SeqCo-DETR: Sequence Consistency Training for Self-supervised Object Detection with Transformers

Guoqiang Jin[1]
jinguoqiang@sensetime.com
Fan Yang[2, 3]
yangfan_2022@ia.ac.cn
Mingshan Sun[1]
sunmingshan@sensetime.com
Ruyi Zhao[1]
zhaoruyi@sensetime.com
Yakun Liu[1]
liuyakun1@sensetime.com
Wei Li[1]
liwei1@sensetime.com
Tianpeng Bao[1]
baotianpeng@sensetime.com
Liwei Wu[1]
wuliwei@sensetime.com
Xingyu Zeng[1]
zengxingyu@sensetime.com
Rui Zhao[1,4]
zhaorui@sensetime.com

[1] SenseTime Research
[2] Institute of Automation, CAS
[3] Peng Cheng Lab
[4] Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, China

Abstract

Self-supervised pre-training and transformer-based architectures have significantly enhanced object detection performance. However, most current self-supervised object detection methods are built on convolutional-based architectures. We believe the transformers' sequence characteristics should be considered when designing a transformer-based self-supervised method for the object detection task. To this end, we propose **SeqCo-DETR**, a novel **Seq**uence **Co**nsistency-based self-supervised method for object **DE**tection with **TR**ansformers. SeqCo-DETR defines a simple yet effective pretext by minimizing the discrepancy of the output sequences of transformers with different image views as input, meanwhile leveraging bipartite matching to find the most relevant sequence pairs that predict the same object. Furthermore, we provide a complementary mask strategy incorporated with the sequence consistency strategy to extract more representative contextual information about the object for the object detection task. Our method achieves state-of-the-art results on MS COCO (45.8 AP) and PASCAL VOC (64.1 AP), demonstrating the effectiveness of our approach.

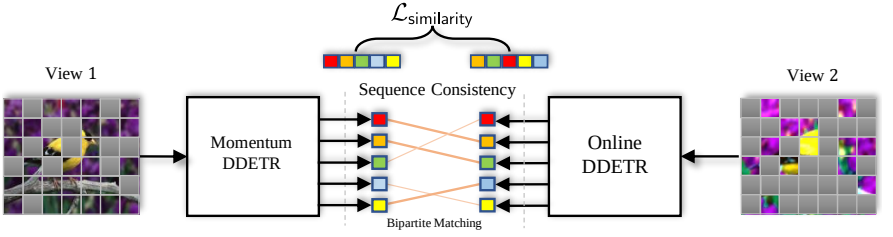


Figure 1: Illustration of the sequence consistency strategy and the complementary mask strategy proposed in our SeqCo-DETR. Note that each output sequence, denoted as colored squares in the figure, contains the object feature and location information.

1 Introduction

Object detection is a prediction-intensive process compared with the image classification task, which involves locating and classifying multiple objects within an image [23]. Existing deep learning-based object detection frameworks can be divided into one-stage methods [20, 29] and two-stage methods [2, 27], either of them requires hand-crafted components. Recently, transformer-based detection methods have emerged as a new object detection paradigm [3, 45], offering a full end-to-end process without hand-crafted components. Unlike convolutional-based architectures, transformer-based architectures define the problem as a sequence-to-sequence process; that is, the transformer converts the input to a sequence and processes the information in the form of a sequence, and the final output is also a sequence [31]. The transformer-based architectures do not rely on the inductive bias characteristics of convolutional-based architectures, such as locality and translation invariance, but instead rely on the global information processing procedure based on attention mechanism [12, 31]. However, these detection methods require supervised training, which demands large amounts of labeled data and extensive human labor since the labeling cost of object detection tasks is much higher than that of image classification tasks.

Self-supervised representation learning is an efficient method to leverage unlabeled data by training models to solve pretext tasks [4, 5, 7, 15, 16]. These customized pretext tasks aim to equip the model with feature representation abilities, which can benefit specific downstream tasks. However, most self-supervised methods are designed for image classification tasks, which consider the image as a whole and only use image-level features. As object detection is a prediction-intensive task that requires object-level features to locate and classify multiple objects in an image, applying these image-level methods directly to object detection leads to limited improvement [39]. Some recent approaches [24, 28, 32, 35, 36, 39, 40, 41, 43] utilize the inductive bias characteristics of convolutional neural networks (CNN) to achieve object-level self-supervised learning, which is not suitable for transformers. Recently, some transformer-based pre-training methods have emerged [1, 8]. However, they handle the pre-training task in an unsupervised way, using hand-crafted pseudo labels to supervise the pre-training process, limiting the model's feature representation ability.

To address the problems mentioned above, we take advantage of the sequence characteristics of transformers and propose a self-supervised object detection pre-training method (SeqCo-DETR) by maintaining consistency of the sequence from different views of an image. As shown in Fig. 1, each output sequence of the transformer decoder stands for an object prediction, which contains the location and category information of the object. Therefore,

the proposed pretext task addresses self-supervised learning on both the location and category of objects, which are the two essential tasks of object detection. Considering that the object prediction of each sequence varies under different image views, we propose to utilize bipartite matching [19] to get the optimal sequence pair to improve the sequence consistency learning process. Additionally, as object detection requires locating the object based on the contextual information around it, we propose adding complementary masks on different image views to help the model learn more global context information about the object. The proposed SeqCo-DETR can pre-train the entire object detection architecture end-to-end, not only the backbone part, thanks to the transformer object detection framework.

In summary, our contributions are: 1) We propose SeqCo-DETR, a transformer-based self-supervised learning method for object detection that maintains sequence consistency between different image views, leveraging the sequence characteristics of transformers and achieving state-of-the-art performance on various benchmarks. 2) We introduce the complementary mask augmentation strategy, designed to complement the proposed sequence consistency strategy by helping the model extract more representative global context information about objects. 3) We adopt bipartite matching to obtain optimal sequence pairs from online and momentum branches with different image views, which boosts the performance of our proposed method.

2 Related Work

Transformer-based object detection methods. The DEtection TRansformers (DETR) [3] brings a new paradigm of object detection tasks, which is a fully end-to-end method without hand-crafted components. DETR is based on the encoder-decoder transformers and defines the object detection problem as a set prediction problem, which does not rely on the inductive bias of convolutional-based architectures. However, the training speed of the original DETR is considerably slow, and the detection results on small objects are limited. To solve the problems, Deformable DETR [45] uses multiscale features and proposes a deformable attention mechanism, which significantly accelerates the convergence speed and improves the overall accuracy. Therefore, our method is based on the Deformable DETR framework.

Self-supervised representation learning. Instance discrimination is one of the competitive pretext tasks for self-supervised visual representation learning, which aims to learn such an embedding space in which similar sample pairs stay close to each other while dissimilar ones are far apart [37]. MoCo [6, 7, 16] improved contrastive methods by storing representations in a momentum structure. SimCLR [5] proved that the memory bank can be replaced with large batch sizes and more image augmentations. SwAV [4] took the features as a set of trainable clustering prototype vectors. BYOL [15] utilized the asymmetric architecture together with the stop gradient design to bootstrap the representations by extracting features from different views of the same instance, which could be trained without negative samples. Although these methods have shown promising performance in image classification tasks, they have a limited improvement in prediction-intensive tasks such as object detection [39].

Self-supervised object detection methods. In order to improve the object detection task via self-supervised learning, several methods have been proposed recently. DetCo [39] proposed a contrastive loss between local patches and the global image to improve the multi-level feature expression ability for detection tasks. ORL [40] achieved object-level representation based on scene images. SCRL [28], ReSim [38], DetCon [18], ContrastiveCrop [26], and SoCo [34] proposed to maintain region-level consistency in the related areas by different

methods. DenseCL [32] and PixPro [41] proposed to utilize the pixel-level to achieve dense contrastive learning. SlotCon [35] proposed to utilize the group/object-level to achieve dense contrastive learning. Self-EMD [24] proposed to utilize the spatial information of CNN features and used Earth Mover's Distance to match features from different views. InsLoc [43] and Align Yourself [36] pasted cropped images at different locations, then minimized the corresponding features extracted from different views, while it failed to consider the localization task in the object detection. However, these methods rely on the inductive bias characteristics of CNN, which is not suitable for transformers.

More recently, UP-DETR [8] designed an unsupervised pre-training pretext based on the transformer architecture, which uses random patches as input to predict patch locations. Since patch locations are known, the proposed pretext task is more like a supervised method with pseudo labels. DETReg [1] used Selective Search [30] to generate region proposals as the location supervision instead of random proposals and used a pre-trained model as the feature supervision, which is an unsupervised method and does not incorporate any self-supervised pre-training part. Thus, the feature representation's ability is limited by the selected pre-trained model and the model's performance would be affected by the handy-crafted supervision. In contrast, our proposed method uses self-supervised learning via maintaining the consistency of transformer sequences, thus the ability to learn feature representation is not limited by a fixed pre-trained model.

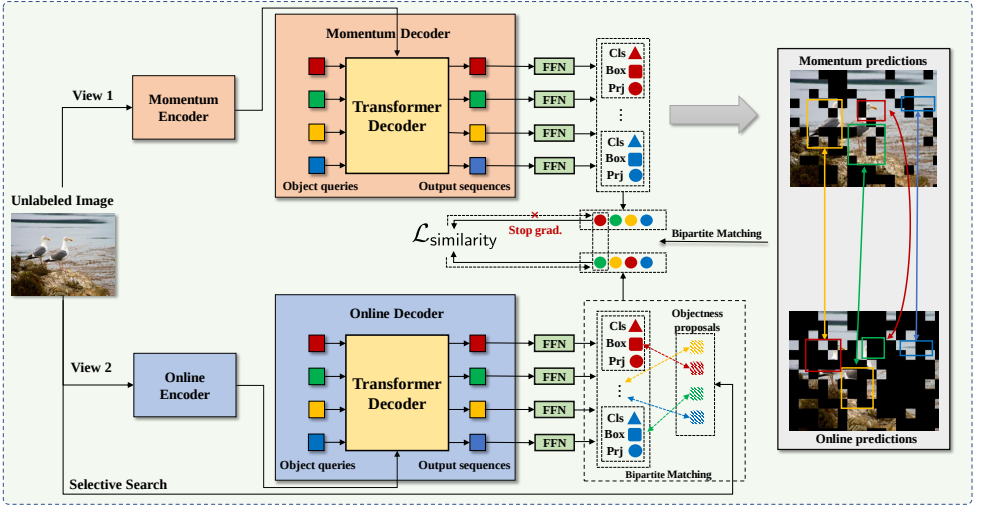


Figure 2: The proposed SeqCo-DETR consists of two branches: the online and momentum branch. The Online/Momentum Encoder in the figure consists of a CNN backbone and a transformer encoder. The input of each branch is a different view of the unlabeled image. Notably, we use complementary image masks for each view to ensure that each branch only sees a non-overlapping part of the image. After processing through the transformer, all object queries are transformed into output sequences. Then the sequences pass through three feedforward networks (FFN) to predict the object's class, bounding box, and feature projection. There are two bipartite matching processes: one between the outputs of two branches, the other between the output of online branches and the "objectness" proposals provided by the Selective Search. Finally, the similarity loss of each paired sequence feature projection is minimized to achieve self-supervised representation learning at the sequence level.

3 Method

3.1 Sequence Consistency Strategy

SeqCo-DETR is designed to leverage the sequence characteristics to achieve self-supervised pre-training for object detection using transformers. The main framework of our SeqCo-DETR is shown in Fig. 2. The core idea of SeqCo-DETR is to maintain consistency between sequences from differently augmented views of the same image, i.e., in a self-supervised learning way. Thus, we utilize the momentum design [16, 25, 42] to achieve self-supervised learning, which contains the online branch and the momentum branch. Both branches share the same structure, including CNN backbone, transformer encoder, transformer decoder, and FFN heads. In particular, we incorporate a projection head [5] after the transformer decoder to generate the feature projection for each sequence. The classification and box regression heads are also used during the pre-training process. After pre-training, the projection head will be removed, and the classification head will be reset and modified to match the number of categories in downstream detection tasks. The remaining pre-trained weights are then loaded as the initial weights during fine-tuning. In line with the momentum design, a stop gradient operator is introduced to ensure that the gradient only updates the weight of the online branch, while the weight of the momentum branch is updated based on the momentum of the weight of the online branch, as per the formula:

$$\Theta_m \leftarrow \beta * \Theta_m + (1-\beta) * \Theta_o, \tag{1}$$

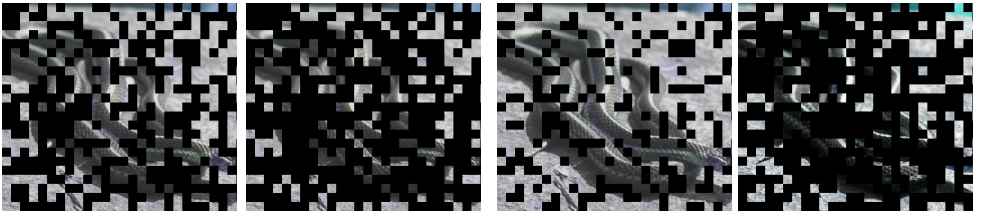
where Θ_m and Θ_o represent the parameters of the momentum branch and the online branch, respectively, β represents the momentum coefficient. Since the momentum branch has a stop gradient design and its parameters are updated by momentum parameters, the two branches would update at different speeds, which can effectively prevent network collapse [15].

The proposed sequence consistency strategy is simple and straightforward. Since each output sequence stands for an object prediction and each sequence contains the most relevant feature description for each object, we apply the consistency constraint on the sequences that predict the same object. Thus, we could maintain the object-level feature consistency instead of the image-level feature, which is more suitable for object detection tasks. The sequence consistency strategy is formulated as follows:

$$\mathcal{L}_{\text{ssl}} = \sum_{i=1}^{N} (\mathcal{L}_{\text{similarity}}(\mathbf{f}_{\text{ffn}}(\mathbf{s}_i), \widehat{\mathbf{f}}_{\text{ffn}}(\widehat{\mathbf{s}}_{\widehat{\sigma}(i)}))), \tag{2}$$

where $\mathbf{s}$ and $\widehat{\mathbf{s}}$ are the output of the sequence by the transformer decoder from the momentum and online branches, respectively; $\mathbf{f}_{\text{ffn}}$ and $\widehat{\mathbf{f}}_{\text{ffn}}$ represent the FFN heads from the momentum and online branches, respectively; N is the number of sequences in one view, and $\widehat{\sigma}$ represents the matching relationship between the two sequences. $\mathcal{L}_{\text{ssl}}$ is the total self-supervised learning loss. $\mathcal{L}_{\text{similarity}}$ is a function that measures the similarity between sequences, and we adopt $\mathcal{L}_2$ as the $\mathcal{L}_{\text{similarity}}$ loss.

As mentioned earlier, since the image views for the two branches are different, the output sequences from each branch would also differ. Therefore, to ensure that we match the sequences that have the same object prediction, we employ bipartite graph matching [19]. The bipartite matching is a classical optimal matching method that is designed to match the elements in two sets, which allows us to obtain the optimal sequence pair from the output of



(a) Mask 50% (b) Mask 70% (c) Complementary mask of 30% and 70%

Figure 3: Examples of different proportions of random masked images.

online and momentum branches, as defined below:

$$\mathcal{L}_{\text{match}}(\mathbf{y}, \widehat{\mathbf{y}}_{\sigma}) = \sum_{i=1}^{N} [-\lambda_{cm} \log \widehat{\mathbf{p}}_{\sigma(i)}(c_i) + \mathbb{1}_{\{c_i \neq \varnothing\}} (\lambda_{bm} \mathcal{L}_{\text{box}}(\mathbf{b}_i, \widehat{\mathbf{b}}_{\sigma(i)}))], \tag{3}$$

$$\widehat{\sigma} = \underset{\sigma \in \Sigma_N}{\arg\min} \sum_{i}^{N} \mathcal{L}_{\text{match}}(\mathbf{y}_i, \widehat{\mathbf{y}}_{\sigma(i)}), \tag{4}$$

where N is the number of sequences in one view; $\mathcal{L}_{\text{match}}$ denotes the Hungarian matching loss; $\mathbf{y}$ and $\widehat{\mathbf{y}}$ are the predicted sequences from momentum and online branches, respectively; $\mathbf{b}$ and c are the location prediction and category prediction, respectively; $\widehat{\mathbf{p}}_{\widehat{\sigma}(i)}(c_i)$ is the probability of class c_i; $\widehat{\sigma}$ denotes the final optimal assignment; Σ_N denotes the set of all possible matches; $\varnothing$ denotes the empty set; λ_{cm} and λ_{bm} are the corresponding weights, which are 2.0 and 5.0, respectively. After the output sequences, which predict objects at the same location, have been matched by the bipartite graph matching, the corresponding features from the sequences are used to calculate the $\mathcal{L}_{\text{ssl}}$. Specifically, the two image views share the same location augmentation parameters, and there are only content differences between the two views, such as color jittering, random erasing, and blur, which reduces the difficulty of the output sequences matching. The comparison results for the different sequence utilization strategies are summarized in Tab. 4.

In order to get more rich semantic features rather than some background features from an image to calculate the self-supervised loss, the transformer needs to predict proposals that contain foreground objects as many as possible. To achieve this, we utilize the Selective Search [30] to initial foreground proposals to train the neural network to have the ability to predict "objectnes" proposals. The Selective Search is an unsupervised method that could only generate the position of foreground proposals, not the category. Thus, there will be only two categories, i.e., foreground and background. The parameters of the Selective Search are the same as in DETReg [1]. The comparison results for the region proposal strategy are summarized in Tab. 3.

3.2 Mask strategy

Object localization is a crucial task in object detection, requiring not only the features of the object but also the context information around the object. Therefore, a strategy to enhance the model's ability to extract global context information is necessary for self-supervised object detection. Mask-based image augmentation, especially in combination with transformers, has proven to be an efficient way to extract global context information [13, 17, 44]. The key

insight is that the attention mechanism in a transformer can effectively process global information, while object detection requires contextual information surrounding the objects. By adding a mask to an image, the network can be forced to use more distant context information to extract features, thereby improving the network's global feature extraction capabilities.

To this end, we design a mask-based image augmentation incorporated with the proposed sequence consistency strategy to enhance the feature representation capability of the transformer, leading to improved object detection performance. Since the gradient updates only the online branch, there is usually a more strong image augmentation used in the online branch [25, 42]. Thus, one straightforward way is to add the mask only to the online branch view, as shown in Fig. 3 (a,b). To further compel the network to exploit contextual information, inspired by [13], we design a complementary mask strategy that adds complementary masks on both branches. Specifically, with the complementary masks, each image view will not have overlapped areas with each other, as shown in Fig. 3 (c). Therefore, the predictions from each branch will not depend on the same local areas, but on the global context information. By minimizing the sequence consistency loss, the network is directed to extract and combine global context information to predict objects, which is useful for object detection. The comparison results for the different mask strategies are summarized in Tab. 2.

The total loss of SeqCo-DETR consists of the region proposals loss $\mathcal{L}_{RPS}$ and the self-supervised learning loss on sequences $\mathcal{L}_{SSL}$, denoted as:

$$\begin{aligned}\mathcal{L}_{\text{total}}(\mathbf{y},\widehat{\mathbf{y}}) =& \mathcal{L}_{RPS}+\mathcal{L}_{SSL}\\ =& \sum_{i=1}^{N}[\lambda_f \mathcal{L}_{\text{focal}}(c_i,\widehat{\mathbf{p}}_{\widehat{\sigma}(i)}) + \mathbb{1}_{\{c_i\neq\varnothing\}}(\lambda_b \mathcal{L}_{\text{box}}(\mathbf{b}_i,\widehat{\mathbf{b}}_{\widehat{\sigma}(i)})]\\ &+\sum_{i=1}^{N}[\lambda_e \mathcal{L}_{\text{ssl}}(\mathbf{z}_i,\widehat{\mathbf{z}}_{\widehat{\sigma}(i)}))],\end{aligned} \tag{5}$$

where $\mathcal{L}_{\text{focal}}$ is the focal loss of classification, $\mathcal{L}_{\text{box}}$ is the location loss of box, $\mathcal{L}_{\text{ssl}}$ is the self-supervised learning loss of sequence, and $\mathbf{z}_i$ stands for the output of $\mathbf{f}_{\text{ffn}}$. The $\mathcal{L}_{RPS}$ consists of $\mathcal{L}_{\text{focal}}$ and $\mathcal{L}_{\text{box}}$, which is the same as in DETReg [1], supervised by the proposals from Selective Search. λ_f, λ_b, and λ_e are the weights of those three losses, which are set to 2.0, 5.0, and 10.0, respectively.

4 Experimental Results

4.1 Experimental Settings

Datasets. Our experiments include the pre-training stage and the fine-tuning stage. First, we pre-train models on the unlabeled dataset. Then, we load the pre-trained weight and fine-tune the network on the downstream object detection tasks following the standard procedure [1]. For the pre-training stage, we use ImageNet (IN1K) [10] and ImageNet100 (IN100) [9] as the main pre-training datasets. ImageNet100 is a subset of ImageNet, which only contains 100 classes. The split of classes is the same as in DETReg [1]. For COCO results, the pre-training dataset is IN1K; for VOC results, the pre-training dataset is IN100. The ablation studies are all based on the IN100. In addition, we also pre-train our model on multi-object datasets on COCO and COCO+ in the ablation studies, where COCO stands for `train2017` without ground truth labels, COCO+ denotes the COCO `train2017` plus the COCO `unlabeled`

dataset. For the fine-tuning stage, we evaluate our method on MS COCO [21] and PASCAL VOC [14]. In particular, we fine-tune the model on COCO `train2017` and evaluate it on COCO `val2017`. As for VOC, we fine-tune on VOC `trainval07+12`, and then evaluate on `test07`. The comparison approaches are DETReg [1], UP-DETR [8] which is based on Deformable DETR, JoinDet [33] which is the latest transformer-based self-supervised method, Deformable DETR [45] with various pre-trained weights, and the common baseline Faster R-CNN [27]. Additional implementation details are in the supplementary material.

Image augmentations. The complementary mask strategy is designed to incorporate with the sequence consistency strategy. Thus, the image augmentations for the two branches are different, where weak image augmentations are used in the momentum branch while strong augmentations are used in the online branch. The complementary mask proportions for the online and momentum branches are 70% and 30%, respectively, with a patch size of 16, which are all chosen by experiments. To ensure the two views have consistency in the location parameters, the image *view* 2 is partially built upon the *view* 1. Following [10], we first generate a base image view from the input unlabeled image, using random flips, random resize, and random resized-crop. The base image view is the same as in DETReg [1]. Then, for the momentum branch, we add the corresponding mask upon the base image view to generate *view* 1. As for the *view* 2, we add more augmentations based on the base image view, including color jitter, random grayscale, random blur, and the corresponding mask. Thus, there is no location coordinates difference between the two views.

4.2 Main Results

Model	COCO `val2017` AP	AP_{50}	AP_{75}	VOC `test07` AP	AP_{50}	AP_{75}
Faster R-CNN [27]	42.0	62.1	45.5	56.1	82.6	62.7
Deformable DETR (Supervised CNN) [45]	43.8	62.6	47.7	59.5	82.6	65.6
Deformable DETR (SimCLR CNN)†	41.5	59.8	45.4	57.3	80.0	63.6
Deformable DETR (BYOL CNN)†	44.7	63.8	48.8	59.9	82.7	66.7
Deformable DETR (MoCo CNN)†	43.1	61.6	46.9	59.6	81.8	66.0
Deformable DETR (SwAV CNN)†	45.0	63.8	49.2	61.0	83.0	68.1
UP-DETR (Deformable DETR)‡	44.7	63.7	48.6	61.8	83.4	69.6
JoinDet [33]	45.6	**64.3**	49.8	63.7	83.8	70.7
DETReg w/o feature embedding†	45.2	63.7	49.5	63.0	83.5	70.2
DETReg [1]	45.5	64.1	49.9	63.5	83.3	70.3
SeqCo-DETR	**45.8**	64.2	**50.0**	**64.1**	**83.8**	**71.6**

Table 1: Comparison results on the object detection datasets. †: We run the method on our codebase. ‡: Results are provided by DETReg [1].

The experiment results are presented in Tab. 1. The above part of the table is the widely referenced baselines in object detection tasks, which are listed here for convenience. The middle part is our baseline, and the bottom part is our approach. As DETReg's training process is supervised by the pre-trained SwAV model [4], the pre-training process can be regarded as learning the fixed features from SwAV. In contrast, our method proposes to use the self-supervised way to learn the features, which could help the network learn more discriminative features during pre-training. To establish a baseline for comparison, we remove the

Model	Mask strategy	AP
DETReg	w/o Mask (baseline) [1]	45.4
	w/ Mask_{50} †	45.0
SeqCo-DETR	w/o Mask	45.6
	$\text{Mask}_{online@50}$	45.6
	$\text{Mask}_{online@50} + \text{Mask}_{momentum@50}$	45.4
	$\text{Mask}_{online@70} + \text{Mask}_{momentum@30}$	45.6
	$\text{Mask}_{online@70} + \text{Mask}_{\neg(online@70)}$	**45.8**

Table 2: Comparison of mask strategies, evaluated on MS COCO `val2017`. †: We run the method on our codebase.

feature embedding learning part of DETReg in the experiment, the results are 45.2 on COCO and 63.0 on VOC. For the final result, our method achieves 45.8 and surpasses DETReg by 0.3 points on COCO and 0.6 points on VOC, proving that our learning-based method has better performance than the method that is supervised by manually defined pseudo-labels. Compared to the methods that only the backbone part is pre-trained, i.e., Deformable DETR (Supervised CNN) and Deformable DETR with different self-supervised pre-trains such as SimCLR [5], BYOL [3], MoCo v2 [6], and SwAV [2], our method could pre-train the entire object detection framework and surpass Deformable DETR (SwAV CNN) by 0.8 points on COCO and 3.1 points on VOC. Thus, our approach is more suitable for the transformer-based object detection task and achieves state-of-the-art results.

4.3 Ablation Study

Mask strategy. We conduct extensive experiments about the mask strategy, as listed in Tab. 2. When the complementary masks are used in online and momentum branches, with mask proportion of 70% and 30%, respectively, denoted as $\text{Mask}_{online@70} + \text{Mask}_{\neg(online@70)}$, we get the best performance of 45.8. To verify whether the complementary characteristic is the most critical design in the mask design, we add the independent random masks for the two branches with the same mask proportion of 70% and 30%, respectively, denoted as $\text{Mask}_{online@70} + \text{Mask}_{momentum@30}$. The corresponding result is 45.6. We also try the same proportion masks for the two branches, with the proportion of 50%, denoted as $\text{Mask}_{online@50} + \text{Mask}_{momentum@50}$. The result is only 45.4. When the mask is only added to the online branch, with the mask proportion of 50%, denoted as $\text{Mask}_{online@50}$ or Mask_{50}, the result of DETReg drops from 45.4 to 45.0. Meanwhile, the performance of ours does not drop. The mask strategy may interfere with the training process that is only supervised by the pseudo labels. Our experiments prove that the complementary mask is a more effective way compared to adding random masks. Additionally, incorporating the complementary mask with the sequence consistency strategy results in improved performance. More experiments on mask parameter selection can be found in the supplementary material.

Pre-training datasets and region proposal strategy. We conduct several experiments to compare the performance with different types of pre-training datasets. `Rnd bbox` stands for the random proposals, and COCO GT stands for the region proposals that come from the COCO ground truth. For COCO and COCO+, we use the images without ground truth and Selective Search to generate initial proposals. Specifically, COCO is a multi-object dataset, while ImageNet is a single-object dataset. As listed in Tab. 3, our method achieves better results against DETReg on both single-object and multi-object datasets, which proves that

Method	IN100	IN100 (Rnd bbox)	COCO	COCO+	COCO GT
DETReg†	45.4	44.1	45.1	45.1	45.6
SeqCo-DETR	**45.8**	**44.3**	**45.6**	**45.6**	**45.8**

Table 3: Comparison of pre-training datasets and region proposal strategies, evaluated on MS COCO `val2017`. †: We run the method on our codebase.

Model	One-by-one matching	Bipartite matching	Multi-feature	AP
SeqCo-DETR	✓			45.6
	✓		✓	45.3
		✓	✓	45.5
		✓		**45.8**

Table 4: Comparison of sequence utilization strategies, evaluated on MS COCO `val2017`.

self-supervised learning has better generality over different types of datasets. Furthermore, we compare the influence of different region proposal strategies. When using the less "objectnes" proposals, i.e., random proposals, both methods drop a lot. When using the ground truth as the region proposal, DETReg improves from 45.1 to 45.6, whereas ours improves from 45.6 to 45.8. This proves that methods that solely rely on hand-crafted pseudo labels are susceptible to the quality of the pseudo labels, while learning-based methods are less affected by them. Our results also indicate that the quality of the region proposals generated by Selective Search is sufficient for our method to learn useful information.

Sequence utilization methods. We conduct several experiments to compare the performance with different types of sequence utilization methods, as listed in Tab. 4. One of the most naive strategies is the one-by-one match, because the sequence output by transformers has sequential characteristics. However, since different branches have different input image views and slightly different network parameters, their predicted results differ even if the output sequences are in the same order. Therefore, bipartite matching is adopted to match the sequences from different branches that predict the same object. As listed in Tab. 4, bipartite matching could improve results from 45.6 to 45.8, compared to the one-by-one matching, proving the effectiveness of the bipartite matching method. Meanwhile, to achieve more sufficient supervision of the sequence, we try to use the outputs of classification head f_{cls}, regression head f_{box}, and projection head f_{prj} in Eq. (2) at the same time. However, it can be seen from the table that the fusion of multiple heads decreases the final results in both matching settings. Perhaps the self-supervision on the projection head is enough; redundancy supervision on three heads would cause a performance drop.

5 Conclusion

In this paper, we introduce SeqCo-DETR, a novel self-supervised learning method for object detection based on transformers. Our approach exploits the sequential nature of transformer networks to achieve self-supervised learning of detection, maintaining sequence consistency under different image views. To extract more global context information and enhance object detection, we propose a complementary mask strategy. Additionally, we use bipartite matching to optimize sequence-level self-supervision. Extensive experiments on various downstream detection tasks and on both single-object and multi-object datasets prove the effectiveness of the SeqCo-DETR.

Acknowledgement

This work is sponsored by Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone (HZQB-KCZYZ-2021045).

References

[1] Amir Bar, Xin Wang, Vadim Kantorov, Colorado J Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Detreg: Unsupervised pretraining with region priors for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14605–14615, 2022.

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020.

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[7] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[8] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. UP-DETR: unsupervised pre-training for object detection with transformers. In *CVPR*, pages 1601–1610, 2021.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255, 2009.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[11] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR*, 2021.

[13] Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pre-training? *arXiv preprint arXiv:2112.10740*, 2021.

[14] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, pages 303–338, 2010.

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[18] Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10086–10096, 2021.

[19] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[20] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018.

[21] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014.

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[23] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International journal of computer vision*, 128(2):261–318, 2020.

[24] Songtao Liu, Zeming Li, and Jian Sun. Self-emd: Self-supervised object detection without imagenet. *CoRR*, 2020.

[25] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[26] Xiangyu Peng, Kai Wang, Zheng Zhu, Mang Wang, and Yang You. Crafting better contrastive views for siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16031–16040, 2022.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[28] Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatially consistent representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1153, 2021.

[29] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.

[30] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104 (2):154–171, 2013.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[32] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021.

[33] Yizhou Wang, Meilin Chen, Shixiang Tang, Feng Zhu, Haiyang Yang, Lei Bai, Rui Zhao, Yunfeng Yan, Donglian Qi, and Wanli Ouyang. Unsupervised object detection pretraining with joint object priors generation and detector learning. *Advances in Neural Information Processing Systems*, 35:12435–12448, 2022.

[34] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[35] Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. In *Advances in Neural Information Processing Systems*, volume 35, pages 16423–16438, 2022.

[36] Di Wu, Siyuan Li, Zelin Zang, Kai Wang, Lei Shang, Baigui Sun, Hao Li, and Stan Z Li. Align yourself: Self-supervised pre-training for fine-grained recognition via saliency alignment. *arXiv preprint arXiv:2106.15788*, 2021.

[37] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[38] Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10539–10548, 2021.

[39] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8392–8401, 2021.

[40] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Unsupervised object-level representation learning from scene images. *Advances in Neural Information Processing Systems*, 34:28864–28876, 2021.

[41] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021.

[42] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3060–3069, 2021.

[43] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3987–3996, 2021.

[44] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

[45] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations, ICLR*, 2021.