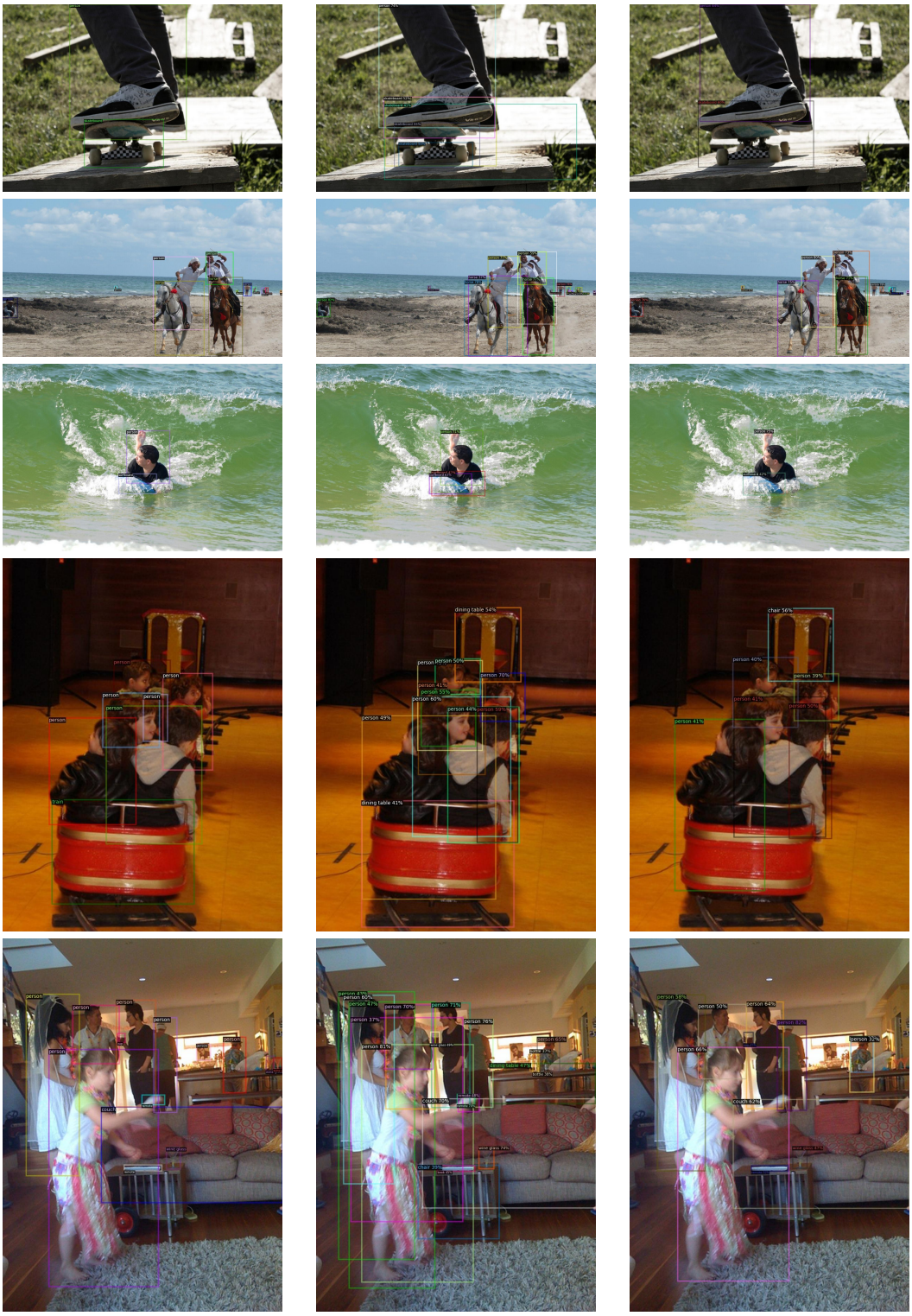
End-to-End Object Detection with Fully Convolutional Network Supplementary Material

1. Visualization

(a) Ground-truth (b) FCOS baseline (c) Ours

Figure 1: The prediction visualizations of different detectors on CrowdHuman *val* set. Our method demonstrates superiority in the crowded scenes. All the models are based on the ResNet-50 backbone. The threshold of the classification score for visualization is set to 0.3.



(a) Ground-truth (b) FCOS baseline (c) Ours

Figure 2: The prediction visualizations of different detectors on COCO *val* set. Compared with the FCOS framework, our end-to-end detector obtains much fewer duplicate predictions, which is crucial for downstream instance-aware tasks. All the models are based on the ResNet-50 backbone. The threshold of the classification score for visualization is set to 0.3.

2. Auxiliary Loss

In this section, we evaluate different one-to-many label assignment rules for the auxiliary loss. The detailed implementations are elaborated as follows:

FCOS. We adopt the assignment rule in FCOS [46].

ATSS. We adopt the assignment rule in ATSS [50].

Quality-ATSS. The rule is elaborated in Sec. 3.2.3.

Quality-FCOS. Similar to FCOS, each ground-truth instance is assigned to the pixels in the pre-defined central area of a specific FPN stage. But the specific FPN stage is selected according to the proposed quality instead of the size of instances.

Quality-Top-k. Each ground-truth instance is assigned to pixels with top-k highest qualities over all the FPN stages. We set $k = 9$ to align with other rules.

As shown in Tab. 1, the results demonstrate the superiority of our proposed prediction-aware quality function over the hand-designed matching metrics. Compared with the standard ATSS framework, the quality based rule can obtain 1.3% mAP absolute gains.

Table 1: The results of different one-to-many label assignment rules for the auxiliary loss on COCO *val* set. All the models are based on the ResNet-50 backbone. '/' is used to distinguish between results without and with NMS.

Method	mAP	AP_{50}	AP_{75}
None	39.8 / 40.0	57.4 / 59.1	43.6 / 43.1
Hand-designed			
FCOS [46]	39.4 / 39.8	57.0 / 59.1	43.4 / 43.0
ATSS [50]	39.8 / 40.1	57.5 / 59.5	44.1 / 43.4
Prediction-aware			
Quality-FCOS	39.7 / 40.0	57.7 / 59.6	43.6 / 43.0
Quality-ATSS	**41.1 / 41.2**	**59.0 / 60.7**	**45.4 / 44.8**
Quality-Top-k	40.7 / 41.0	58.7 / 60.4	44.9 / 44.3

3. Comparison to DETR

As shown in Tab. 2 and Tab. 3, we give the comparison of different methods based on ResNet-50 backbone, where the NMS is not utilized except for FCOS.

Table 2: The comparison on COCO *val* set.

Method	Epochs	mAP	AP_s	AP_m	AP_l	#Param
DETR [3]	500	42.0	20.5	45.8	**61.1**	41.5 M
FCOS [46]	36	41.1	25.9	44.8	52.3	36.4 M
Ours	36	41.5	**26.4**	44.7	52.8	37.0 M
Ours*	36	**43.5**	26.3	**46.6**	55.4	40.3 M

* adopts two extra deformable convolutions in the head.

Table 3: The comparison on CrowdHuman *val* set.

Method	Queries	Epochs	AP_{50}	mMR	Recall
DETR [3]	100	300	72.8	80.1	82.7
DETR	200	300	78.8	66.3	90.2
DETR	300	300	70.6	79.1	89.7
Ours	-	32	**89.1**	**48.9**	**96.5**

Compared with transformers, convolutions have been extensively tested in vision applications and have many variants for better performance than the DETR, *e.g.*, deformable convolutions [59] in Tab. 2. Moreover, as shown in Tab. 3, our framework has great advantages over the DETR [3] in convergence speed and crowded scenes.