You Only Look One-level Feature (YOLOF)

[Paper]

Overview

1. Utilizing only **one** level feature for detection
2. Dilated encoder
3. Uniform matching

Single-input-single-out

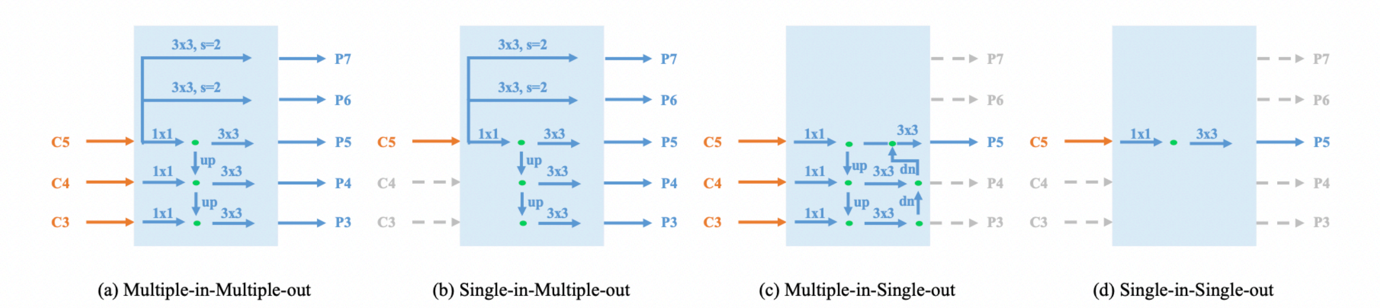


Figure 8. Detailed Structures of Multiple-in-Multiple-out (MiMo), Single-in-Multiple-out (SiMo), Multiple-in-Single-out (MiSo), and Single-in-Single-out (SiSo) encoders.

Dilated encoder

multiple scale range in from only C5 feature

Uniform Matching

adopting the **k nearest** anchors as positive anchors for each gt box

Framework

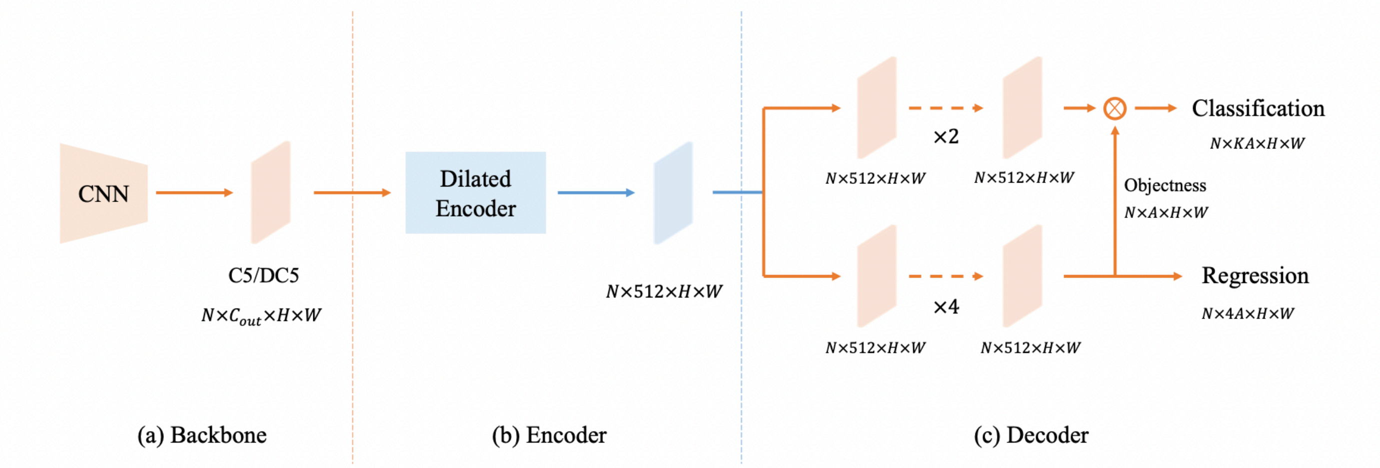


Figure 9. The sketch of YOLOF, which consists of three main components: the backbone, the encoder, and the decoder. In the figure, 'C5/DC5' represents the output feature of the backbone with downsample rate of $32/16$. 'C_{out}' means the number of channels of the feature. We set the number of channels as 512 for feature maps in the encoder and the decoder. $H \times W$ is the height and width of feature maps.

Efficient Decoder-free Object Detection with Transformers (DFFT)

[Paper]

Chunhua Shen

Tecent Youtu lab

Overview

They say:

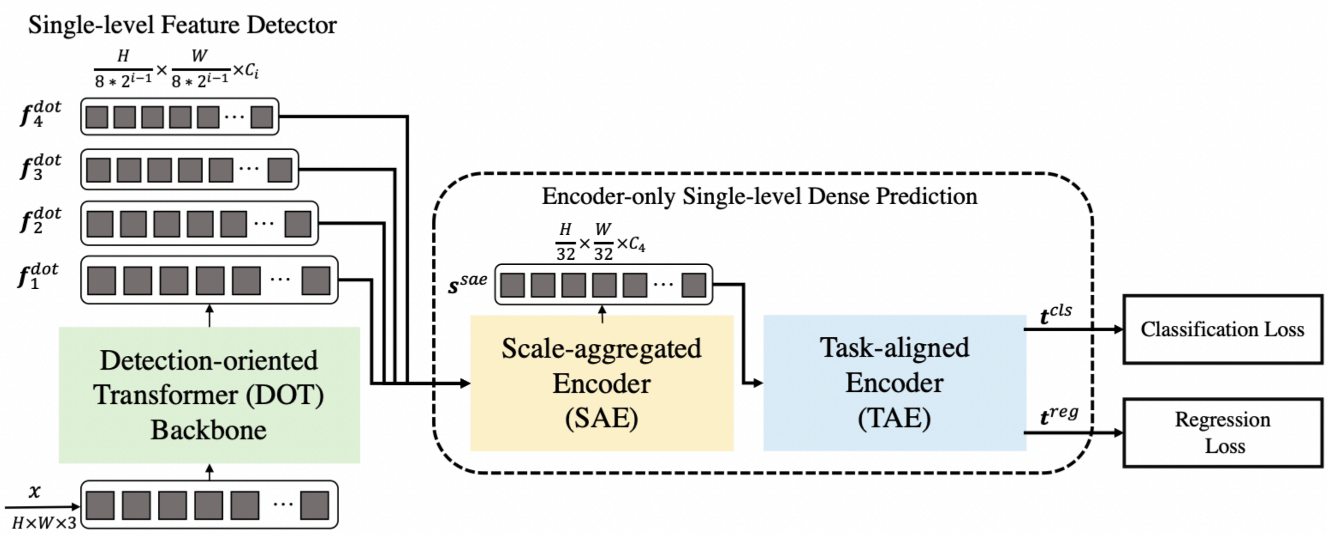
1. Eliminate the training **inefficient decoder**.
2. Leverage **two strong encoders**.
3. Explore **low-level** semantic features with limited computation.

I say:

1. It makes a **trick** in name. It is actually a traditional encoder-neck-decoder structure.
2. It uses **YOLOF SiSo** (single-in single-out) structure.
3. The major contribution is an **efficient vision transformer backbone**.
4. A combination of backbone design and an existing decoder.

Method

Framework



(b) Overview of our proposed DFFT.

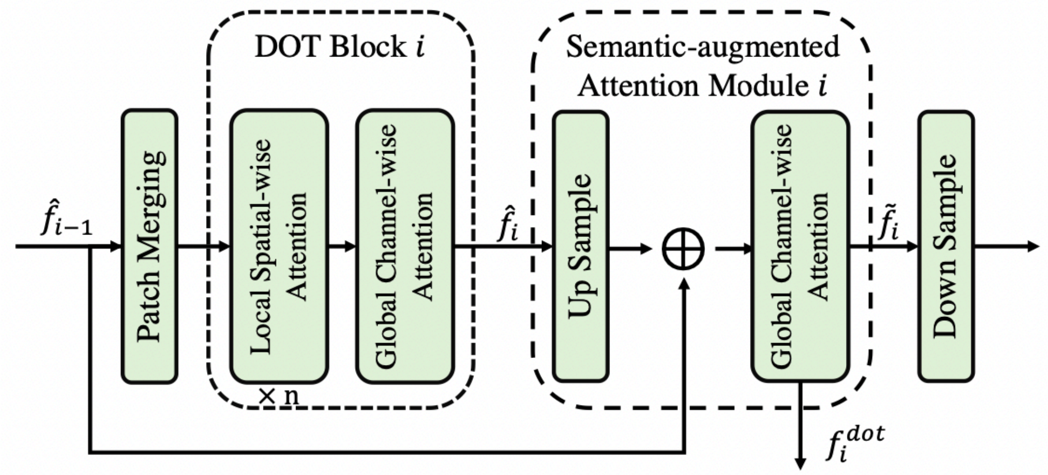
- **Backbone+neck:** Detection-oriented Transformer Backbone (DOT)
- **Encoder:** Scale-aggregated Encoder (SAE)
- **Decoder:** Task-aligned encoder

Backbone: DOT

Patch embedding

- Patch size $\frac{H}{8} \times \frac{W}{8}$.

DOT Stage



(a) The i-th DOT Backbone Stage

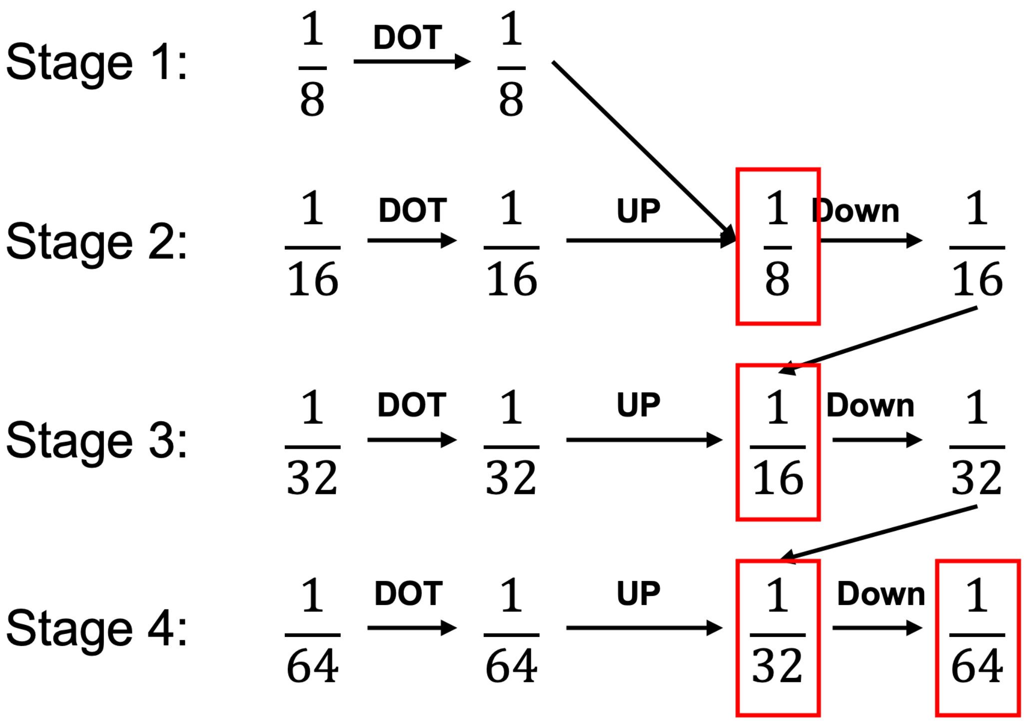
- Each DOT stage: **one DOT block**.
- Each DOT block: **multiple Swin** block (SW-MSA) and **one Xcit** block (global chennel-wise attention block).
- Semantic-augmented Attention (SAA): exchange information between scale levels.

notation

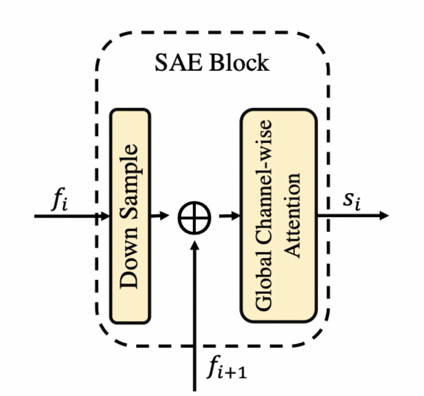
- $\hat{f}_i$ is Denoted as **DOT block** output.
- $\tilde{f}_i$ is denoted as **SAA** output.
- f_i^{dot} is the final multiscale output
- F_{block} is DOT block.
- F_{se-att} is SAA.

instantiations

- **Four** DOT stages
- First stage: one DOT block but **no** SAA
- Other stages: one patch merging, one DOT block, one SAA and one downsampling

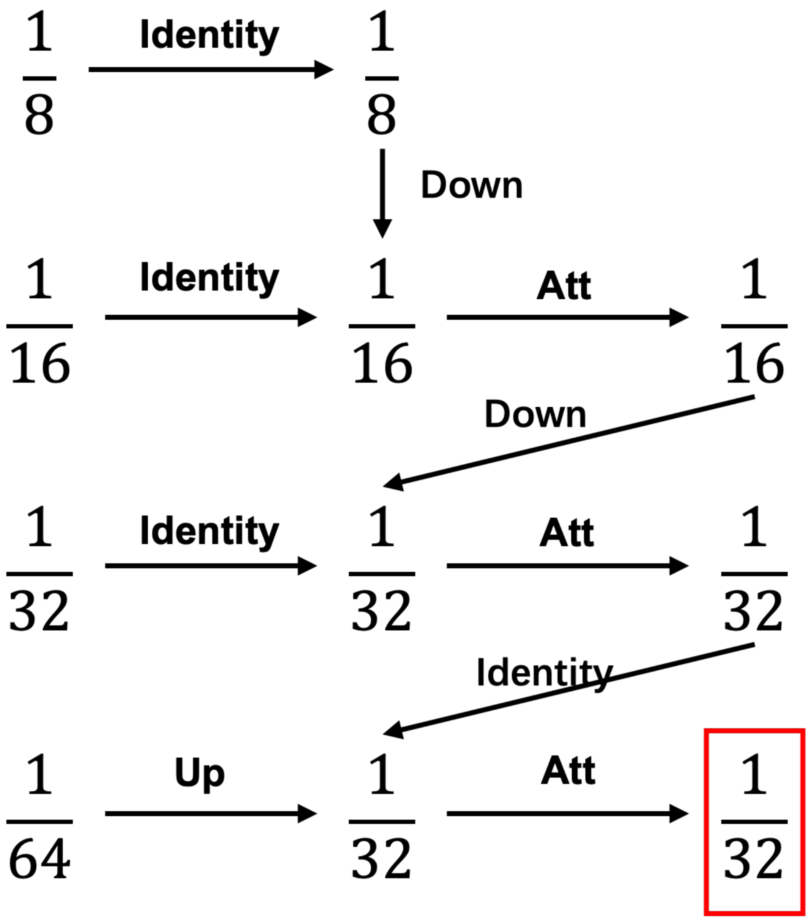


Encoder: Scale-aggregated encoder

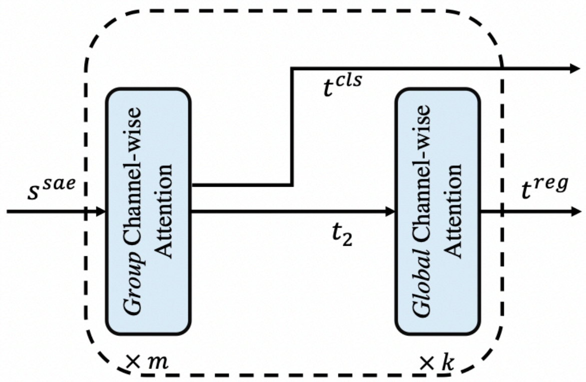


(b) SAE

- Acutally plays as an FPN
- Aggregate at $\frac{1}{32}$



Task-alinged Encoder



(c) TAE

- Except QKV embbeding, all the linear projections of **Group Channelwise Attnetion** are conducted in two groups.
- Predict a **single-level** dense prediction with a **single** feature map. (raised by YOLOF)

Experiments

ImageNet pretrain

Models	Backbone Settings		Effectiveness (%)		Efficiency (GFLOPs)	
	Value of C_i	Number of SA	Accuracy	AP	Backbone	DFFT
$\text{DFFT}_{\text{NANO}}$	$(3, 3, 6, 9)$	$(2, 2, 6, 2)$	80.0	42.8	26	42
$\text{DFFT}_{\text{TINY}}$	$(4, 4, 8, 12)$	$(1, 1, 5, 1)$	81.1	43.5	39	57
$\text{DFFT}_{\text{SMALL}}$	$(4, 4, 8, 12)$	$(2, 2, 6, 2)$	82.1	44.5	44	62
$\text{DFFT}_{\text{MEDIUM}}$	$(4, 4, 7, 12)$	$(2, 2, 18, 2)$	82.7	45.7	48	67
$\text{DFFT}_{\text{LARGE}}$	$(6, 6, 8, 12)$	$(2, 2, 18, 2)$	83.1	46.0	83	101

Table 1. The definition and performance of DFFT models with different magnitudes. In the backbone setting, we list the output feature's number of channels C_i and the number of SA blocks in all four backbone stages. In the effectiveness evaluation, we report the accuracy of the pre-trained backbone on ImageNet and the detection AP of DFFT after training on the MS COCO dataset.

Main results

Methods	Epochs	AP (%)	AP_{50} (%)	AP_{75} (%)	AP_{S} (%)	AP_{M} (%)	AP_{L} (%)	GFLOPs
Faster RCNN-FPN-R50 [26]	36	40.2	61.0	43.8	24.2	43.5	52.0	180
RetinaNet [20]	12	35.9	55.7	38.5	19.4	39.5	48.2	201
YOLOF-R50 [4]	12	37.7	56.9	40.6	19.1	42.5	53.2	86
Swin-Tiny-RetinaNet [24]	12	42.0	-	-	-	-	-	245
Focal-Tiny-RetinaNet [33]	12	43.7	-	-	-	-	-	265
Mobile-Former [5]	12	34.2	53.4	36.0	19.9	36.8	45.3	322
$\text{DFFT}_{\text{NANO}}$	12	39.1	58.3	41.7	19.0	42.9	51.2	42
$\text{DFFT}_{\text{SMALL}}$	12	41.4	60.9	44.5	20.1	45.4	58.9	62
$\text{DFFT}_{\text{MEDIUM}}$	12	42.6	62.5	45.5	22.6	46.7	61.4	67
DETR-R50 [2]	500	42.0	62.4	44.2	20.5	45.8	61.1	86
WB-DETR [22]	500	39.6	58.4	43.8	18.2	42.7	54.9	62
YOLOS [11].	150	37.6	-	-	-	-	-	172
Deformable DETR [36]	50	43.8	62.6	47.7	26.4	47.1	58.0	173
SMCA-R50 [13]	50	43.7	63.6	47.2	24.2	47.0	60.4	152
Anchor DETR-DC5-R50 [31]	50	44.2	64.7	47.5	24.7	48.2	60.6	151
Conditional DETR-R50 [25]	50	40.9	61.8	43.3	20.8	44.6	59.2	90
TSP-FCOS-R50 [28]	36	43.1	62.3	47.0	26.6	46.8	55.9	189
Efficient DETR-R50 [34]	36	44.2	62.2	48.0	**28.4**	47.5	56.6	159
$\text{DFFT}_{\text{NANO}}$	36	42.8	61.9	46.2	23.4	46.8	59.7	42
$\text{DFFT}_{\text{SMALL}}$	36	44.5	63.6	48.0	24.5	49.0	60.7	62
$\text{DFFT}_{\text{MEDIUM}}$	36	**45.7**	**64.8**	**49.7**	25.5	**50.4**	**63.1**	67

Table 2. Comparison of our DFFT and modern detection methods on the MS COCO benchmark [21]. The table is divided into four sections from top to bottom: (1) anchor-based methods, (2) DFFT trained for 12 epochs, (3) DETR-based methods, and (4) DFFT trained for 36 epochs. DFFT achieves competitive precision with significantly fewer training epochs and inference GFLOPs.

Ablation study

Major components

1. Replace DOT backbone with swin
2. Disable SAE by directly upsampling the $\frac{1}{64}$ to $\frac{1}{32}$
3. Replace TAE module with YOLOF head

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

Table 4. Ablation study of the three major modules in DFFT.

SAA

Comparing with FPN on Retina head

SAE

Comparing with dilated encoder in YOLOF

SAA	FPN	AP (%)	GFLOPs
-	-	37.4	319
✔	-	38.9	332
-	✔	38.4	341

Table 5. Analysis of SAA.

Method	AP (%)	GFLOPs
CONCAT	39.6	56
YOLOF	40.3	58
DFFT	41.4	62

Table 6. Analysis of SAE