Accelerating DETR Convergence via Semantic-Aligned Matching

Gongjie Zhang[1] Zhipeng Luo[1,2] Yingchen Yu[1] Kaiwen Cui[1] Shijian Lu*[1]
[1]Nanyang Technological University, Singapore [2]SenseTime Research
{gongjiezhang, shijian.lu}@ntu.edu.sg {zhipeng001, yingchen001, kaiwen001}@e.ntu.edu.sg

Abstract

The recently developed DEtection TRansformer (DETR) establishes a new object detection paradigm by eliminating a series of hand-crafted components. However, DETR suffers from extremely slow convergence, which increases the training cost significantly. We observe that the slow convergence is largely attributed to the complication in matching object queries with target features in different feature embedding spaces. This paper presents SAM-DETR, a Semantic-Aligned-Matching DETR that greatly accelerates DETR's convergence without sacrificing its accuracy. SAM-DETR addresses the convergence issue from two perspectives. First, it projects object queries into the same embedding space as encoded image features, where the matching can be accomplished efficiently with aligned semantics. Second, it explicitly searches salient points with the most discriminative features for semantic-aligned matching, which further speeds up the convergence and boosts detection accuracy as well. Being like a plug and play, SAM-DETR complements existing convergence solutions well yet only introduces slight computational overhead. Extensive experiments show that the proposed SAM-DETR achieves superior convergence as well as competitive detection accuracy. The implementation codes are publicly available at https://github.com/ZhangGongjie/SAM-DETR.

1. Introduction

Object detection is one of the most fundamental tasks in computer vision and has achieved unprecedented progress with the development of deep learning [27]. However, most object detectors often suffer from complex detection pipelines and sub-optimal performance due to their over-reliance on hand-crafted components such as anchors, rule-based target assignment, and non-maximum suppression (NMS). The recently proposed DEtection TRansformer (DETR) [3] removes the need for such hand-designed components and establishes a fully end-to-end framework for object detection. Despite its simple design and promising results, one of the most significant drawbacks of DETR is its extremely slow convergence on training, which requires 500 epochs to converge on the COCO benchmark [26], while Faster R-CNN [35] only takes 12∼36 epochs instead. This slow convergence issue significantly increases the training cost and thus hinders its more comprehensive applications.

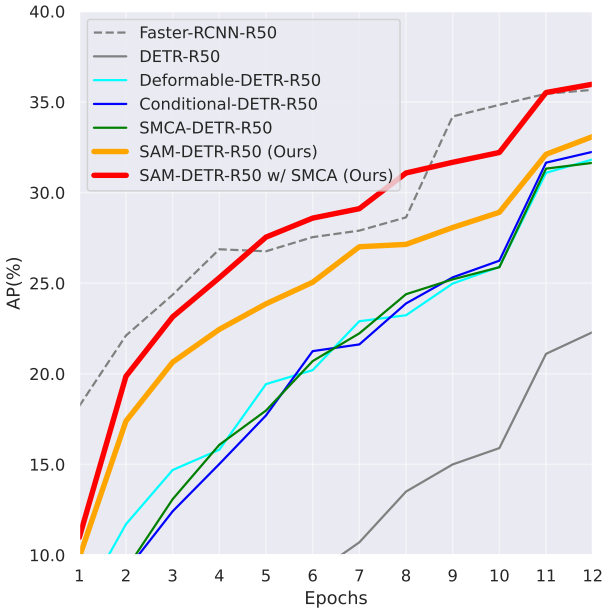

Figure 1. Convergence curves of our proposed SAM-DETR and other detectors on COCO val 2017 under the 12-epoch training scheme. All competing methods are single-scale. SAM-DETR converges much faster than the original DETR, and can work in complementary with existing convergence-boosting solutions, reaching a comparable convergence speed with Faster R-CNN.

DETR employs a set of object queries in its decoder to detect target objects at different spatial locations. As shown in Fig. 2, in the cross-attention module, these object queries are trained with a set-based global loss to match the target objects and distill corresponding features from the matched regions for subsequent prediction. However, as pointed out in [10, 31, 63], each object query is almost equally matched to all spatial locations at initialization, thus

* Corresponding author.

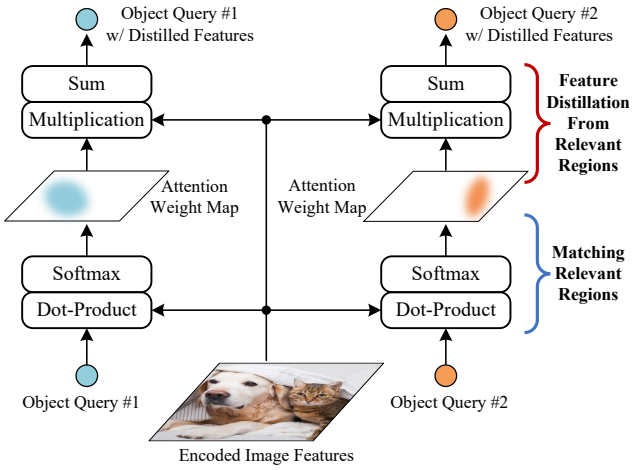

Figure 2. The cross-attention module in DETR's decoder can be interpreted as a 'matching and feature distillation' process. Each object query first matches its own relevant regions in encoded image features, and then distills features from the matched regions, generating output for subsequent prediction.

requiring tedious training iterations to learn to focus on relevant regions. The matching difficulty between object queries and corresponding target features is the major reason for DETR's slow convergence.

A few recent works have been proposed to tackle the slow convergence issue of DETR. For example, Deformable DETR [63] replaces the original global dense attention with deformable attention that only attends to a small set of features to lower the complexity and speed up convergence. Conditional DETR [31] and SMCA-DETR [10] modify the cross-attention module to be spatially conditioned. In contrast, our approach works from a different perspective without modifying the attention mechanism.

Our core idea is to ease the matching process between object queries and their corresponding target features. One promising direction for matching has been defined by Siamese-based architecture, which aligns the semantics of both matching sides via two identical sub-networks to project them into the same embedding space. Its effectiveness has been demonstrated in various matching-involved vision tasks, such as object tracking [1, 4, 20, 21, 46, 47], re-identification [5, 37, 38, 48, 59], and few-shot recognition [15, 19, 39, 41, 55]. Motivated by this observation, we propose *Semantic-Aligned-Matching DETR (SAM-DETR)*, which appends a plug-and-play module ahead of the cross-attention module to semantically align object queries with encoded image features, thus facilitating the subsequent matching between them. This imposes a strong prior for object queries to focus on semantically similar regions in encoded image features. In addition, motivated by the importance of objects' keypoints and extremities in recognition and localization [3, 31, 62], we propose to explicitly search multiple salient points and use them for semantic-aligned matching, which naturally fits in the DETR's original multi-head attention mechanism. Our approach only introduces a plug-and-play module into the original DETR while leaving most other operations unchanged. Therefore, the proposed method can be easily integrated with existing convergence solutions in a complementary manner.

In summary, the contributions of this work are four-fold. *First*, we propose *Semantic-Aligned-Matching DETR (SAM-DETR)*, which significantly accelerates DETR's convergence by innovatively interpreting its cross-attention as a 'matching and distillation' process and semantically aligning object queries with encoded image features to facilitate their matching. *Second*, we propose to explicitly search for objects' salient points with the most discriminative features and feed them to the cross-attention module for semantic-aligned matching, which further boosts the detection accuracy and speeds up the convergence of our model. *Third*, experiments validate that our proposed SAM-DETR achieves significantly faster convergence compared with the original DETR. *Fourth*, as our approach only adds a plug-and-play module into the original DETR and leaves other operations mostly unchanged, the proposed SAM-DETR can be easily integrated with existing solutions that modify the attention mechanism to further improve DETR's convergence, leading to a comparable convergence speed with Faster R-CNN even within 12 training epochs.

2. Related Work

Object Detection. Modern object detection methods can be broadly classified into two categories: two-stage and single-stage detectors. Two-stage detectors mainly include Faster R-CNN [35] and its variants [2, 9, 16, 23, 32, 44, 49, 51, 54], which employ a Region Proposal Network (RPN) to generate region proposals and then make per-region predictions over them. Single-stage detectors [17, 28, 29, 33, 34, 43, 57, 61, 62] skip the proposal generation and directly perform object classification and localization over densely placed sliding windows (anchors) or object centers. However, most of these approaches still rely on many hand-crafted components, such as anchor generation, rule-based training target assignment, and non-maximum suppression (NMS) post-processing, thus are not fully end-to-end.

Distinct from the detectors mentioned above, the recently proposed DETR [3] has established a new paradigm for object detection [50, 55, 56, 60, 63]. It employs a Transformer [45] encoder-decoder architecture and a set-based global loss to replace the hand-crafted components, achieving the first fully end-to-end object detector. However, DETR suffers from severe low convergence and requires extra-long training to reach good performance compared with those two-stage and single-stage detectors. Several works have been proposed to mitigate this issue: De-

formable DETR [63] replaces the original dense attention with sparse deformable attention; Conditional DETR [31] and SMCA-DETR [10] propose conditioned cross-attention and Spatially Modulated Co-Attention (SMCA), respectively, to replace the cross-attention module in DETR's decoder, aiming to impose spatial constraints to the original cross-attention to better focus on prominent regions. In this work, we also aim to improve DETR's convergence, but from a different perspective. Our approach does not modify the original attention mechanism in DETR, thus can work in complementary with existing methods.

Siamese-based Architecture for Matching. Matching is a common concept in vision tasks, especially in contrastive tasks such as face recognition [36, 40], re-identification [5, 14, 22, 37, 38, 48, 59], object tracking [1, 4, 8, 11, 20, 21, 42, 46, 47, 52, 58, 64], few-shot recognition [15, 19, 39, 41, 53, 55], *etc*. Its core idea is to predict the similarity between two inputs. Empirical results have shown that Siamese-based architectures, which project both matching sides into the same embedding space, perform exceptionally well on the tasks involving matching. Our work is motivated by this observation to interpret DETR's cross-attention as a 'matching and feature distillation' process. To achieve fast convergence, it is crucial to ensure the aligned semantics between object queries and encoded image features, *i.e.*, both of them are projected into the same embedding space.

3. Proposed Method

In this section, we first review the basic architecture of DETR, and then introduce the architecture of our proposed *Semantic-Aligned-Matching DETR (SAM-DETR)*. We also show how to integrate our approach with existing convergence solutions to boost DETR's convergence further. Finally, we present and analyze the visualization of a few examples to illustrate the mechanism of our approach and demonstrate its effectiveness.

3.1. A Review of DETR

DETR [3] formulates the task of object detection as a set prediction problem and addresses it with a Transformer [45] encoder-decoder architecture. Given an image $\mathbf{I} \in \mathbb{R}^{H_0 \times W_0 \times 3}$, the backbone and the Transformer encoder produce the encoded image features $\mathbf{F} \in \mathbb{R}^{HW \times d}$, where d is the feature dimension, and H_0, W_0 and H, W denote the spatial sizes of the image and the features, respectively. Then, the encoded image features $\mathbf{F}$ and a small set of object queries $\mathbf{Q} \in \mathbb{R}^{N \times d}$ are fed into the Transformer decoder to produce detection results, where N is the number of object queries, typically $100 \sim 300$.

In the Transformer decoder, object queries are sequentially processed by a self-attention module, a cross-attention module, and a feed-forward network (FFN) to produce the outputs, which further go through a Multi-Layer Perceptron (MLP) to generate prediction results. A good way to interpret this process is: object queries denote potential objects at different spatial locations; the self-attention module performs message passing among different object queries; and in the cross-attention module, object queries first search for the corresponding regions to match, then distill relevant features from the matched regions for the subsequent predictions. The cross-attention mechanism is formulated as:

$$\mathbf{Q}' = \overbrace{\underbrace{\mathrm{Softmax}(\frac{(\mathbf{Q}\mathbf{W}_{\mathrm{q}})(\mathbf{F}\mathbf{W}_{\mathrm{k}})^{\mathrm{T}}}{\sqrt{d}})}_{\text{to distill features from matched regions}}(\mathbf{F}\mathbf{W}_{\mathrm{v}})}^{\text{to match relevant regions}}, \quad (1)$$

where $\mathbf{W}_{\mathrm{q}}$, $\mathbf{W}_{\mathrm{k}}$, and $\mathbf{W}_{\mathrm{v}}$ are the linear projections for query, key, and value in the attention mechanism. Ideally, the cross-attention module's output $\mathbf{Q}' \in \mathbb{R}^{N \times d}$ should contain relevant information distilled from the encoded image features to predict object classes and locations.

However, as pointed out in [10, 31, 63], the object queries are initially equally matched to all spatial locations in the encoded image features, and it is very challenging for the object queries to learn to focus on specific regions properly. The matching difficulty is the key reason that causes the slow convergence issue of DETR.

3.2. SAM-DETR

Our proposed SAM-DETR aims to relieve the difficulty of the matching process in Eq. 1 by semantically aligning object queries and encoded image features into the same embedding space, thus accelerating DETR's convergence. Its major difference from the original DETR [3] lies in the Transformer decoder layers. As illustrated in Fig. 3 (a), the proposed SAM-DETR appends a *Semantics Aligner* module ahead of the cross-attention module and models learnable *reference boxes* to facilitate the matching process. Same as DETR, the decoder layer is repeated six times, with zeros as input for the first layer and previous layer's outputs as input for subsequent layers.

The learnable reference boxes $\mathbf{R}_{\mathrm{box}} \in \mathbb{R}^{N \times 4}$ are modeled at the first decoder layer, representing the initial locations of the corresponding object queries. With the localization guidance of these reference boxes, the proposed Semantics Aligner takes the previous object query embeddings $\mathbf{Q}$ and the encoded image features $\mathbf{F}$ as inputs to generate new object query embeddings $\mathbf{Q}^{\mathrm{new}}$ and their position embeddings $\mathbf{Q}_{\mathrm{pos}}^{\mathrm{new}}$, feeding to the subsequent cross-attention module. The generated embeddings $\mathbf{Q}^{\mathrm{new}}$ are enforced to lie in the same embedding space with the encoded image features $\mathbf{F}$, which facilitates the subsequent matching process between them, making object queries able to quickly and properly attend to relevant regions in the encoded image features.

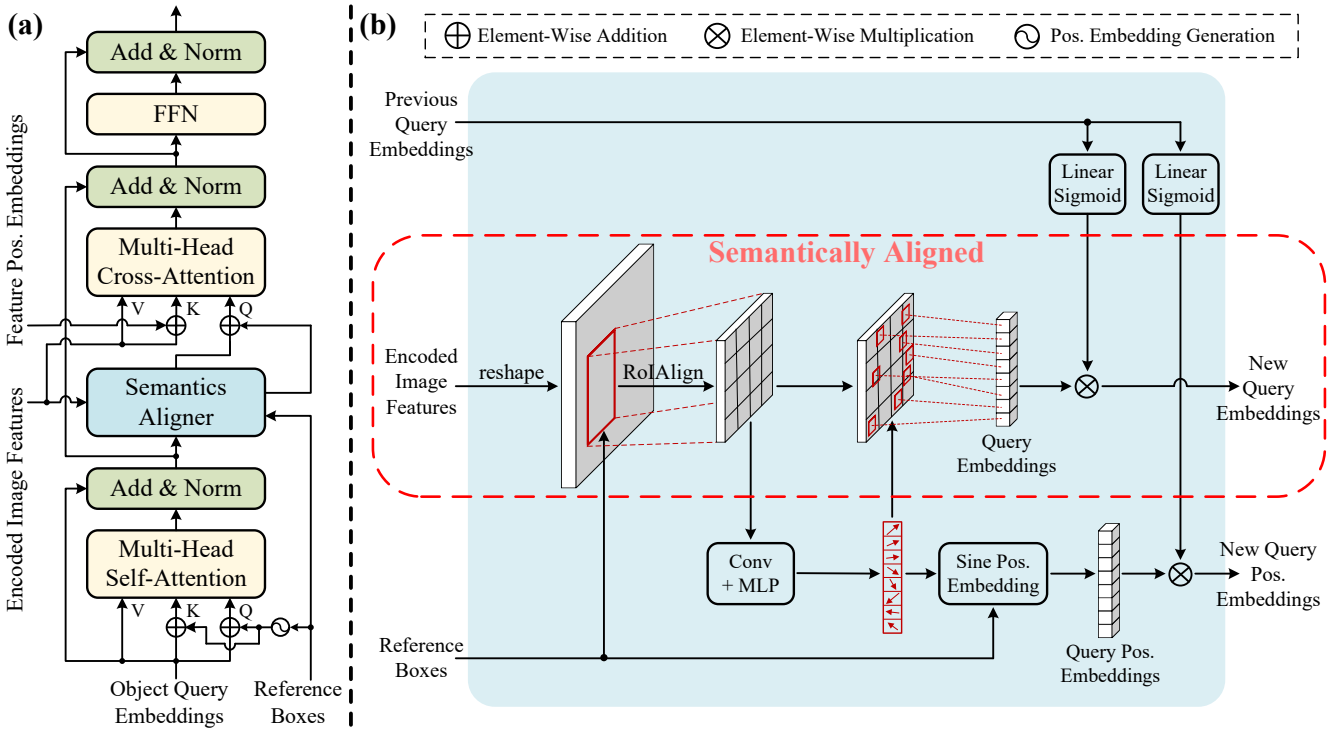

Figure 3. The proposed *Semantic-Aligned-Matching DETR (SAM-DETR)* appends a *Semantics Aligner* into the Transformer decoder layer. **(a) The architecture of one decoder layer in SAM-DETR.** It models a learnable *reference box* for each object query, whose center location is used to generate corresponding position embeddings. With the guidance of the reference boxes, Semantics Aligner generates new object queries that are semantically aligned with the encoded image features, thus facilitating their subsequent matching. **(b) The pipeline of the proposed *Semantics Aligner*.** For simplicity, only one object query is illustrated. It first leverages the reference box to extract features from the corresponding region via RoIAlign. The region features are then used to predict the coordinates of salient points with the most discriminative features. The salient points' features are then extracted as the new query embeddings with aligned semantics, which are further reweighted by previous query embeddings to incorporate useful information from them.

3.2.1 Semantic-Aligned Matching

As shown in Eq. 1 and Fig. 2, the cross-attention module applies dot-product to object queries and encoded image features, producing attention weight maps indicating the matching between object queries and target regions. It is intuitive to use dot-product since it measures similarity between two vectors, encouraging object queries to have higher attention weights for more similar regions. However, the original DETR [3] does not enforce object queries and encoded image features being semantically aligned, *i.e.*, projected into the same embedding space. Therefore, the object query embeddings are randomly projected to an embedding space at initialization, thus are almost equally matched to the encoded image features' all spatial locations. Consequently, extremely long training is needed to learn a meaningful matching between them.

With the above observation, the proposed Semantics Aligner designs a semantic alignment mechanism to ensure object query embeddings are in the same embedding space with encoded image features, which guarantees the dot-product between them is a meaningful measurement of similarity. This is accomplished by resampling object queries from the encoded image features based on the reference boxes, as shown in Fig. 3 (b). Given the encoded image features $\mathbf{F}$ and object queries' reference boxes $\mathbf{R}_{\text{box}}$, the Semantics Aligner first restores the spatial dimensions of the encoded image features from 1D sequences $HW \times d$ to 2D maps $H \times W \times d$. Then, it applies RoIAlign [12] to extract region-level features $\mathbf{F}_{\text{R}} \in \mathbb{R}^{N \times 7 \times 7 \times d}$ from the encoded image features. The new object queries $\mathbf{Q}^{\text{new}}$ and $\mathbf{Q}^{\text{new}}_{\text{pos}}$ are then obtained via resampling from $\mathbf{F}_{\text{R}}$. More details are to be discussed in the ensuing subsection.

$$\mathbf{F}_{\text{R}} = \text{RoIAlign}(\mathbf{F}, \mathbf{R}_{\text{box}}) \tag{2}$$

$$\mathbf{Q}^{\text{new}}, \mathbf{Q}^{\text{new}}_{\text{pos}} = \text{Resample}(\mathbf{F}_{\text{R}}, \mathbf{R}_{\text{box}}, \mathbf{Q}) \tag{3}$$

Since the resampling process does not involve any projection, the new object query embeddings $\mathbf{Q}^{\text{new}}$ share the exact same embedding space with the encoded image features $\mathbf{F}$, yielding a strong prior for object queries to focus on semantically similar regions.

3.2.2 Matching with Salient Point Features

Multi-head attention plays an indispensable role in DETR, which allows each head to focus on different parts and thus significantly strengthens its modeling capacity. Besides, prior works [3, 31, 62] have identified the importance of objects' most discriminative salient points in object detection. Inspired by these observations, instead of naively resampling by average-pooling or max-pooling, we propose to explicitly search for multiple salient points and employ their features for the aforementioned semantic-aligned matching. Such design naturally fits in the multi-head attention mechanism [45] without any modification.

Let us denote the number of attention heads as M, which is typically set to 8. As shown in Fig. 3 (b), after retrieving region-level features $\mathbf{F}_{\mathrm{R}}$ via RoIAlign, we apply a ConvNet followed by a multi-layer perception (MLP) to predict M coordinates $\mathbf{R}_{\mathrm{SP}} \in \mathbb{R}^{N \times M \times 2}$ for each region, representing the salient points that are crucial for recognizing and localizing the objects.

$$\mathbf{R}_{\mathrm{SP}} = \mathrm{MLP}(\mathrm{ConvNet}(\mathbf{F}_{\mathrm{R}})) \tag{4}$$

It is worth noting that we constrain the predicted coordinates to be within the reference boxes. This design choice has been empirically verified in Section 4.3. Salient points' features are then sampled from $\mathbf{F}_{\mathrm{R}}$ via bilinear interpolation. The M sampled feature vectors corresponding to the M searched salient points are finally concatenated as the new object query embeddings, so that each attention head can focus on features from one salient point.

$$\mathbf{Q}^{\mathrm{new}\prime} = \mathrm{Concat}(\{\mathbf{F}_{\mathrm{R}}[..., x, y, ...] \text{ for } x, y \in \mathbf{R}_{\mathrm{SP}}\}) \tag{5}$$

The new object queries' position embeddings are generated using sinusoidal functions with salient points' image-scale coordinates as input. Similarly, position embeddings corresponding to M salient points are also concatenated to feed to the subsequent multi-head cross-attention module.

$$\mathbf{Q}^{\mathrm{new}\prime}_{\mathrm{pos}} = \mathrm{Concat}(\mathrm{Sinusoidal}(\mathbf{R}_{\mathrm{box}}, \mathbf{R}_{\mathrm{SP}})) \tag{6}$$

3.2.3 Reweighting by Previous Query Embeddings

The Semantics Aligner effectively generates new object queries that are semantically aligned with encoded image features, but also brings one issue: previous query embeddings $\mathbf{Q}$ that contain valuable information for detection are not leveraged at all in the cross-attention module. To mitigate this issue, the proposed Semantics Aligner also takes previous query emebddings $\mathbf{Q}$ as inputs to generate reweighting coefficients via a linear projection followed by a sigmoid function. Through element-wise multiplication with the reweighting coefficients, both new query embeddings and their position embeddings are reweighted to highlight important features, thus effectively leveraging useful information from previous query embeddings. This process can be formulated as:

$$\mathbf{Q}^{\mathrm{new}} = \mathbf{Q}^{\mathrm{new}\prime} \otimes \sigma(\mathbf{Q}\mathbf{W}_{\mathrm{RW1}}) \tag{7}$$

$$\mathbf{Q}^{\mathrm{new}}_{\mathrm{pos}} = \mathbf{Q}^{\mathrm{new}\prime}_{\mathrm{pos}} \otimes \sigma(\mathbf{Q}\mathbf{W}_{\mathrm{RW2}}), \tag{8}$$

where $\mathbf{W}_{\mathrm{RW1}}$ and $\mathbf{W}_{\mathrm{RW2}}$ denote linear projections, $\sigma(\cdot)$ denotes sigmoid function, and $\otimes$ denotes element-wise multiplication.

3.3. Compatibility with SMCA-DETR

As illustrated in Fig. 3 (a), our proposed SAM-DETR only adds a plug-and-play module with slight computational overhead, leaving most other operations like the attention mechanism unchanged. Therefore, our approach can easily work with existing convergence solutions in a complementary manner to facilitate DETR's convergence further. We demonstrate the excellent compatibility of our approach by integrating it with SMCA-DETR [10], a state-of-the-art method to accelerate DETR's convergence.

SMCA-DETR [10] replaces the original cross-attention with Spatially Modulated Co-Attention (SMCA), which estimates the spatial locations of object queries and applies 2D-Gaussian weight maps to constrain the attention responses. In SMCA-DETR [10], both the center locations and the scales for the 2D-Gaussian weight maps are predicted from the object query embeddings. To integrate our proposed SAM-DETR with SMCA, we make slight modifications: we adopt the coordinates of M salient points predicted by Semantics Aligner as the center locations for the 2D Gaussian-like weight maps, and simultaneously predict the scales of weight maps from pooled RoI features. Experimental results demonstrate the complementary effect between our proposed approach and SMCA-DETR [10].

3.4. Visualization and Analysis

Fig. 4 visualizes the salient points searched by the proposed Semantics Aligner, as well as their attention weight maps generated from the multi-head cross-attention module. We also compare them with the original DETR's attention weight maps. Both models are trained for 12 epochs with ResNet-50 [13] as their backbones.

It can be observed that the searched salient points mostly fall within the target objects and typically are the most distinctive locations that are crucial for object recognition and localization. This illustrates the effectiveness of our approach in searching salient features for the subsequent matching process. Besides, as shown in the attention weight maps from different heads, the sampled features from each salient point can effectively match target regions and narrow down the search range as reflected by the area of attention maps. Consequently, the model can effectively and efficiently attend to the extremities of the target objects as

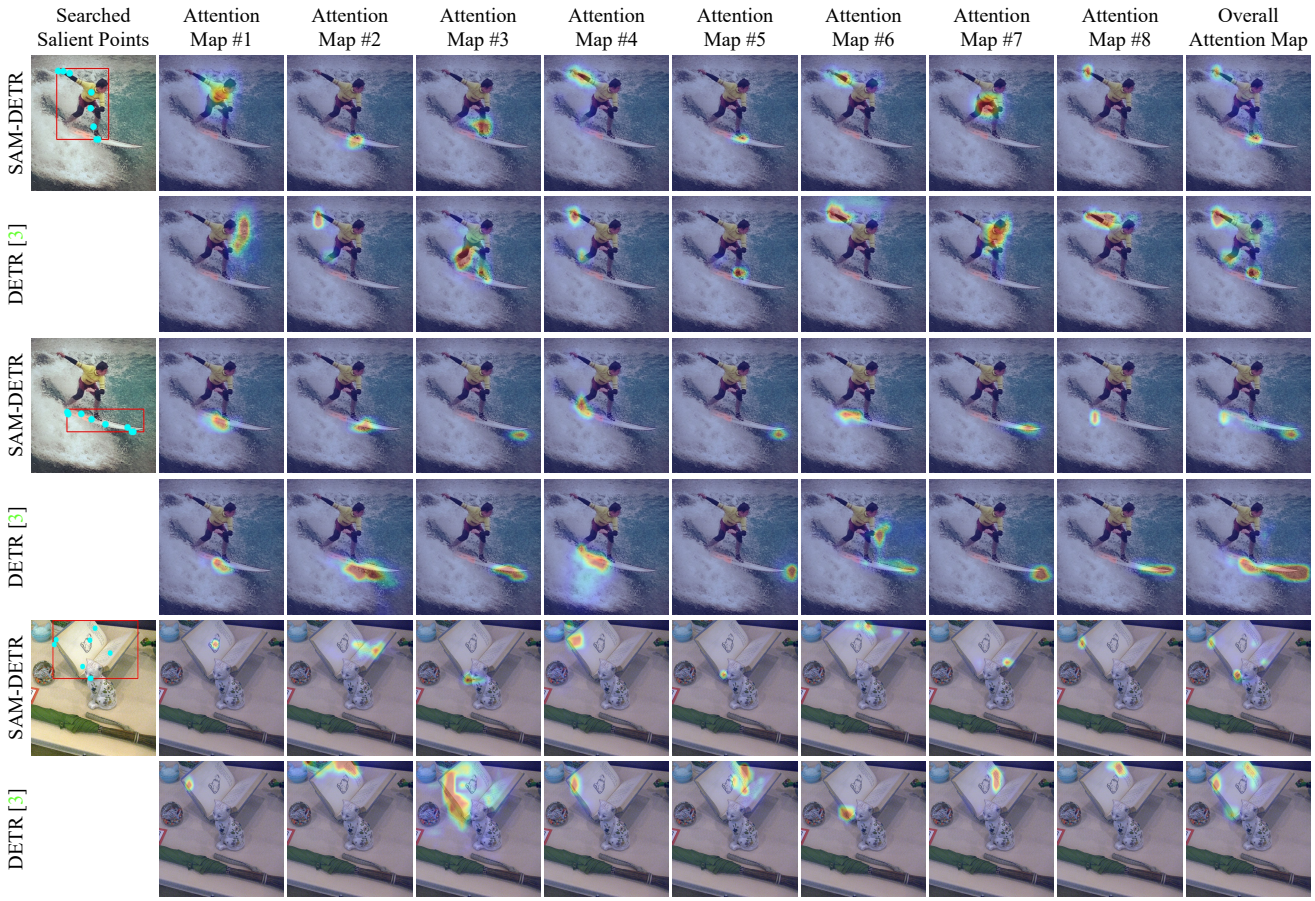

Figure 4. Visualization of SAM-DETR's searched salient points and their attention weight maps. The searched salient points mostly fall within the target objects and precisely indicate the locations with the most discriminative features for object recognition and localization. Compared with the original DETR, SAM-DETR's attention weight maps are more precise, demonstrating that our method effectively narrows down the search space for matching and facilitates convergence. In contrast, the original DETR's attention weight maps are more scattered, suggesting its inefficiency for matching relevant regions and distilling distinctive features.

shown in the overall attention maps, which greatly facilitates the convergence. In contrast, the attention maps generated from the original DETR are much more scattered, failing to locate the extremities efficiently and accurately. Such observation aligns with our motivation that the complication in matching object queries to target features is the primary reason for DETR's slow convergence. The visualization also proves the effectiveness of our proposed design in easing the matching difficulty via semantic-aligned matching and explicitly searched salient features.

4. Experiments

4.1. Experiment Setup

Dataset and Evaluation Metrics. We conduct experiments on the COCO 2017 dataset [26], which contains ∼117k training images and 5k validation images. Standard evaluation metrics for COCO are adopted to evaluate the performance of object detection.

Implementation Details. The implementation details of SAM-DETR mostly align with the original DETR [3]. We adopt ImageNet-pretrained [7] ResNet-50 [13] as the backbone, and train our model with 8 × Nvidia V100 GPUs using the AdamW optimizer [18, 30]. The initial learning rate is set as 1×10^{-5} for the backbone and 1×10^{-4} for the Transformer encoder-decoder framework, with a weight decay of 1×10^{-4}. The learning rate is decayed at a later stage by 0.1. The batch size is set to 16. When using ResNet-50 with dilations (R50-DC5), the batch size is 8. Model-architecture-related hyper-parameters stay the same with DETR, except we increase the number of object queries N from 100 to 300, and replace cross-entropy loss for classification with sigmoid focal loss [25]. Both design changes align with the recent works to facilitate DETR's convergence [10, 31, 63].

We adopt the same data augmentation scheme as DETR [3], which includes horizontal flip, random crop, and random resize with the longest side at most 1333 pixels and the shortest side at least 480 pixels.

Method	multi-scale	#Epochs	#Params (M)	GFLOPs	AP	$AP_{0.5}$	$AP_{0.75}$	AP_S	AP_M	AP_L
Baseline methods trained for long epochs:										
Faster-RCNN-R50-DC5 [35]		108	166	320	41.1	61.4	44.3	22.9	45.9	55.0
Faster-RCNN-FPN-R50 [24, 35]	✓	108	42	180	42.0	62.1	45.5	26.6	45.4	53.4
DETR-R50 [3]		500	41	86	42.0	62.4	44.2	20.5	45.8	61.1
DETR-R50-DC5 [3]		500	41	187	43.3	63.1	45.9	22.5	47.3	61.1
Comparison of SAM-DETR with other detectors under shorter training schemes:										
Faster-RCNN-R50 [35]		12	34	547	35.7	56.1	38.0	19.2	40.9	48.7
DETR-R50 [3] ‡		12	41	86	22.3	39.5	22.2	6.6	22.8	36.6
Deformable-DETR-R50 [63]		12	34	78	31.8	51.4	33.5	15.0	35.7	44.7
Conditional-DETR-R50 [31]		12	44	90	32.2	52.1	33.4	13.9	34.5	48.7
SMCA-DETR-R50 [10]		12	42	86	31.6	51.7	33.1	14.1	34.4	46.5
SAM-DETR-R50 (Ours)		12	58	100	33.1	54.2	33.7	13.9	36.5	51.7
SAM-DETR-R50 w/ SMCA (Ours)		12	58	100	36.0	56.8	37.3	15.8	39.4	55.3
Faster-RCNN-R50-DC5 [35]		12	166	320	37.3	58.8	39.7	20.1	41.7	50.0
DETR-R50-DC5 [3] ‡		12	41	187	25.9	44.4	26.0	7.9	27.1	41.4
Deformable-DETR-R50-DC5 [63]		12	34	128	34.9	54.3	37.6	19.0	38.9	47.5
Conditional-DETR-R50-DC5 [31]		12	44	195	35.9	55.8	38.2	17.8	38.8	52.0
SMCA-DETR-R50-DC5 [10]		12	42	187	32.5	52.8	33.9	14.2	35.4	48.1
SAM-DETR-R50-DC5 (Ours)		12	58	210	38.3	59.1	40.1	21.0	41.8	55.2
SAM-DETR-R50-DC5 w/ SMCA (Ours)		12	58	210	40.6	61.1	42.8	21.9	43.9	58.5
Faster-RCNN-R50 [35]		36	34	547	38.4	58.7	41.3	20.7	42.7	53.1
DETR-R50 [3] ‡		50	41	86	34.9	55.5	36.0	14.4	37.2	54.5
Deformable-DETR-R50 [63]		50	34	78	39.4	59.6	42.3	20.6	43.0	55.5
Conditional-DETR-R50 [31]		50	44	90	40.9	61.8	43.3	20.8	44.6	59.2
SMCA-DETR-R50 [10]		50	42	86	41.0	-	-	21.9	44.3	59.1
SAM-DETR-R50 (Ours)		50	58	100	39.8	61.8	41.6	20.5	43.4	59.6
SAM-DETR-R50 w/ SMCA (Ours)		50	58	100	41.8	63.2	43.9	22.1	45.9	60.9
Deformable-DETR-R50 [63]	✓	50	40	173	43.8	62.6	47.7	26.4	47.1	58.0
SMCA-DETR-R50 [10]	✓	50	40	152	43.7	63.6	47.2	24.2	47.0	60.4
Faster-RCNN-R50-DC5 [35]		36	166	320	39.0	60.5	42.3	21.4	43.5	52.5
DETR-R50-DC5 [3] ‡		50	41	187	36.7	57.6	38.2	15.4	39.8	56.3
Deformable-DETR-R50-DC5 [63]		50	34	128	41.5	61.8	44.9	24.1	45.3	56.0
Conditional-DETR-R50-DC5 [31]		50	44	195	43.8	64.4	46.7	24.0	47.6	60.7
SAM-DETR-R50-DC5 (Ours)		50	58	210	43.3	64.4	46.2	25.1	46.9	61.0
SAM-DETR-R50-DC5 w/ SMCA (Ours)		50	58	210	45.0	65.4	47.9	26.2	49.0	63.3
Accelerating DETR's convergence with self-supervised learning:										
UP-DETR-R50 [6]		150	41	86	40.5	60.8	42.6	19.0	44.4	60.0
UP-DETR-R50 [6]		300	41	86	42.8	63.0	45.3	20.8	47.1	61.7

Table 1. Comparison of the proposed SAM-DETR, other DETR-like detectors, and Faster R-CNN on COCO 2017 val set. ‡ denotes the original DETR [3] with aligned setups, including increased number of object queries (100→300) and focal loss for classification.

We adopt two training schemes for experiments, which include a 12-epoch scheme where the learning rate decays after 10 epochs, as well as a 50-epoch scheme where the learning rate decays after 40 epochs.

4.2. Experiment Results

Table 1 presents a thorough comparison of the proposed SAM-DETR, other DETR-like detectors [3, 6, 10, 31, 63], and Faster R-CNN [35]. As shown, Faster R-CNN and DETR can both achieve impressive performance when trained for long epochs. However, when trained for only 12 epochs, Faster R-CNN still achieves good performance, while DETR performs substantially worse due to its slow convergence. Several recent works [10, 31, 63] modify the original attention mechanism and effectively boost DETR's performance under the 12-epoch training scheme, but still have large gaps compared with the strong Faster R-CNN baseline. For standalone usage, our proposed SAM-DETR can achieve a significant performance gain compared with the original DETR baseline (+10.8% AP) and outperform all DETR's variants [10, 31, 63]. Furthermore, the proposed SAM-DETR can be easily integrated with existing

SAM	Query Resampling Strategy				RW	AP	$AP_{0.5}$	$AP_{0.75}$
	Avg	Max	SP x1	SP x8				
						22.3	39.5	22.2
✓	✓					25.2	48.9	23.3
✓		✓				27.0	50.2	25.8
✓			✓			28.6	50.3	28.1
✓			✓		✓	30.3	52.0	29.8
✓				✓		32.0	53.4	32.8
✓				✓	✓	33.1	54.2	33.7

Table 2. Ablation studies on our proposed design choices. Results are obtained on COCO val 2017. 'SAM' denotes the proposed Semantic-Aligned Matching. 'RW' denotes reweighting by previous query embeddings. Different resampling strategies for SAM are studied, including average-pooling (Avg), max-pooling (Max), one salient point (SP x1), and eight salient points (SP x8).

Salient Point Search Range		AP	$AP_{0.5}$	$AP_{0.75}$
within ref box	within image			
✓		33.1	54.2	33.7
	✓	30.0	52.3	29.2

Table 3. Ablation study on the salient point search range. Results are obtained on COCO val 2017.

convergence-boosting methods for DETR to achieve even better performance. Combining our proposed SAM-DETR with SMCA [10] brings an improvement of +2.9% AP compared with the standalone SAM-DETR, and +4.4% AP compared with SMCA-DETR [10], leading to performance on par with Faster R-CNN within 12 epochs. The convergence curves of the competing methods under the 12-epoch scheme are also presented in Fig. 1.

We also conduct experiments with a stronger backbone R50-DC5 and with a longer 50-epoch training scheme. Under various setups, the proposed SAM-DETR consistently improves the original DETR's performance and achieves state-of-the-art accuracy when further integrated with SMCA [10]. The superior performance under various setups demonstrates the effectiveness of our approach.

4.3. Ablation Study

We conduct ablation studies to validate the effectiveness of our proposed designs. Experiments are performed with ResNet-50 [13] under the 12-epoch training scheme.

Effect of Semantic-Aligned Matching (SAM). As shown in Table 2, the proposed SAM, together with any query resampling strategy, consistently achieves superior performance than the baseline. We highlight that even with the naive max-pooling resampling, $AP_{0.5}$ improves by 10.7%, a considerable margin. The results strongly support our claim that SAM effectively eases the complication in matching object queries to their corresponding target features, thus accelerating DETR's convergence.

Effect of Searching Salient Points. As shown in Table 2, different query resampling strategies lead to large variance in detection accuracy. Max-pooling performs better than average-pooling, suggesting that detection relies more on key features rather than treating all features equally. This motivates us to explicitly search salient points and use their features for semantic-aligned matching. Results show that searching just one salient point and resampling its features as new object queries outperforms the naive resampling strategies. Furthermore, sampling multiple salient points can naturally work with the multi-head attention mechanism, further strengthening the representation capability of the new object queries and boosting performance.

Searching within Boxes *vs.* Searching within Images. As introduced in Section 3.2.2, salient points are searched within the corresponding reference boxes. As shown in Table 3, searching salient points at the image scale (allowing salient points outside their reference boxes) degrades the performance. We suspect the performance drop is due to increased difficulty for matching with a larger search space. It is noteworthy that the original DETR's object queries do not have explicit search ranges, while our proposed SAM-DETR models learnable reference boxes with interpretable meanings, which effectively narrows down the search space, resulting in accelerated convergence.

Effect of Reweighting by Previous Embeddings. We believe previous object queries' embeddings contain helpful information for detection that should be effectively leveraged in the matching process. To this end, we predict a set of reweighting coefficients from previous query embeddings to apply to the newly generated object queries, highlighting critical features. As shown in Table 2, the proposed reweighting consistently boosts performance, indicating effective usage of knowledge from previous object queries.

4.4. Limitation

Compared with Faster R-CNN [35], SAM-DETR inherits from DETR [3] superior accuracy on large objects and degraded performance on small objects. One way to improve accuracy on small objects is to leverage multi-scale features, which we will explore in the future.

5. Conclusion

This paper proposes SAM-DETR to accelerate DETR's convergence. At the core of SAM-DETR is a plug-and-play module that semantically aligns object queries and encoded image features to facilitate the matching between them. It also explicitly searches salient point features for semantic-aligned matching. The proposed SAM-DETR can be easily integrated with existing convergence solutions to boost performance further, leading to a comparable accuracy with Faster R-CNN within 12 training epochs. We hope our work paves the way for more comprehensive research and applications of DETR.

References

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *ECCV*, 2016. 2, 3

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *CVPR*, 2018. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[4] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *CVPR*, 2021. 2, 3

[5] Dahjung Chung, Khalid Tahboub, and Edward J Delp. A two stream siamese convolutional neural network for person re-identification. In *ICCV*, 2017. 2, 3

[6] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. UP-DETR: Unsupervised pre-training for object detection with transformers. In *CVPR*, 2021. 7

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[8] Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *ECCV*, 2018. 3

[9] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-RPN and multi-relation detector. In *CVPR*, 2020. 2

[10] Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fast convergence of DETR with spatially modulated co-attention. In *ICCV*, 2021. 1, 2, 3, 5, 6, 7, 8

[11] Anfeng He, Chong Luo, Xinmei Tian, and Wenjun Zeng. A twofold siamese network for real-time object tracking. In *CVPR*, 2018. 3

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 4

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 6, 8

[14] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. TransReID: Transformer-based object re-identification. In *ICCV*, 2021. 3

[15] Ting-I Hsieh, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. One-shot object detection with co-attention and co-excitation. In *NeurIPS*, 2019. 2, 3

[16] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *CVPR*, 2018. 2

[17] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *ICCV*, 2019. 2

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[19] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, 2015. 2, 3

[20] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. SiamRPN++: Evolution of siamese visual tracking with very deep networks. In *CVPR*, 2019. 2, 3

[21] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, 2018. 2, 3

[22] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *CVPR*, 2021. 3

[23] Minghui Liao, Pengyuan Lyu, Minghang He, Cong Yao, Wenhao Wu, and Xiang Bai. Mask TextSpotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):532–548, 2021. 2

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 7

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 6

[26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 1, 6

[27] 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*, 128:261–318, 2020. 1

[28] Songtao Liu, Di Huang, and Yunhong Wang. Receptive field block net for accurate and fast object detection. In *ECCV*, 2018. 2

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *ECCV*, 2016. 2

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[31] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional DETR for fast training convergence. In *ICCV*, 2021. 1, 2, 3, 5, 6, 7

[32] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra R-CNN: Towards balanced learning for object detection. In *CVPR*, 2019. 2

[33] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M Hospedales, and Tao Xiang. Incremental few-shot object detection. In *CVPR*, 2020. 2

[34] Joseph Redmon and Ali Farhadi. YOLO 9000: Better, faster, stronger. In *CVPR*, 2017. 2

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1, 2, 7, 8

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 3

[37] Chen Shen, Zhongming Jin, Yiru Zhao, Zhihang Fu, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. Deep siamese network with multi-level similarity perception for person re-identification. In *ACM MM*, 2017. 2, 3

[38] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. Learning deep neural networks for vehicle Re-ID with visual-spatio-temporal path proposals. In *ICCV*, 2017. 2, 3

[39] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017. 2, 3

[40] Lingxue Song, Dihong Gong, Zhifeng Li, Changsong Liu, and Wei Liu. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In *ICCV*, 2019. 3

[41] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018. 2, 3

[42] Ran Tao, Efstratios Gavves, and Arnold WM Smeulders. Siamese instance search for tracking. In *CVPR*, 2016. 3

[43] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *ICCV*, 2019. 2

[44] Lachlan Tychsen-Smith and Lars Petersson. Improving object localization with fitness NMS and bounded IoU loss. In *CVPR*, 2018. 2

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, L. Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 3, 5

[46] Paul Voigtlaender, Jonathon Luiten, Philip HS Torr, and Bastian Leibe. Siam R-CNN: Visual tracking by re-detection. In *CVPR*, 2020. 2, 3

[47] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *CVPR*, 2021. 2, 3

[48] Lin Wu, Yang Wang, Junbin Gao, and Xue Li. Where-and-when to look: Deep siamese attention networks for video-based person re-identification. *IEEE Transactions on Multimedia*, 21(6):1412–1424, 2018. 2, 3

[49] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In *ECCV*, 2020. 2

[50] Chuhui Xue, Shijian Lu, Song Bai, Wenqing Zhang, and Changhu Wang. I2C2W: Image-to-character-to-word transformers for accurate scene text recognition. *arXiv preprint arXiv:2105.08383*, 2021. 2

[51] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta R-CNN: Towards general solver for instance-level low-shot learning. In *ICCV*, 2019. 2

[52] Fangao Zeng, Bin Dong, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. MOTR: End-to-end Multiple-Object tracking with TRansformer. *arXiv preprint arXiv:2105.03247*, 2021. 3

[53] Gongjie Zhang, Kaiwen Cui, Rongliang Wu, Shijian Lu, and Yonghong Tian. PNPDet: Efficient few-shot detection without forgetting via plug-and-play sub-networks. In *WACV*, 2021. 3

[54] Gongjie Zhang, Shijian Lu, and Wei Zhang. CAD-Net: A context-aware detection network for objects in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):10015–10024, 2019. 2

[55] Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, and Shijian Lu. Meta-DETR: Image-level few-shot object detection with inter-class correlation exploitation. *arXiv preprint arXiv:2103.11731*, 2021. 2, 3

[56] Jingyi Zhang, Jiaxing Huang, Zhipeng Luo, Gongjie Zhang, and Shijian Lu. DA-DETR: Domain adaptive detection transformer by hybrid attention. *arXiv preprint arXiv:2103.17084*, 2021. 2

[57] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Single-shot refinement neural network for object detection. In *CVPR*, 2018. 2

[58] Zhipeng Zhang and Houwen Peng. Deeper and wider siamese networks for real-time visual tracking. In *CVPR*, 2019. 3

[59] Meng Zheng, Srikrishna Karanam, Ziyan Wu, and Richard J Radke. Re-identification with consistent attentive siamese networks. In *CVPR*, 2019. 2, 3

[60] Changqing Zhou, Zhipeng Luo, Yueru Luo, Tianrui Liu, Liang Pan, Zhongang Cai, Haiyu Zhao, and Shijian Lu. PTTR: Relational 3D point cloud object tracking with transformer. In *CVPR*, 2022. 2

[61] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019. 2

[62] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *CVPR*, 2019. 2, 5

[63] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 1, 2, 3, 6, 7

[64] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *ECCV*, 2018. 3