Efficient Decoder-free Object Detection with Transformers

Peixian Chen[1†] Mengdan Zhang[1†] Yunhang Shen[1] Kekai Sheng[1] Yuting Gao[1] Xing Sun[1] Ke Li[12] Chunhua Shen[2]

[1]Tencent Youtu Lab, [2]Zhejiang University

{peixianchen, davinazhang, saulsheng, yutinggao, tristanli}@tencent.com {shenyunhang01, winfred.sun, chhshen}@gmail.com

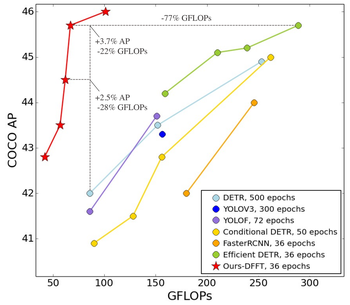


Figure 1: The trade-off between performance (AP) and efficiency (Epochs & GFLOPs) for detection methods.

Highlights

- We eliminate the training-inefficient decoder and leverage two strong encoders to preserve the accuracy of single-level feature map prediction.
- We explore low-level semantic features as much as possible for the detection task with limited computational resources.
- We conduct comprehensive experiments to verify the superiority of DFFT as well as the effectiveness.

DFFT

Framework Overview

- DFFT trims the training-inefficient decoder and simplifies object detection to an encoder-only single-level dense prediction framework.
- DFFT proposed a Detection-oriented transformer (DOT) backbone to extract multi-scale features with strong semantics.
- DFFT proposed two strong encoder (SAE and TAE) to conduct fast but accurate inference on a single-level feature map.

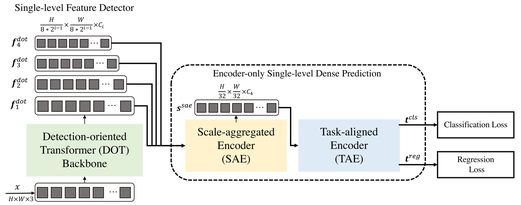


Figure 2: Overview of our proposed DFFT.

Detection-oriented Transformer Backbone aims to extract multi-scale features with strong semantics. As shown in the Fig. 3(a), it hierarchically stacks one embedding module and four DOT stages, where a novel semantic-augmented attention module aggregates the low-level semantic information of every two consecutive DOT stages.

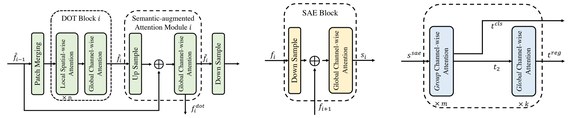
For each input image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times 3}$, the DOT backbone extracts features at four different scales:

$$\boldsymbol{f}_1^{\text{dot}}, \boldsymbol{f}_2^{\text{dot}}, \boldsymbol{f}_3^{\text{dot}}, \boldsymbol{f}_4^{\text{dot}} = F(\boldsymbol{x}), \tag{1}$$

scale-aggregated encoder is proposed to aggregate the multi-scale features $\boldsymbol{f}_i^{\text{dot}}$ from the DOT backbone into one feature map $\boldsymbol{s}^{\text{sae}}$. As shown in the Fig 3(b), each SAE block takes two features as the input and aggregates the features step by step across all SAE blocks.

$$\begin{aligned}
\boldsymbol{s}_0 &= \boldsymbol{f}_1^{\text{dot}}, \\
\boldsymbol{s}_1 &= S_{\text{att}}\big(\text{down}(\boldsymbol{s}_0) + \boldsymbol{f}_2^{\text{dot}}\big), \\
\boldsymbol{s}_2 &= S_{\text{att}}\big(\text{down}(\boldsymbol{s}_1) + \boldsymbol{f}_3^{\text{dot}}\big), \\
\boldsymbol{s}_3 &= S_{\text{att}}\big(\boldsymbol{s}_2 + \text{up}(\boldsymbol{f}_4^{\text{dot}})\big),
\end{aligned} \tag{2}$$

where S_{att} is the global channel-wise attention block and $\boldsymbol{s}^{\text{sae}} = \boldsymbol{s}_3$ is the final aggregated feature map.



(a) The i-th DOT Backbone Stage (b) SAE (c) TAE

Figure 3: Illustration of the three major modules in our proposed DFFT. DFFT contains a light-weight **D**etection-**O**riented **T**ransformer backbone with four DOT stages to extract features with rich semantic information, a **S**cale-**A**ggregated **E**ncoder (SAE) with three SAE blocks to aggregate multi-scale features into one feature map for efficiency, and a **T**ask-**A**ligned **E**ncoder (TAE) to resolve conflicts between classification and regression tasks in the coupled detection head.

Task-aligned Encoder offers a better balance between object classification and localization tasks by learning task-interactive and task-specific features via stacking group channel-wise attention blocks in a coupled head.

$$\begin{aligned}
\boldsymbol{t}_1, \boldsymbol{t}_2 &= T_{\text{group}}(\boldsymbol{s}^{\text{sae}}), \\
\boldsymbol{t}^{\text{cls}} &= \boldsymbol{t}_1, \\
\boldsymbol{t}^{\text{reg}} &= T_{\text{global}}(\boldsymbol{t}_2),
\end{aligned} \tag{3}$$

where $\boldsymbol{t}_1, \boldsymbol{t}_2 \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 256}$ are the split features, and $\boldsymbol{t}^{\text{cls}} \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 256}$ and $\boldsymbol{t}^{\text{reg}} \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 512}$ are the final features for the classification and regression tasks, respectively.

Experiment

Models	Backbone Settings		AP	GFLOPs	FPS
	Value of C_i	Number of SA			
DETR-R50	-	-	42.0	86	24
Deformable DETR	-	-	43.8	173	14
NANO	(3,3,6,9)	(3,3,7,3)	42.8	42	22
TINY	(4,4,8,12)	(2,2,6,2)	43.5	57	24
SMALL	(4,4,8,12)	(3,3,7,3)	44.5	62	22
MEDIUM	(4,4,7,12)	(3,3,19,3)	45.7	67	17
LARGE	(6,6,8,12)	(3,3,19,3)	46.0	101	17

Table 1: The performance, GFLOPs and FPS of DFFT models with different magnitudes on the MS COCO benchmark. C_i means the output feature's number of channels. All the results are measured on the same machine with a V100 GPU using mmdetection.

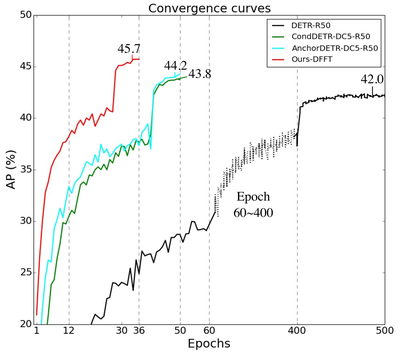


Figure 4: The convergence curves of DFFT and DETR-based methods on the COCO 2017 validation set. Our DFFT converges significantly faster than the counterparts.

DOT	SAE	TAE	AP (%)	GFLOPs
-	-	-	33.8	45
✔	-	-	37.9	47
✔	✔	-	39.9	58
✔	-	✔	39.8	51
✔	✔	✔	41.4	62

Table 2: Ablation study of the three major modules in DFFT.

Source code is available at:
`https://github.com/PeixianChen/DFFT`.