR exhibits longer and less ordered arrows, while DisCo DETR shows shorter and more organized arrows, indicating that our approach assigns objects to nearby queries for easier box regression. In fine-tuning (Figure 6), RT-DETR pre-trained by DisCo DETR achieves a lower average distance between queries and their matched objects, showing that the model has learned to detect objects with nearby queries without injecting object features after pre-training.

CLD for Discriminative Semantic Features We demonstrate the contribution of the CLD module to learning discriminative semantic features by visualizing encoder attention maps in Figure 7. To isolate its impact, the model used in this analysis was pre-trained without the DMOQF module. For each image feature, we average its attention score matrix over the entire image to generate a heatmap. The visualization reveals that, compared to Siamese DETR, our CLD-only model produces attention maps with much clearer object boundaries under identical settings. This indicates a stronger ability to discriminate features between different objects and validates that our CLD module effectively encourages the learning of discriminative representations.

method	DETR	Pretrain Dataset	Backbone	Schedule	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
$from\ scratch$	DAB DETR	-	SwAV 50	20/24	29.7	50.2	29.5	12.4	33.8	43.9
DETReg	DAB DETR	miniImageNet	SwAV 50	20/24	29.8	50.2	29.8	12.1	34.5	45.1
Siamese DETR	DAB DETR	miniImageNet	SwAV 50	20/24	30.5	52.4	31.9	13.1	36.1	48.0
DisCo DETR	DAB DETR	miniImageNet	SwAV 50	20/24	**33.9**	**54.5**	**34.5**	**15.4**	**37.8**	**49.5**
$from\ scratch$	DAB DETR	-	SwAV 50	40/50	36.4	57.6	38.6	17.2	40.4	52.7
DETReg	DAB DETR	miniImageNet	SwAV 50	40/50	36.0	56.7	37.9	16.8	39.8	52.0
Siamese DETR	DAB DETR	miniImageNet	SwAV 50	40/50	38.0	58.4	39.7	17.4	42.1	55.3
DisCo DETR	DAB DETR	miniImageNet	SwAV 50	40/50	**39.7**	**59.4**	**41.1**	**19.0**	**42.7**	**56.3**
$from\ scratch$	conditional	-	SwAV R50	40/50	39.2	60.1	41.6	19.7	42.8	55.5
Siamese DETR	conditional	miniImageNet	SwAV R50	40/50	40.6	61.3	43.1	20.4	44.5	57.7
DisCo DETR	conditional	miniImageNet	SwAV R50	40/50	**41.2**	**61.9**	**44.9**	**22.2**	**45.0**	**59.0**
$from\ scratch$	RT-DETR	-	SSLD R50vd	20/24	43.2	60.8	47.2	23.5	47.5	62.1
DETReg	RT-DETR	miniImageNet	SSLD R50vd	20/24	40.5	57.7	44.1	21.0	44.1	58.9
Siamese DETR	RT-DETR	miniImageNet	SSLD R50vd	20/24	44.1	61.6	47.5	25.2	48.3	62.9
DisCo DETR	RT-DETR	miniImageNet	SSLD R50vd	20/24	**46.7**	**63.9**	**50.1**	**27.4**	**50.5**	**65.1**
$from\ scratch$	RT-DETR	-	SSLD R50vd	40/50	49.2	67.4	52.9	29.6	53.4	67.8
DETReg	RT-DETR	miniImageNet	SSLD R50vd	40/50	47.5	63.4	50.2	27.6	51.1	65.4
Siamese DETR	RT-DETR	miniImageNet	SSLD R50vd	40/50	48.2	66.2	51.9	28.6	52.3	66.6
DisCo DETR	RT-DETR	miniImageNet	SSLD R50vd	40/50	**49.8**	**68.1**	**53.2**	**30.1**	**53.4**	**68.1**
$from\ scratch$	DINO	-	SSLD R50vd	20/24	48.4	65.7	52.9	30.6	52.4	62.7
Siamese DETR	DINO	miniImageNet	SSLD R50vd	20/24	48.1	65.1	52.7	30.2	51.6	62.8
DisCo DETR	DINO	miniImageNet	SSLD R50vd	20/24	**49.0**	**66.0**	**53.8**	**31.3**	**52.5**	**63.3**

Table 1: Comparisons of DisCo DETR, $from\ scratch$ and SiameseDETR on the COCO detection benchmark. The results of all models are achieved by officially released repositories and pre-trained models.

method	DETR	Schedule	AP	AP_{50}	AP_{75}
$from\ scratch$	DAB DETR	40/50	41.1	70.0	43.3
DETReg	DAB DETR	40/50	47.3	73.2	50.4
Siamese DETR	DAB DETR	40/50	48.7	75.1	53.2
DisCo DETR	DAB DETR	40/50	**50.2**	**76.6**	**54.5**
$from\ scratch$	conditional	40/50	44.3	70.4	47.3
Siamese DETR	conditional	40/50	50.3	74.3	55.0
DisCo DETR	conditional	40/50	**51.7**	**75.9**	**56.2**
$from\ scratch$	Deform-SS	40/50	38.5	65.8	40.3
Siamese DETR	Deform-SS	40/50	35.9	60.5	37.8
DisCo DETR	Deform-SS	40/50	**36.6**	**60.9**	**39.0**
$from\ scratch$	RT-DETR	40/50	59.8	80.4	65.1
DETReg	RT-DETR	40/50	59.0	79.1	63.4
Siamese DETR	RT-DETR	40/50	60.4	81.3	65.9
DisCo DETR	RT-DETR	40/50	**61.6**	**82.2**	**67.2**
$from\ scratch$	DINO	30/36	59.5	81.1	64.8
Siamese DETR	DINO	30/36	60.4	81.0	66.1
DisCo DETR	DINO	30/36	**61.7**	**82.2**	**67.7**

Table 2: Comparisons on the PASCAL VOC benchmark.

Main Results

COCO Detection As shown in Table 1, we achieve state-of-the-art performance on the challenging COCO benchmark, consistently outperforming $from\ scratch$, DETReg, and Siamese DETR across multiple DETR variants and training schedules. The results highlight two key strengths of DisCo DETR. Firstly, it accelerates convergence, a benefit derived from higher convergence stability provided by our DMOQF module. This is evidenced by the larger performance gains in shorter training schedules. For instance, compared to Siamese DETR, the AP uplift for DAB-DETR is +3.4 points in a 24-epoch schedule (vs. +1.7 in a 50-epoch one). This trend of accelerated convergence is even more pronounced when compared against DETReg. On the RT-DETR architecture, our method's advantage over DETReg is a massive +6.2 AP in the short 24-epoch schedule (vs. +2.3 in a 50-epoch one). This clearly demonstrates our method's superior ability to bootstrap performance and converge faster. Secondly, it significantly enhances localization ability, and this advantage remains robust even with longer training. This improvement is particularly notable not only for the demanding AP_{75} metric but also across different object scales (AP_s and AP_l). Focusing specifically on the 50-epoch schedule, the improvements on AP_{75} are substantial. Our method consistently outperforms Siamese DETR by +1.3 to +1.8 points, and the advantage over DETReg is even more significant, with gains of over +3.0 points. These targeted improvements in both training efficiency and localization precision validate our core design of encouraging models to detect objects via spatially close queries.

PASCAL VOC Detection Our method's effectiveness is further demonstrated on the PASCAL VOC benchmark. The results in Table 2 showcase the comprehensive advantages of DisCo DETR across five different DETR architectures, where it consistently surpasses all baselines: training $from\ scratch$, DETReg, and Siamese DETR. Notably, DisCo DETR proves more robust and effective than prior pre-training work. For instance, on the advanced RT-DETR, DETReg fails to match the $from\ scratch$ performance (59.0 vs 59.8 AP), while our method provides a significant boost to 61.6 AP. Even on the strong DINO baseline, our method improves performance from 60.4 AP to 61.7 AP. The key to this success is improved localization, confirmed by the demanding AP_{75} metric, where DisCo DETR con-

Method	DETR	DMOQF	CLD	λ_3	AP
Siamese DETR	conditional	×	×	-	50.3
ours(a)	conditional	✓	×	-	51.4
ours(b)	conditional	×	✓	2	51.14
ours(c)	conditional	✓	✓	2	**51.7**
Siamese DETR	Defom-MS	×	×	-	48.7
ours(a)	Defom-MS	✓	×	1	49.8
ours(b)	Defom-MS	×	✓	2	49.6
ours(c)	Defom-MS	✓	✓	5	**50.2**

Table 3: Ablations on DMOQF, CLD and its loss weight λ_3.

Method	f_{proj}	f_{pred}	DMOQF	AP	AP_{50}	AP_{75}
Siamese DETR	×	×	×	46.9	72.7	50.4
DisCo DETR (f)	×	×	✓	43.9	70.9	46.8
DisCo DETR (e)	✓	×	✓	47.5	73.1	51.0
DisCo DETR (d)	×	✓	✓	46.8	73.0	50.6
DisCo DETR	✓	✓	✓	**47.8**	**73.7**	**51.2**

Table 4: More ablations about f_{pred} and f_{proj} of CLD using Conditional DETR with 100 queries.

sistently achieves the largest gains over the strong Siamese DETR baseline (improving by +1.2 to +1.6 points). This fundamental boost in localization, in turn, elevates the overall detection performance, with standard AP increasing by up to +1.5 points (e.g., 50.2 vs. 48.7 on DAB DETR). These combined improvements directly validate our core design principle: pre-train the model to utilize spatially close object queries to detect objects.

Ablations

We conduct comprehensive ablation studies to validate our design choices. All experiments are performed on PASCAL VOC with Conditional DETR, unless otherwise specified.

Effectiveness of Core Components As shown in Table 3, we first validate the individual and combined contributions of our proposed modules: DMOQF and CLD.

- **Individual Gains:** Applied alone, DMOQF (ours(a)) consistently improves AP by +1.1 over the Siamese DETR on both Conditional DETR and Deformable DETR (Defom-MS). Similarly, CLD alone (ours(b)) provides significant gains of +0.8 to +0.9. This confirms the standalone effectiveness of both components.
- **Synergistic Effect:** Combining both modules (ours(c)) yields the best performance on both architectures, achieving 51.7 AP and 50.2 AP, respectively. This demonstrates that DMOQF and CLD are complementary.
- **Loss Weight λ_3:** We analyzed the sensitivity of the CLD loss weight λ_3 and found a performance peak at $\lambda_3 = 2$, with stable results for nearby values. This indicates the robustness of our design, and we recommend setting $\lambda_3 = 2$ as a default, though our method is generally robust to variations in this hyperparameter.

Dissection of the CLD To further analyze CLD's design, we evaluate variants with/without projection head f_{proj} and prediction head f_{pred}. Table 4 reveals that removing both

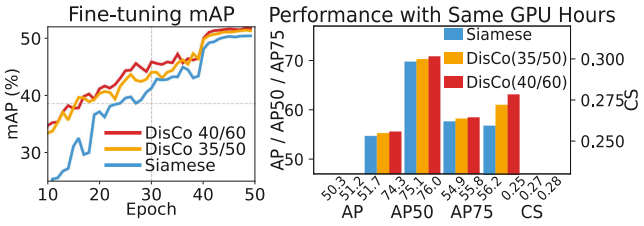


Figure 8: Comparison of performance under the Same GPU hours. Conditional DETR is used for fine-tuning.

components (denoted as DisCo DETR(f)) causes a significant performance drop (-3.0 AP vs. Siamese DETR), which might be attributable to conflicting optimization objectives when directly aligning raw cross-view positional information and local image features. Introducing f_{proj} (denoted as DisCo DETR(e)) recovers 3.6 AP by aligning after projection to shared embedding space, while f_{pred} (DisCo DETR(d)) adds 2.9 AP through predictive embedding alignment in shared space. The full model (with both f_{proj} and f_{pred} achieves 47.8 AP, with AP_{75} improving by 4.4 points (51.2 vs. 46.8)), demonstrating the necessity of shared embedding space and predictive embedding alignment.

Computation Overhead Our method introduces a minor overhead from CLD's MLP heads (adding 725K parameters, ∼1.8% of total) and DMOQF's matching, resulting in a ∼0.16% increase in per-epoch pre-training time. To further validate that this cost is worthwhile, we compare models pre-trained under the same-GPU-hour setting. As shown in Figure 8, the model pre-trained by our method converges faster and achieves higher mAP (51.2 vs. 50.3 AP) in fine-tuning. Achieving a better result in fewer pre-training epochs demonstrates that the per-epoch benefit gained from DisCo DETR is substantially greater than that of Siamese DETR (under 40/60 schedule). Therefore, the minor computational overhead per epoch in pre-training is a worthwhile investment for achieving significant improvements in both fine-tuning efficiency and final performance.

Conclusion

Without changing the intrinsic structure of DETR, we introduce a novel self-supervised pre-training method to enhance DETR's both localization ability and semantic understanding. We propose **Distance-aware Multi-view Object Query Fusion**, which stabilizes the convergence of object queries and teaches DETR to detect objects with spatially close queries. We also introduce **Contrastive Learning for DETR**, which provides additional supervision for semantically discriminative features. DisCo DETR achieves better transfer performance with various DETR variants on COCO and PASCAL VOC benchmarks compared to previous work, providing an easy and extensive method to enhance convergence speed and performance for DETR variants in fine-tuning. Admittedly, our method is limited by its reliance on the pre-training backbone without a uniform pre-training paradigm. Future work will explore more uniform, efficient, and effective self-supervised pre-training for DETR.

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.

Bouniot, Q.; Audigier, R.; Loesch, A.; and Habrard, A. 2023. Proposal-contrastive pretraining for object detection from fewer data. arXiv:2310.16835.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *Computer Vision – ECCV 2020*, 213–229. Cham: Springer International Publishing.

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.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, Q.; Chen, X.; Wang, J.; Zhang, S.; Yao, K.; Feng, H.; Han, J.; Ding, E.; Zeng, G.; and Wang, J. 2023a. 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.

Chen, Z.; Huang, G.; Li, W.; Teng, J.; Wang, K.; Shao, J.; Loy, C. C.; and Sheng, L. 2023b. Siamese detr. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15722–15731.

Chen, Z.; Huang, G.; Li, W.; Teng, J.; Wang, K.; Shao, J.; Loy, C. C.; and Sheng, L. 2023c. Siamese DETR. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15722–15731.

Cui, C.; Guo, R.; Du, Y.; He, D.; Li, F.; Wu, Z.; Liu, Q.; Wen, S.; Huang, J.; Hu, X.; et al. 2021. Beyond Self-Supervision: A Simple Yet Effective Network Distillation Alternative to Improve Backbones. arXiv:2103.05959.

Dai, Z.; Cai, B.; Lin, Y.; and Chen, J. 2021. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1601–1610.

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.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Hou, X.; Liu, M.; Zhang, S.; Wei, P.; and Chen, B. 2024. Salience detr: Enhancing detection transformer with hierarchical salience filtering refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17574–17583.

Huang, S.; Lu, Z.; Cun, X.; Yu, Y.; Zhou, X.; and Shen, X. 2024. DEIM: DETR with Improved Matching for Fast Convergence. arXiv:2412.04234.

Jäckl, B.; Metz, Y.; Schlegel, U.; Keim, D. A.; and Fischer, M. T. 2024. Leveraging Color Channel Independence for Improved Unsupervised Object Detection. arXiv:2412.15150.

Jia, D.; Yuan, Y.; He, H.; Wu, X.; Yu, H.; Lin, W.; Sun, L.; Zhang, C.; and Hu, H. 2023. Detrs with hybrid matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19702–19712.

Jin, G.; Yang, F.; Sun, M.; Zhao, R.; Liu, Y.; Li, W.; Bao, T.; Wu, L.; Zeng, X.; and Zhao, R. 2023. Seqco-detr: Sequence consistency training for self-supervised object detection with transformers. arXiv:2303.08481.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.

Kumar, C.; Herrera-Gerena, J.; Just, J.; Darr, M.; and Jannesari, A. 2024. Unsupervised learning based object detection using Contrastive Learning. arXiv:2402.13465.

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.

Li, M.; Wu, J.; Wang, X.; Chen, C.; Qin, J.; Xiao, X.; Wang, R.; Zheng, M.; and Pan, X. 2023. Aligndet: Aligning pretraining and fine-tuning in object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6866–6876.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Lin, Y.; Yuan, Y.; Zhang, Z.; Li, C.; Zheng, N.; and Hu, H. 2023. Detr does not need multi-scale or locality design. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6545–6554.

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

Melas-Kyriazi, L.; Rupprecht, C.; Laina, I.; and Vedaldi, A. 2022. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8364–8375.

Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; and Wang, J. 2021. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3651–3660.

Metaxas, I. M.; Bulat, A.; Patras, I.; Martinez, B.; and Tzimiropoulos, G. 2024. Aligned Unsupervised Pretraining of Object Detectors with Self-training. arXiv:2307.15697.

Siméoni, O.; Sekkat, C.; Puy, G.; Vobeckỳ, A.; Zablocki, É.; and Pérez, P. 2023. Unsupervised object localization: Observing the background to discover objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3176–3186.

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.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Wang, X.; Girdhar, R.; Yu, S. X.; and Misra, I. 2023a. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3124–3134.

Wang, X.; Zhang, R.; Shen, C.; Kong, T.; and Li, L. 2021. Dense contrastive learning for self-supervised visual pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3024–3033.

Wang, Y.; Chen, M.; Tang, S.; Zhu, F.; Yang, H.; Bai, L.; Zhao, R.; Yan, Y.; Qi, D.; and Ouyang, W. 2022a. Unsupervised object detection pretraining with joint object priors generation and detector learning. *Advances in neural information processing systems*, 35: 12435–12448.

Wang, Y.; Shen, X.; Yuan, Y.; Du, Y.; Li, M.; Hu, S. X.; Crowley, J. L.; and Vaufreydaz, D. 2023b. TokenCut: Segmenting Objects in Images and Videos With Self-Supervised Transformer and Normalized Cut. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 15790–15801.

Wang, Y.; Zhang, X.; Yang, T.; and Sun, J. 2022b. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 2567–2575.

Xie, E.; Ding, J.; Wang, W.; Zhan, X.; Xu, H.; Sun, P.; Li, Z.; and Luo, P. 2021. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8392–8401.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv:2203.03605.

Zhao, C.; Sun, Y.; Wang, W.; Chen, Q.; Ding, E.; Yang, Y.; and Wang, J. 2024a. Ms-detr: Efficient detr training with mixed supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17027–17036.

Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; and Chen, J. 2024b. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16965–16974.

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:2010.04159.

Zitnick, C. L.; and Dollár, P. 2014. Edge boxes: Locating object proposals from edges. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 391–405. Springer.

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.