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

Zhi Cai[1,2], Songtao Liu[2], Guodong Wang[1], Zheng Ge[2], Zeming Li[2], Xiangyu Zhang[2], Di Huang[1]

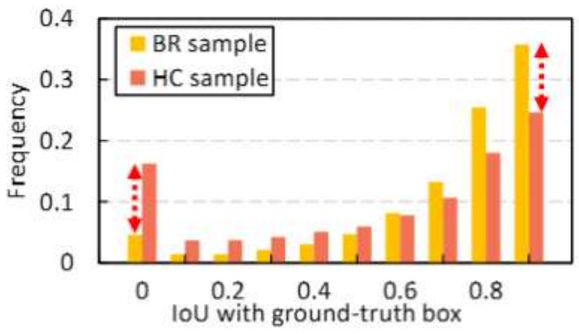
1: IRIP Lab, Beihang University, 2: Megvii Technology

Github: https://github.com/FelixCaae/AlignDETR

➢ Motivation

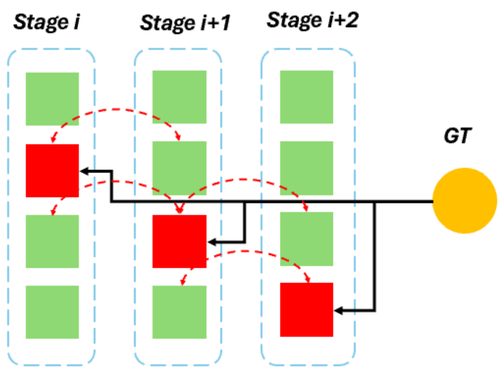
We observe two **mis-alignment problems** in the DETR-like object detectors

- **Classification-regression misalignment**



HC Sample are predictions with **high confidence**
BR samples are **best regressed predictions**

- **Cross-layer target misalignment**



During **label assignment**, the same sample in **different stages** can receive **contradictory targets, which hinders the convergence**

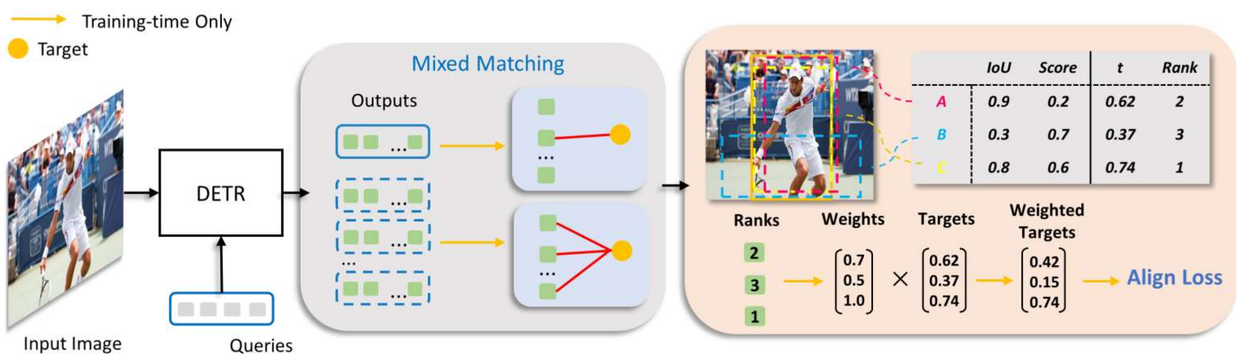
How previous methods handle these problems?

Adopting IoU-aware loss or hybrid matching

Weakness: They do not solve both of the issues jointly!

➢ Our Methods

DETR with mixed matching and aligned loss



Pros:

✓ A unified solution to both of the misalignment problems mentioned above

✓ No extra computation burden (parameters, FLOPS) in training and inference

Mixed Matching:

- One-to-many matching in intermediate layers and one-to-one matching in the output layer
- Similar to hybrid layer matching but we do not change the total query number and thus is more efficient

Align Loss:

- Inspired by QFL, we adopt a quality metric by multiplying confidence and IoU $q = p^{\alpha} \cdot u^{(1-\alpha)}$,
- We sort the predictions according to quality metric and compute a final weighting factor: $t_c = e^{-r/\tau} \cdot q$,

- We define Align loss as:

$$\mathcal{L}_{Align\_cls} = -t_c(1-p)^{\gamma} logp - (1-t_c)\, p^{\gamma} \log(1-p)$$

$$\mathcal{L}_{Align\_reg} = e^{-\frac{r}{\tau}}(\mathcal{L}_{l1}(b,\hat{b}) + \mathcal{L}_{GIoU}(b,\hat{b}))$$

➢ Experiments

- **1x and 2x results on COCO validation set**

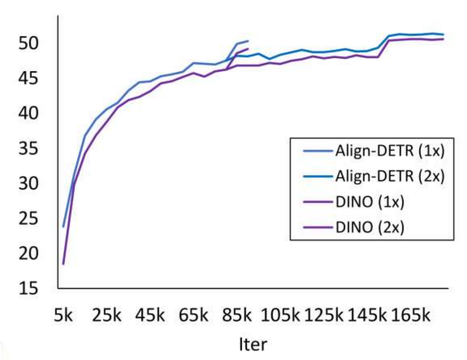
Model	#epochs	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
SMCA-DETR [6]	50	R50	43.7	63.6	47.2	24.2	47.0	60.4
SAM-DETR [44]	50	R50	45.0	65.4	47.9	26.2	49.0	63.3
Def.DETR [54]	50	R50	45.4	64.7	49.0	26.8	48.3	61.7
AdaMixer[8]	36	R50	47.0	66.0	51.1	30.1	50.2	61.8
SD-DETR [47]	50	R50	45.5	65.4	48.5	25.6	49.9	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	56.3	35.2	54.7	66.5
Align-DETR (Ours)	12	R50	**50.5**	**67.7**	**55.3**	**34.7**	53.6	64.6
Align-DETR (Ours)	24	R50	51.7	69.0	56.3	35.5	55.0	66.1

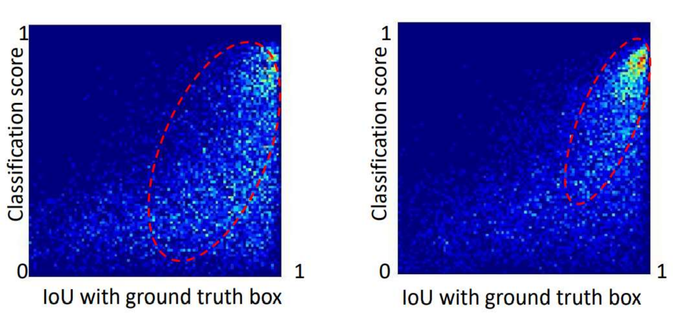
- **Other Experiments**

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**

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

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**





Left: Focal Loss
Right: Align Loss

- **Visualization**