STUDY AND EVALUATION OF RT-DETR MODEL FOR REAL-TIME OBJECT DETECTION

Huy-Hoan Nguyen

University of Information Technology - VNUHCM

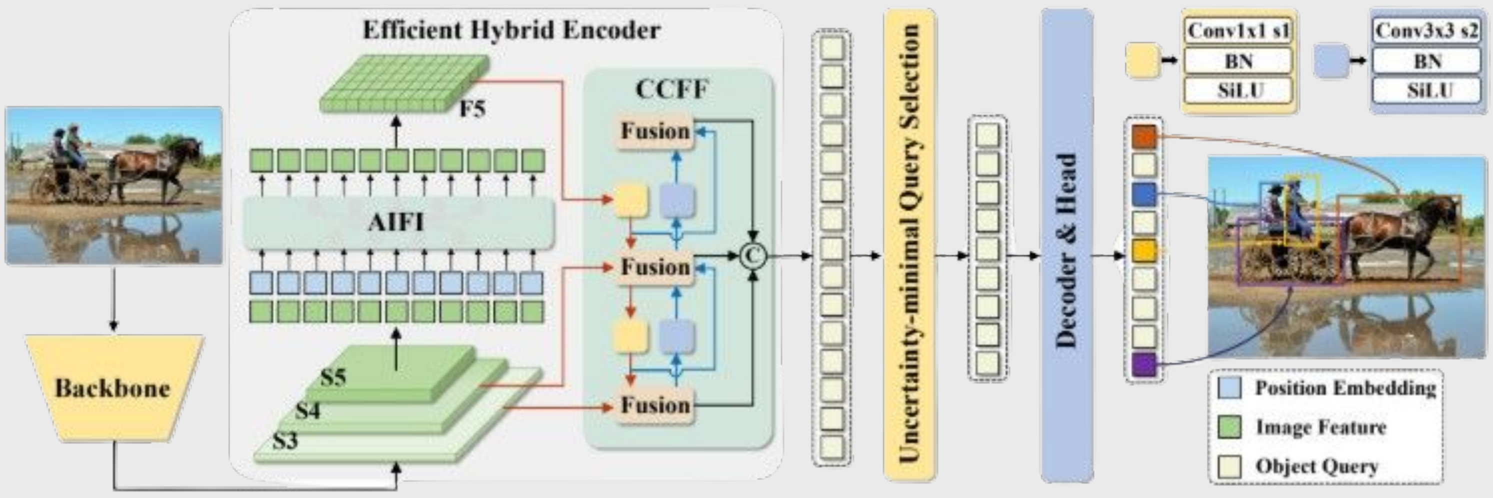
What ?

- We reproduce and evaluate **RT-DETR**, the first real-time End-to-End Object Detector.
- The study benchmarks RT-DETR against state-of-the-art **YOLOv8** models on the **MS COCO val2017** dataset.
- **Goal:** Verify if Transformer architecture can replace CNNs in real-time scenarios by eliminating post-processing bottlenecks.

Why ?

- **Problem:** YOLO models rely on **Non-Maximum Suppression (NMS)**, which causes variable latency and requires manual hyperparameter tuning.
- **Gap:** Previous Transformers (DETR) remove NMS but suffer from high computational costs (low FPS).
- **Solution:** RT-DETR introduces a **Hybrid Encoder** to balance speed and accuracy, achieving real-time performance without NMS.

Overview



The overall architecture of RT-DETR with the Efficient Hybrid Encoder and Transformer Decoder.

Description

1. Efficient Hybrid Encoder

- **AIFI (Attention-based Intra-scale Feature Interaction):** Applies self-attention only on high-level features S5 to capture semantic concepts without heavy computation.
- **CCFF (CNN-based Cross-scale Feature Fusion):** Fuses multi-scale features using efficient convolution blocks instead of heavy attention mechanisms.

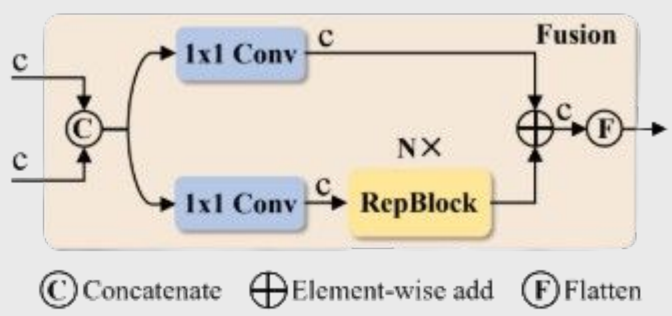


Figure 1: Structure of the Efficient Hybrid Encoder with AIFI and CCFF modules.

2. Uncertainty-minimal Query Selection

- **Problem:** Traditional selection solely based on classification scores often picks features with poor localization accuracy.
- **Solution:** Define an uncertainty metric U that minimizes the discrepancy between the predicted distribution and the ground truth.
- Selects top-K features that act as high-quality initial queries for the Decoder.

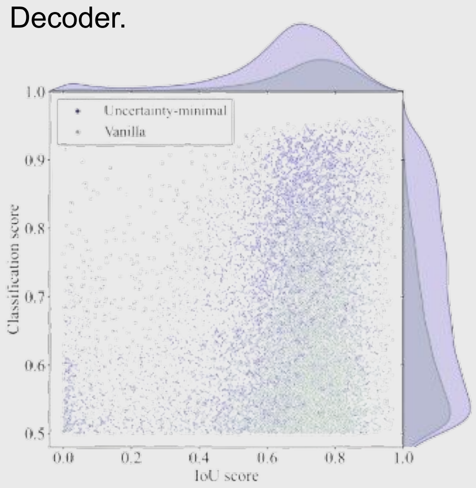


Figure 2: Analysis of query selection. Purple points (Ours) show better alignment between classification and localization.

3. Experimental Results (Benchmark)

- **Problem:** Traditional selection solely based on classification scores often picks features with poor localization accuracy.
- **Solution:** Define an uncertainty metric U that minimizes the discrepancy between the predicted distribution and the ground truth.
- Selects top-K features that act as high-quality initial queries for the Decoder.

Model	Backbone	#Epochs	#Params (M)	GFLOPs	$FPS_{bs=1}$	AP^{val}	AP^{val}_{50}	AP^{val}_{75}	AP^{val}_{S}	AP^{val}_{M}	AP^{val}_{L}
Real-time Object Detectors											
YOLOv5-L [11]	-	300	46	109	54	49.0	67.3	-	-	-	-
YOLOv5-X [11]	-	300	86	205	43	50.7	68.9	-	-	-	-
PPYOLOE-L [40]	-	300	52	110	94	51.4	68.9	55.6	31.4	55.3	66.1
PPYOLOE-X [40]	-	300	98	206	60	52.3	69.9	56.5	33.3	56.3	66.4
YOLOv6-L [16]	-	300	59	150	99	52.8	70.3	57.7	34.4	58.1	70.1
YOLOv7-L [38]	-	300	36	104	55	51.2	69.7	55.5	35.2	55.9	66.7
YOLOv7-X [38]	-	300	71	189	45	52.9	71.1	57.4	36.9	57.7	68.6
YOLOv8-L [12]	-	-	43	165	71	52.9	69.8	57.5	35.3	58.3	69.8
YOLOv8-X [12]	-	-	68	257	50	53.9	71.0	58.7	35.7	59.3	70.7
End-to-end Object Detectors											
DETR-DC5 [4]	R50	500	41	187	-	43.3	63.1	45.9	22.5	47.3	61.1
DETR-DC5 [4]	R101	500	60	253	-	44.9	64.7	47.7	23.7	49.5	62.3
Anchor-DETR-DC5 [39]	R50	50	39	172	-	44.2	64.7	47.5	24.7	48.2	60.6
Anchor-DETR-DC5 [39]	R101	50	-	-	-	45.1	65.7	48.8	25.8	49.4	61.6
Conditional-DETR-DC5 [27]	R50	108	44	195	-	45.1	65.4	48.5	25.3	49.0	62.2
Conditional-DETR-DC5 [27]	R101	108	63	262	-	45.9	66.8	49.5	27.2	50.3	63.3
Efficient-DETR [42]	R50	36	35	210	-	45.1	63.1	49.1	28.3	48.4	59.0
Efficient-DETR [42]	R101	36	54	289	-	45.7	64.1	49.5	28.2	49.1	60.2
SMCA-DETR [9]	R50	108	40	152	-	45.6	65.5	49.1	25.9	49.3	62.6
SMCA-DETR [9]	R101	108	58	218	-	46.3	66.6	50.2	27.2	50.5	63.2
Deformable-DETR [45]	R50	50	40	173	-	46.2	65.2	50.0	28.8	49.2	61.7
DAB-Deformable-DETR [23]	R50	50	48	195	-	46.9	66.0	50.8	30.1	50.4	62.5
DAB-Deformable-DETR++ [23]	R50	50	47	-	-	48.7	67.2	53.0	31.4	51.6	63.9
DN-Deformable-DETR [17]	R50	50	48	195	-	48.6	67.4	52.7	31.0	52.0	63.7
DN-Deformable-DETR++ [17]	R50	50	47	-	-	49.5	67.6	53.8	31.3	52.6	65.4
DINO-Deformable-DETR [44]	R50	36	47	279	5	50.9	69.0	55.3	34.6	54.1	64.6
Real-time End-to-end Object Detector (ours)											
RT-DETR	R50	72	42	136	**108**	**53.1**	**71.3**	**57.7**	34.8	58.0	70.0
RT-DETR	R101	72	76	259	**74**	**54.3**	**72.7**	58.6	36.0	58.8	**72.1**

Table 1: Comparison with SOTA methods. RT-DETR outperforms YOLOv8 in both speed and accuracy.

NII

Nguyễn Huy Hoàn - ĐH Công Nghệ Thông Tin - ĐHQG Thành Phố Hồ Chí Minh

Email : hoannh.20@grad.uit.edu.vn