Align-DETR: Enhancing End-to-end Object Detection with Aligned Loss

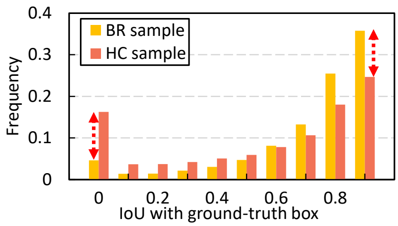
Zhi Cai[1,2]
caizhi97@buaa.edu.cn
Songtao Liu[3]
liusongtao@megvii.com
Guodong Wang[1,2]
wangguodong@buaa.edu.cn
Zheng Ge[3]
gezheng@megvii.com
Zeming Li[3]
lizeming@megvii.com
Xiangyu Zhang[3]
zhangxiangyu@megvii.com
Di Huang[1,2,†]
dhuang@buaa.edu.cn

[1] SKLCCSE, Beihang University
Beijing, China
[2] IRIP Lab, SCSE, Beihang University
Beijing, China
[3] Megvii Inc.

Abstract

DETR has set up a simple end-to-end pipeline for object detection by formulating this task as a set prediction problem, showing promising potential. Despite its notable advancements, this paper identifies two key forms of misalignment within the model: classification-regression misalignment and cross-layer target misalignment. Both issues impede DETR's convergence and degrade its overall performance. To tackles both issues simultaneously, we introduce a novel loss function, termed as Align Loss, designed to resolve the discrepancy between the two tasks. Align Loss guides the optimization of DETR through a joint quality metric, strengthening the connection between classification and regression. Furthermore, it incorporates an exponential down-weighting term to facilitate a smooth transition from positive to negative samples. Align-DETR also employs many-to-one matching for supervision of intermediate layers, akin to the design of $\mathcal{H}$-DETR , which enhances robustness against instability. We conducted extensive experiments, yielding highly competitive results. Notably, our method achieves a 49.3% $(+0.6)$ AP on the $\mathcal{H}$-DETR baseline with the ResNet-50 backbone. It also sets a new state-of-the-art performance, reaching 50.5% AP in the $1\times$ setting and 51.7% AP in the $2\times$ setting, surpassing several strong competitors. Our code is available at https://github.com/FelixCaae/AlignDETR.

† Indicates corresponding author.



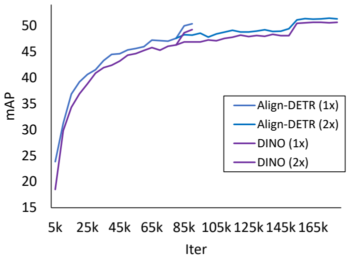


Figure 1: **Left:** Intersection over Union (IoU) distribution of two types of samples. There is a notable gap between best regressed samples (oracle) and the high confident samples, indicating a discrepancy between these two tasks. **Right:** The convergence curve of Align-DETR and DINO where Align-DETR converges faster significantly.

1 Introduction

Recently, transformer-based methods have garnered significant attention in the object detection community, largely due to the introduction of the DETR paradigm by [1]. Unlike previous CNN-based detectors [24, 33, 48, 53], DETR approaches object detection as a set prediction problem, utilizing learnable queries to represent each object in one-to-one correspondence. Such unique correspondence derives from bipartite graph matching by means of label assignment during training. It bypasses hand-crafted components such as non-maximum suppression (NMS) and anchor generation. With this simple and extensible pipeline, DETR shows great potential in a wide variety of areas, including 2D segmentation [3, 4, 19], 3D detection [27, 30, 41], in addition to 2D detection [8, 13, 26, 37, 45, 54].

During the past few years, the successors have advanced DETR in many ways. For instance, some methods attempt to incorporate local operators, such as ROI pooling [37] or deformable attention [8, 54], to increase the convergence speed and reduce the computational cost; some methods indicate that those learnable queries can be improved through extra physical embeddings [25, 29, 40]; and some methods [2, 14, 17, 45] notice the defect of one-to-one matching and introduce more positive samples by adding training-only queries. Box refinement [37, 45, 54] is another helpful technique, which explicitly takes previous predictions as priors at the next stages.

Despite the recent progress in DETR-based detectors [1, 8, 13, 22, 25, 45, 54], the misalignment problem of DETR has received insufficient attention. There are two key aspects to this misalignment issue in recent DETR-like methods. Firstly, there exists a misalignment between classification confidence and localization precision, stemming from inconsistent loss design. This discrepancy is highlighted through an analysis conducted on the output of a prominent end-to-end detector, DINO [45], revealing a significant dissonance between high-confidence samples (HC samples) and best-regressed samples (BR samples), as depicted in Fig. 1 (Left). Such a discrepancy significantly impacts model performance, particularly in ranking-based metrics such as mean average precision (mAP). Secondly, there is a misalignment in training targets across layers. This arises from the dynamic matching design of DETR [17, 26, 45], wherein samples are assigned different targets in different layers, leading to confusion within the optimizer, as highlighted by Stable-DINO [26]. These misalignment issues impede the convergence of DETR and hinder it from realizing its full potential as shown in Fig. 1 (Right).

The current solutions to the misalignment problem in DETR-like methods typically address either the first misalignment issue [15, 21, 46] or the second [26, 45]. To tackle both simultaneously, we introduce a novel approach called Align-DETR. It makes use of the standard focal loss [24] with an IoU-aware target on foreground samples, which we term the Align Loss. To overcome the first misalignment problem, Align Loss dynamically adjusts the target for foreground samples according to their classification confidence and regression precision thus they are aligned during optimization [26]. For the second problem, Align-DETR enlarges the range of positive samples by adopting a mixed-matching strategy. This approach allows multiple candidates to be considered for each ground truth. Subsequently, to mitigate conflicts arising from this expanded range of positive samples, the targets of the additional positive samples are smoothed using an exponential weight decay. By incorporating these mechanisms, Align-DETR aims to effectively address both misalignment issues encountered in DETR-based detectors.

Overall, Align-DETR offers a straightforward yet effective solution to the misalignment problem, enhancing DETR with aligned training targets. Equipped with a ResNet-50 [11] backbone and a $\mathcal{H}$-DETR [14] baseline, our method achieves +0.6% AP gain. We also combine it with the strong baseline DINO [45] and establish a new state-of-the-art performance with 50.5% AP in the $1\times$ and 51.7% AP in $2\times$ setting on the COCO [23] validation set.

2 Related Work

2.1 Label Assignment in Object Detection

As the CNN-based object detectors develop from the anchor-based framework to the anchor-free one, many works realize the importance of label assignment (which is previously hidden by anchor and IoU matching) during training. Some works [9, 16, 48] identify positive samples by measuring their dynamic prediction quality for each object. Others [7, 20, 21, 53] learn the assignment in a soft way and achieve better alignment on prediction quality by incorporating IoU [21, 46] or a combination of IoU and confidence [7, 20, 53].

The misalignment problem in object detection has been addressed by various traditional solutions, such as incorporating an additional IoU branch to fine-tune the confidence scores [15] or integrating the IoU prediction branch into classification losses [21]. In contrast, the misalignment problem in DETR is under-explored, despite it sharing some ideas from CNN-based detectors. However, the optimization target between many-to-one and one-to-one is the key difference as we will illustrate in the next section.

2.2 End-to-end Object Detection

The pursuit of end-to-end object detection or segmentation dates back to several early efforts [32, 35, 36]. They rely on recurrent neural network (RNN)[35] to remove duplicates or adopt complex subnets [32, 36] to replace NMS. Different from them, DETR [1] has established a set-prediction framework based on the transformer [38]. Compared to previous work, DETR is rather simpler but still suffers from the downside of slow convergence with a number of subsequent DETR variants [6, 8, 17, 25, 37, 45] working on this issue. Some methods make improvements on the cross-attention in decoders [6, 29]. Deformable DETR [54] presents a deformable-attention module that only scans a small set of points near the reference point, while AdaMixer[8] further extends the 2D offset to 3D for better multi-scale feature fusion.

Besides, some works also pay attention to improve the inference efficiency of DETR [18, 39, 42, 51, 52]. Efficient DETR [42] advocates for a 1-layer-only decoder structure to largely reduce the computation burden by initializing the query precisely. Notably, RT-DETR [51] represents a significant advancement by enabling real-time inference for DETR, surpassing the performance of other rapid detectors such as the YOLO series series [10].

The optimization of DETR also attracts the attention of many researchers [17, 26, 45]. Specifically, DN-DETR [17] relies on a denoising mechanism to stabilize the training, which is further refined by DINO [45] through introducing a contrastive denoising mechanism. Additionally, Stable-DINO [26] introduces a position-guided loss that mitigates the instability incurred by a standard loss, *i.e.* focal loss[24]. Meanwhile, a few recent studies have noticed limitations of one-to-one matching and have proposed many-to-one assigning strategies to ameliorate DETR regarding training efficiency. Group-DETR [2] and $\mathcal{H}$-DETR [14] accelerate the training process with multiple groups of samples and ground truths. DAC-DETR[13] proposes a decoupled training strategy that focuses on the learning of cross-attention layers with many-to-one matching.

Despite the strides made, it is evident that many contemporary approaches [2, 13, 14, 17, 45] either overlook the misalignment issue highlighted earlier or offer only partial remedies [26]. In contrast to these approaches, our work offers a comprehensive and unified solution to address this challenge consistently.

3 Method

3.1 Preliminaries

DETR. The original DETR [1] framework consists of three main components: a CNN-backbone, an encoder-decoder transformer [38], and a prediction head. The backbone processes the input image first, and the resulting feature is flattened into a series of tokens $X = \{x_1, x_2, ..., x_m\}$. Then the transformer extracts information from X with a group of learnable queries $Q = \{q_1, q_2, ..., q_n\}$ as containers. At last, the updated queries are transformed into predictions $P = \{p_1, p_2, ..., p_n\}$ through the prediction head. In most cases, m is much less than n, making DETR a sparse object detection pipeline.

The focal loss [24] is adopted by DETR in classification optimization to help focus on important samples. Given a binary label $y \in \{0, 1\}$ and a logit $p \in [0, 1]$, it is defined as:

$$\mathcal{L}_{focal} = -y \cdot (1-p)^{\gamma} \cdot logp - (1-y) \cdot p^{\gamma} \cdot log(1-p), \tag{1}$$

where γ is the hyper-parameter to control the degree of weight decay.

DETR adopts the one-to-one label assignment on all layers to help eliminate redundant predictions. However, this strategy is inefficient compared to many-to-one label assignment used in CNN-based detectors[33, 48, 53]. To overcome this issue, $\mathcal{H}$-DETR[14] proposes a hybrid layer matching strategy that applies many-to-one matching on some shallow layers and one-to-one matching on deep layers. Hybrid matching ensures DETR's final outputs are unique while it allows more efficient training on intermediate layers.

3.2 Motivation and Framework

The motivation for the proposed Align-DETR comes from the hypothesis that a consistent and aligned optimization target can benefit the training of object detectors like DETR [17, [26, 45, 54]. There are two concerns for the current optimization method: (i) the alignment between classification and regression is essential for the optimization of DETR, which is not considered in current design and (ii) the matching mechanism of DETR is unstable across layers. To mitigate the concerns, we propose a unified solution, namely Align-DETR.

We illustrate our framework in Fig. 2 and introduce the detailed implementations in the following sections. Overall, our key insight is to design a dynamic and accurate training target for DETR. For the first concern, we build a strong connection between the classification and regression by adopting a regression-aware classification loss. To mitigate the second issue , we adopt many-to-one matching along with a ranking & weighting strategy. In this way, both of the two misalignment issues can be solved jointly.

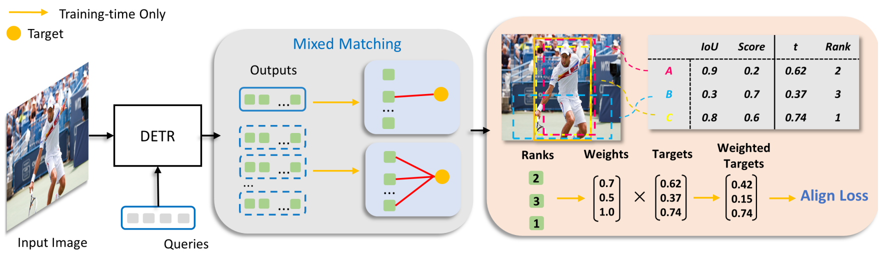


Figure 2: The architecture overview of the proposed approach Align-DETR . Align-DETR adopts many-to-one matching where each GT is assigned multiple queries. These queries are sorted according to their quality . Then, we compute an alignment score for each query according to their rank, classification confidence and IoU with the GT. The alignment score is used in the loss computation for both classification and regression.

3.3 Align-DETR

Driven by the aforementioned concerns, our objective is to enhance the optimization of DETR by addressing the misalignment issue. Initially, we present our matching strategy, followed by the introduction of our proposed loss function, denoted as the Align Loss. This sequential approach is aimed at systematically mitigating misalignment and thereby improving the overall efficacy of DETR optimization.

Mixed Matching and Ranking Strategy. DETR [1] and most of its variants [29, 54] adopt Hungarian Matching to learn a unique association between GT and predictions. However, this approach assigns only one positive sample for each GT annotation, rendering it susceptible to the instability inherent in matching, as noted in previous works [26, 45]. To address this challenge, we propose a gradual transition from positive to negative samples which involves implementing a mixed-matching and ranking strategy.

Given predictions P and ground truth G, each comprising N instances, we employ a modified version of Hungarian Matching to assign k predictions to each ground truth, resulting in a total of kN matched samples, termed candidates. These candidates are subsequently arranged based on their distances from the GT. We propose defining a quality metric q as inspired previous studies [5, 53], which represents the geometric average of classification accuracy (p) and regression precision (u):

$$q = p^{\alpha} \cdot u^{(1-\alpha)}, \tag{2}$$

where p denotes the binary classification score, u signifies the IoU between the predicted bounding box and the ground truth, and α serves as a hyper-parameter to balance these factors. We denote the ranking of each candidate as $r \in \{0,1,2,3,k-1\}$. We set $k > 1$ for intermediate predictions and expect the change of matching happens within a candidate bag. As for the last decoder layer, we set $k = 1$ for one-to-one association.

Align-DETR shares some similarities with $\mathcal{H}$-DETR [14] but diverges in both motivation and implementation: **(a)** While $\mathcal{H}$-DETR utilizes many-to-one matching primarily to expedite convergence, we employ it to ensure consistent optimization across layers; **(b)** $\mathcal{H}$-DETR treats all positive samples equally, whereas we introduce an adaptive target mechanism, as detailed in Section 3.3.

Align Loss. To promote more consistent and efficient optimization, we outline two guiding principles to inform the loss design of DETR. Firstly, the target of the classification loss should be adaptive and position-guided, echoing findings in prior literature [26]. Secondly, there should be a smooth transition from positive samples to negative samples.

In accordance with these principles, we propose a straightforward yet effective loss function for DETR, defined as follows:

$$\mathcal{L}_{align} = -t_c \cdot (1-p)^{\gamma} \cdot logp - (1-t_c) \cdot p^{\gamma} \cdot log(1-p), \tag{3}$$

wherein the hard label y in Eq. 1 is substituted with a soft target t_c. As shown in Eq. 3, adjusting t_c from 1 to 0 makes a smooth transition from a positive target to a negative target. This property makes it perfectly compatible to achieve a transition between positive sample and negative sample. We define t_c as follows :

$$t_c = e^{-r/\tau} \cdot q, \tag{4}$$

where $e^{-r/t}$ τ is an exponential down-weighting term controlled by a hyper-parameter τ. By associating the t_c with the joint quality, Align Loss can guide the learning of classification with regression precision simultaneously and thus build a strong connection between these two tasks [6, 49]. In the literature, [26] employs a position-supervised classification loss to establish a unified optimization framework, which bears similarities to our approach. However, our method approaches the problem from a distinct perspective, emphasizing the alignment of the two tasks. Thus, we utilize the classification confidence in t_c, which our experiments in Section 4 have validated as crucial. Another noteworthy distinction lies in our identification of misalignment in the classification target across layers due to the unstable matching phenomenon. This issue is addressed in our method through a gradual positive-to-negative transition.

Given that the Align Loss functions as a "soft" variant of focal loss [24], seamlessly integrating with any DETR-variant compatible with focal loss is feasible. To leverage this capability, we propose an asymmetric classification loss by applying Align Loss on selected candidates and focal loss on background samples:

$$\mathcal{L}_{cls} = \sum_{i}^{N_{pos}} \mathcal{L}_{align}(p_i, y_i) + \sum_{j}^{N_{neg}} \mathcal{L}_{focal}(p_j, 0), \tag{5}$$

where N_{pos} and N_{neg} denote the number of total positive samples and negative samples, respectively.

In the context of regression tasks, though not obviously influenced by the misalignment issue aforementioned, we opt to implement a regression loss consistent with Align Loss.

This helps achieves a consistent optimization in both tasks. Given predicted bounding box b_i and GT box $\hat{b}_i$, our regression loss is defined as follows:

$$\mathcal{L}_{reg} = \sum_{i}^{N_{pos}} e^{(-r_i/\tau)} \cdot (\mathcal{L}_{l1}(b_i, \hat{b}_i) + \mathcal{L}_{GIoU}(b_i, \hat{b}_i)) \quad (6)$$

Ultimately, our loss is defined as:

$$\mathcal{L} = \sum_{l=1}^{L-1} \mathcal{L}_{task}(P_l, G^{(k)}) + \mathcal{L}_{task}(P, G), \quad (7)$$

where $\mathcal{L}_{task}$ is a weighted combination of classification loss $\mathcal{L}_{cls}$ and $\mathcal{L}_{reg}$ [1], $G^{(k)}$ is an augmented version of GT by copying k times and L is the total number of decoder layers.

In summary, the Align-DETR introduces an Align Loss along with a matching strategy to solve the misalignment issue for higher precision on localization of DETR. Without loss of generality, our method can be integrated into any DETR-like architecture.

4 Experiments

4.1 Setup

Datasets. We conduct all our experiments on MS-COCO 2017 [23] Detection Track and report our results with the mean average precision metric on the validation dataset.

Implementation details. We use DINO [45] as the baseline method, along with their default hyper-parameter settings. The DINO baseline adopts deformable-transformer [54] and multi-scale features as inputs. For the hyper-parameters introduced in Align-DETR, we set $k = 4$, $\alpha = 0.25$, and $\tau = 1.5$. To ensure a fair comparison with recent methods [13, 26, 45], we train Align-DETR for $1\times$ and $2\times$ schedules. We implement our methods with the help of open-source library detrex [34]. To optimize the model, we set the initial learning rate to 1×10^{-4} and decay it by multiplying 0.1 for backbone learning. We use AdamW [28] as the optimizer with 1×10^{-4} weight decay and set batch size to 16 for all our experiments.

4.2 Main Results

We conduct experiments using DINO [45] and $\mathcal{H}$-DETR [14] as the baselines, which adopts the deformable-transformer as the backbone. DINO uses tricks such as CDN, look forward-twice, and bounding box refinement for better performance. We follow DINO's approach and adopt its tricks. Regarding to backbone, we use an Resnet-50 (R-50) [11] backbone with 4-scale features (P3, P4, P5, and P6) as input.

The results are presented in Tab. 1 and Tab. 2. Despite the highly optimized structure of DINO [45], our method still outperforms it by 1.5% and 1.3% AP in $1\times$ and $2\times$ schedules, respectively. This indicates that even the advanced DETR-variant can be affected by the misalignment problem. Then we compare Align-DETR to two recent state-of-the-art methods, DAC-DETR [13] and Stable-DINO [26], and find that Align-DETR achieves higher AP while using fewer tricks such as memory fusion, demonstrating its superior effectiveness. It's noteworthy that Align-DETR exhibits highly competitive performance, particularly in detecting small objects. In this domain, it surpasses Stable-DINO by 1.8% in terms of AP, indicating that small objects are more susceptible to the misalignment issue. This

Model	#epochs	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
SMCA-DETR [6]	50	R50	43.7	63.6	47.2	24.2	47.0	60.4
SAM-DETR [44]	50	R50	45.0	65.4	47.9	26.2	49.0	63.3
Def.DETR [54]	50	R50	45.4	64.7	49.0	26.8	48.3	61.7
AdaMixer[8]	36	R50	47.0	66.0	51.1	30.1	50.2	61.8
SD-DETR [47]	50	R50	45.5	65.4	48.5	25.6	49.9	64.2
DAB-Def.DETR [25]	50	R50	46.9	66.0	50.8	30.1	50.4	62.5
DN-Def.DETR [17]	12	R50	43.4	61.9	47.2	24.8	46.8	59.4
DN-Def.DETR [17]	50	R50	48.6	67.4	52.7	31.0	52.0	63.7
DINO [45]	12	R50	49.0	66.6	53.5	32.0	52.3	63
DINO [45]	24	R50	50.4	68.3	54.8	33.3	53.7	64.8
Co-DETR [55]	12	R50	49.5	67.6	54.3	32.4	52.7	63.7
Cascade-DETR [43]	12	R50	49.7	67.1	54.1	32.4	53.5	65.1
Group-DETR [2]	12	R50	49.8	−−	−−	32.4	53.0	64.2
$\mathcal{H}$-DETR [14]	12	R50	48.7	66.4	52.9	31.2	51.5	63.5
DAC-DETR [13]	12	R50	50.0	67.6	54.7	32.9	53.1	64.2
DAC-DETR [13]	24	R50	51.2	68.9	56.0	34.0	54.6	65.4
Salience-DETR [12]	12	R50	50.0	67.7	54.2	33.3	54.4	64.4
Salience-DETR [12]	24	R50	51.2	68.9	55.7	33.9	55.5	65.6
Rank-DETR [31]	12	R50	50.2	67.7.	55.0	34.1	53.6	64.0
MS-DETR [50]	12	R50	50.0	67.3	54.4	31.6	53.2	64.0
MS-DETR [50]	24	R50	50.9	68.4	56.1	34.7	54.3	65.1
Focus-DETR [52]	36	R50	50.4	68.5	55.0	34.0	53.5	64.4
Stable-DINO [26]	12	R50	50.4	67.4	55.0	32.9	**54.0**	**65.5**
Stable-DINO [26]	24	R50	51.5	68.5	<u>56.3</u>	35.2	54.7	<u>66.5</u>
Align-DETR (Ours)	12	R50	**50.5**	**67.7**	**55.3**	**34.7**	53.6	64.6
Align-DETR (Ours)	24	R50	<u>51.7</u>	<u>69.0</u>	<u>56.3</u>	<u>35.5</u>	<u>55.0</u>	66.1

Table 1: Comparisons (%) of Align-DETR and other DETR-like methods on COCO *val* set. Def.DETR is the abbreviation of Deformable DETR. **Bold** and <u>underlined</u> text are best results under $1\times$ and $2\times$ schedule setting, respectively.

highlights the effectiveness of Align-DETR in addressing the challenges posed by misalignment, particularly in scenarios where precise localization is crucial, such as detecting small objects. Align-DETR also outperforms other competitors such as SMCA [6], Faster RCNN-FPN [33], Deformable-DETR [54] and Focus-DETR [52] with much less training schedule. At last, we compare our methods to two DETR-variants that also focus on improving the assignment of DETR, *i.e.* $\mathcal{H}$-DETR [14] and Group-DETR [2], and find Align-DETR leads them by a large margin of 1.8% AP and 0.7% AP, respectively, with fewer queries used in training. These results suggest that Align-DETR is a highly effective and efficient method for object detection tasks.

4.2.1 Comparison with Related Methods

In addition to comparison with state-of-the-art DETR-variants, we also implement methods like Quality Focal Loss (QFL) [21], and Varifocal Loss (VFL) [46], Position-Supervised Loss (PSL) [26] on DINO [45], and the results are presented in Tab. 3. Interestingly, we find that the the IoU-branch [15], which is a widely adopted component in CNN-based detectors [10], brings limited improvement to the performance. QFL [21] and VFL [46] also perform poorly in our experiments, which suggests that they are not designed for end-to-end

Method	w/ Align Loss	AP	AP_{50}	AP_{75}
$\mathcal{H}$-DETR [14]		48.7	66.4	52.9
Align-$\mathcal{H}$-DETR	✓	**49.3**	**67.2**	**53.7**

Table 2: Comparisons (%) of Align-$\mathcal{H}$-DETR and $\mathcal{H}$-DETR on COCO *val* set with 1x schedule.

Method	AP	AP_{50}	AP_{75}
Focal Loss [24]	49.0	66.0	53.5
IoU branch [15]	49.2	66.3	53.5
QFL [21]	47.6	64.3	51.8
VFL [46]	48.7	67.0	52.3
PSL [26]	49.8	66.7	54.5
PSL + PMC [26]	50.2	66.7	55.0
Align Loss (Ours)	**50.5**	**67.8**	**55.3**

Table 3: Comparison (%) with other methods on the misalignment problem on COCO *val*. We use "PSL" and "PMC" for position-supervised loss and position-modulated matching in Stable-DINO [26]

detectors. Compared to the most closely related method, PSL [26], Align Loss demonstrated a significant improvement of 0.7% AP. Even when augmented with PMC, PSL still falls short of matching the performance of Align Loss. We attribute this discrepancy to PSL's focus on optimizing paths individually for each layer, without addressing the issue of misalignment across layers. This is likely a contributing factor to the superior performance of our method.

4.3 Ablation Study

We conduct a series of ablation studies with DINO baseline to validate the effectiveness of the components. All experiments here use an R50 backbone and a schedule of standard 1x training schedule.

Firstly, we validate the effectiveness of the proposed loss design and the results are summarized in Tab. 4. It is observed that both classification loss and regression loss contribute to the final performance, with the primary contribution stemming from the classification loss,

Cls Loss	Reg Loss	Matching	AP	AP_{50}	AP_{75}
✓	✓	✓	**50.5**	**67.8**	**55.3**
✓	✓		50.1	67.2	54.8
✓		✓	49.7	66.9	54.1
	✓	✓	49.1	67.5	53.4
		✓	49.0	66.0	53.5

Table 4: Ablation study (%) of Align-DETR on each component in terms of AP on COCO *val*. The results demonstrate the effectiveness of our proposed component.

α	0	0.25	0.5	0.75
AP	50.0	**50.5**	49.2	47.6

k	1	2	3	4	5
AP	50.1	50.2	50.4	**50.5**	50.2

τ	1.5	3	6	9
AP	**50.5**	50.1	50.0	49.7

Table 5: Influence (%) of hyper-paramters α, k and τ on our approach on COCO *val*.

as anticipated. Notably, when the regression loss is deprecated, the performance experiences a 0.8% AP drop, underscoring the importance of consistency in Eq.7 in the loss design. To further investigate the impact of the hyper-parameters we introduced, *i.e.* α, k and τ, we conduct sensitivity analysis by changing one variable and keeping other variables controlled. Our default values are $k = 4$, $\alpha = 0.25$, and $\tau = 1.5$. As shown in Tab. 5, α is has the greatest influence on the performance while τ and k have moderate effects. This sensitivity analysis supports our hypothesis that α should be kept small to prevent effective training signals from suppression.

5 Conclusion

This paper investigates the optimization of DETR and identifies two aspects of the misalignment issue that could impede performance. To address these challenges, we propose a unified and straightforward solution named Align-DETR, comprising a many-to-one matching strategy and a novel loss function, referred to as Align Loss. To mitigate the side effects of misaligned targets across layers, our matching strategy expands the number of samples assigned to a ground truth, which we term as candidates. We anticipate the matching changes to occur within a group of candidates. The Align Loss is designed as a "soft" variant of focal loss, employing a quality metric to guide the learning of classification with respect to position. Additionally, we implement a gradual transition from positive to negative samples within a group of candidates to smooth the conflict caused by matching change. Competitive experimental results are achieved on the common COCO benchmark, demonstrating the superiority of Align-DETR in terms of effectiveness.

Acknowledgements

This work is partly supported by the National Natural Science Foundation of China (No. 62022011), the Research Program of State Key Laboratory of Complex and Critical Software Environment, and the Fundamental Research Funds for the Central Universities.

References

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

[2] Qiang Chen, Xiaokang Chen, Jian Wang, Shan Zhang, Kun Yao, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. Group detr: Fast detr training with group-wise one-to-many assignment. In *ICCV*, 2023.

[3] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021.

[4] Bin Dong, Fangao Zeng, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Solq: Segmenting objects by learning queries. In *NeurIPS*, 2021.

[5] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *ICCV*, pages 3490–3499, 2021.

[6] Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fast convergence of detr with spatially modulated co-attention. In *ICCV*, 2021.

[7] Ziteng Gao, Limin Wang, and Gangshan Wu. Mutual supervision for dense object detection. In *CVPR*, pages 3641–3650, 2021.

[8] Ziteng Gao, Limin Wang, Bing Han, and Sheng Guo. Adamixer: A fast-converging query-based object detector. In *CVPR*, 2022.

[9] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *CVPR*, 2021.

[10] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

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

[12] Xiuquan Hou, Meiqin Liu, Senlin Zhang, Ping Wei, and Badong Chen. Salience detr: Enhancing detection transformer with hierarchical salience filtering refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17574–17583, June 2024.

[13] Zhengdong Hu, Yifan Sun, Jingdong Wang, and Yi Yang. Dac-detr: Divide the attention layers and conquer. In *NeurIPS*, 2024.

[14] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. In *CVPR*, 2023.

[15] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *ECCV*, pages 784–799, 2018.

[16] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. In *ECCV*, 2020.

[17] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *CVPR*, 2022.

[18] Feng Li, Ailing Zeng, Shilong Liu, Hao Zhang, Hongyang Li, Lei Zhang, and Lionel M Ni. Lite detr: An interleaved multi-scale encoder for efficient detr. In *CVPR*, pages 18558–18567, 2023.

[19] Feng Li, Hao Zhang, Huaizhe xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *CVPR*, 2023.

[20] Shuai Li, Chenhang He, Ruihuang Li, and Lei Zhang. A dual weighting label assignment scheme for object detection. In *CVPR*, 2022.

[21] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *NeurIPS*, 2020.

[22] Junyu Lin, Xiaofeng Mao, Yuefeng Chen, Lei Xu, Yuan He, and Hui Xue. Dˆ 2etr: Decoder-only detr with computationally efficient cross-scale attention. *arXiv preprint arXiv:2203.00860*, 2022.

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

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[25] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. In *ICLR*, 2022.

[26] Shilong Liu, Tianhe Ren, Jiayu Chen, Zhaoyang Zeng, Hao Zhang, Feng Li, Hongyang Li, Jun Huang, Hang Su, Jun Zhu, et al. Detection transformer with stable matching. In *ICCV*, pages 6491–6500, 2023.

[27] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *ECCV*, 2022.

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

[29] 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.

[30] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *ICCV*, 2021.

[31] Yifan Pu, Weicong Liang, Yiduo Hao, Yuhui Yuan, Yukang Yang, Chao Zhang, Han Hu, and Gao Huang. Rank-detr for high quality object detection. In *NeurIPS*, 2023.

[32] Mengye Ren and Richard S Zemel. End-to-end instance segmentation with recurrent attention. In *CVPR*, 2017.

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[34] Tianhe Ren, Shilong Liu, Feng Li, Hao Zhang, Ailing Zeng, Jie Yang, Xingyu Liao, Ding Jia, Hongyang Li, He Cao, Jianan Wang, Zhaoyang Zeng, Xianbiao Qi, Yuhui Yuan, Jianwei Yang, and Lei Zhang. detrex: Benchmarking detection transformers, 2023.

[35] Amaia Salvador, Miriam Bellver, Victor Campos, Manel Baradad, Ferran Marques, Jordi Torres, and Xavier Giro-i Nieto. Recurrent neural networks for semantic instance segmentation. *arXiv preprint arXiv:1712.00617*, 2017.

[36] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *CVPR*, 2016.

[37] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *CVPR*, 2021.

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

[39] Tao Wang, Li Yuan, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. Pnp-detr: Towards efficient visual analysis with transformers. In *ICCV*, pages 4661–4670, 2021.

[40] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *AAAI*, 2022.

[41] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *CoRL*, 2022.

[42] Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient detr: improving end-to-end object detector with dense prior. *arXiv preprint arXiv:2104.01318*, 2021.

[43] Mingqiao Ye, Lei Ke, Siyuan Li, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu. Cascade-detr: delving into high-quality universal object detection. In *ICCV*, pages 6704–6714, 2023.

[44] Gongjie Zhang, Zhipeng Luo, Yingchen Yu, Kaiwen Cui, and Shijian Lu. Accelerating detr convergence via semantic-aligned matching. In *CVPR*, pages 949–958, 2022.

[45] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2023.

[46] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. Varifocalnet: An iou-aware dense object detector. In *CVPR*, pages 8514–8523, 2021.

[47] Manyuan Zhang, Guanglu Song, Yu Liu, and Hongsheng Li. Decoupled detr: Spatially disentangling localization and classification for improved end-to-end object detection. In *ICCV*, pages 6601–6610, 2023.

[48] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *CVPR*, 2020.

[49] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye. Freeanchor: Learning to match anchors for visual object detection. In *NeurIPS*, 2019.

[50] Chuyang Zhao, Yifan Sun, Wenhao Wang, Qiang Chen, Errui Ding, Yi Yang, and Jingdong Wang. Ms-detr: Efficient detr training with mixed supervision. In *CVPR*, pages 17027–17036, 2024.

[51] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *CVPR*, 2024.

[52] Dehua Zheng, Wenhui Dong, Hailin Hu, Xinghao Chen, and Yunhe Wang. Less is more: Focus attention for efficient detr. In *ICCV*, pages 6674–6683, 2023.

[53] Benjin Zhu, Jianfeng Wang, Zhengkai Jiang, Fuhang Zong, Songtao Liu, Zeming Li, and Jian Sun. Autoassign: Differentiable label assignment for dense object detection. *arXiv preprint arXiv:2007.03496*, 2020.

[54] 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.

[55] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *ICCV*, pages 6748–6758, 2023.