DEIM: DETR with Improved Matching for Fast Convergence

Supplementary Material

1. Experimental Settings

Dataset and metric. We evaluate our method on the COCO [20] dataset, training DEIM on `train2017` and validating it on `val2017`. Standard COCO metrics are reported, including AP (averaged over IoU thresholds from 0.50 to 0.95 with a step size of 0.05), AP$_{50}$, AP$_{75}$, and AP at different object scales: AP$_S$, AP$_M$, and AP$_L$.

Table 9. **Different hyperparameters for D-FINE models trained with DEIM**.

D-FINE	**X**	**L**	**M**	**S**
Base LR	5e-4	5e-4	4e-4	4e-4
Min LR	2.5e-4	2.5e-4	2e-4	2e-4
Backbone LR	5e-6	2.5e-5	4e-5	2e-4
Backbone MinLR	2.5e-6	1.25e-5	2e-5	1e-4
Weight of MAL	1	1	1	1
γ in MAL	1.5	1.5	1.5	1.5
Freeze Backbone BN	False	False	False	False
Decoder Act.	SiLU	SiLU	SiLU	SiLU
Epochs	50	50	90	120

Table 10. **Different hyperparameters for RT-DETRv2 models trained with DEIM**.

RT-DETRv2	**X**	**L**	**M***	**M**	**S**
Base LR	2e-4	2e-4	2e-4	2e-4	2e-4
Min LR	1e-4	1e-4	1e-4	1e-4	1e-4
Backbone LR	2e-6	2e-5	2e-5	1e-4	2e-4
Backbone MinLR	1e-6	1e-5	1e-5	5e-5	1e-4
Weight of MAL	1	1	1	1	1
γ in MAL	1.5	1.5	1.5	1.5	1.5
Freeze Backbone BN	False	False	False	False	False
Decoder Act.	SiLU	SiLU	SiLU	SiLU	SiLU
Epochs	60	60	60	120	120

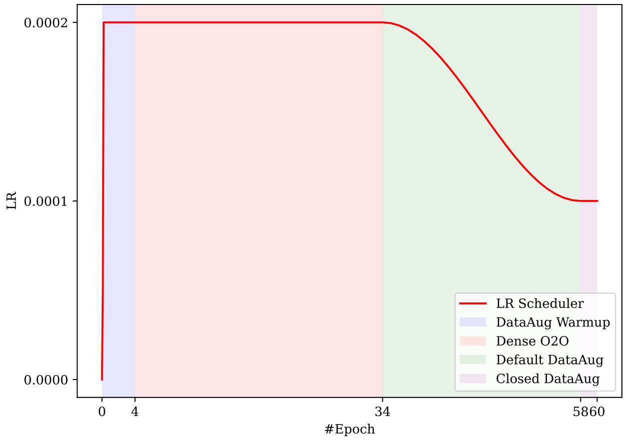

Figure 5. An illustrated example of our proposed novel training scheme for learning rate and data augmentation scheduler.

Implementation details. We implement and validate our method using the D-FINE [27] and RT-DETRv2 [24, 43] frameworks. Most hyperparameters follow their original settings, with differences detailed in Tab. 9 and Tab. 10, respectively. Inspired by the FlatCosine LR scheduler in RT-MDet [25], we propose a novel data augmentation scheduler tailored for Dense O2O. Attention mechanisms in DETRs are critical for extracting accurate object features for localization and classification. However, learning attention from scratch without inductive biases can be challenging. To mitigate this, we introduce a data augmentation warmup strategy, referred to as DataAug Warmup, which simplifies the learning by disabling advanced data augmentations during the initial epochs. An example of the FlatCosine LR and proposed DataAug schedulers for 60 training epochs is shown in Fig. 5.

2. Comparison with Lighter YOLO Detectors

We present the results of comparisons with more lightweight real-time models (S and M sizes) in the Table 11. Based on the strong real-time detectors RT-DETRv2 [24] and D-FINE [27], our DEIM achieves significant improvements across the board. Notably, in RT-DETRv2, all three model sizes show an approximately 1 AP improvement, with the DEIM-RT-DETRv2-M* achieving a remarkable 1.3 AP gain. Compared to other methods, our approach achieves the latest state-of-the-art results.

3. Additional Results

Effectiveness of the minor modifications. We incorporate minor modifications, including unfreezing the BN layers in the Backbone, adopting the FlatCosine LR scheduler, and replacing the Decoder activation function with SiLU, into both D-FINE-L and D-FINE-X. After training for 36 epochs, we observe that these changes have no impact on D-FINE-L but lead to a 0.1 AP improvement for D-FINE-X (55.4 vs. 55.5). This configuration is used as the new baseline for our experiments.

Number of positive samples between with/without Dense O2O. During one epoch of training, we compared the number of positive samples in the same training images

Table 11. **Comparison with S and M sized real-time object detectors on COCO [20] `val2017`.** ⋆ indicates that the NMS is tuned with a confidence threshold of 0.01.

Model	#Epochs	#Params.	GFLOPs	Latency (ms)	AP^{val}	AP$^{val}_{50}$	AP$^{val}_{75}$	AP$^{val}_{S}$	AP$^{val}_{M}$	AP$^{val}_{L}$
YOLO-based Real-time Object Detectors										
YOLOv8-S [12]	500	11	29	6.96	44.9	61.8	48.6	25.7	49.9	61.0
YOLOv8-M [12]	500	26	79	9.66	50.2	67.2	54.6	32.0	55.7	66.4
YOLOv9-S [34]	500	7	26	8.02	46.8	61.8	48.6	25.7	49.9	61.0
YOLOv9-M [34]	500	20	76	10.15	51.4	67.2	54.6	32.0	55.7	66.4
Gold-YOLO-S [33]	300	22	46	2.01	46.4	63.4	-	25.3	51.3	63.6
Gold-YOLO-M [33]	300	41	88	3.21	51.1	68.5	-	32.3	56.1	68.6
YOLOv10-S [32]	500	7	22	2.65	46.3	63.0	50.4	26.8	51.0	63.8
YOLOv10-M [32]	500	15	59	4.97	51.1	68.1	55.8	33.8	56.5	67.0
YOLO11-S⋆ [13]	500	9	22	2.86	47.0	63.9	50.7	29.0	51.7	64.4
YOLO11-M⋆ [13]	500	20	68	4.95	51.5	68.5	55.7	33.4	57.1	67.9
DETR-based Real-time Object Detectors										
RT-DETR-R18 [43]	72	20	61	4.63	46.5	63.8	50.4	28.4	49.8	63.0
RT-DETR-R34 [43]	72	31	93	6.43	48.9	66.8	52.9	30.6	52.4	66.3
RT-DETRv2-S [24]	120	20	60	4.59	48.1	65.1	57.4	36.1	57.9	70.8
DEIM-RT-DETRv2-S	120	20	60	4.59	49.0	66.1	53.3	32.6	52.5	64.1
RT-DETRv2-M [24]	120	31	92	6.40	49.9	67.5	58.6	35.8	58.6	72.1
DEIM-RT-DETRv2-M	120	31	92	6.40	50.9	68.6	55.2	34.3	54.4	67.1
RT-DETRv2-M* [24]	72	33	100	6.90	51.9	69.9	56.5	33.5	56.8	69.2
DEIM-RT-DETRv2-M*	60	33	100	6.90	53.2	71.2	57.8	35.3	57.6	70.2
D-FINE-Nano [27]	148	4	7	2.12	42.8	60.3	45.5	22.9	46.8	62.1
DEIM-D-FINE-Nano	148	4	7	2.12	43.0	60.4	46.2	24.5	47.1	62.1
D-FINE-S [27]	120	10	25	3.49	48.5	65.6	52.6	29.1	52.2	65.4
DEIM-D-FINE-S	120	10	25	3.49	49.0	65.9	53.1	30.4	52.6	65.7
D-FINE-M [27]	120	19	57	5.55	52.3	69.8	56.4	33.2	56.5	70.2
DEIM-D-FINE-M	90	19	57	5.55	52.7	70.0	57.3	35.3	56.7	69.5

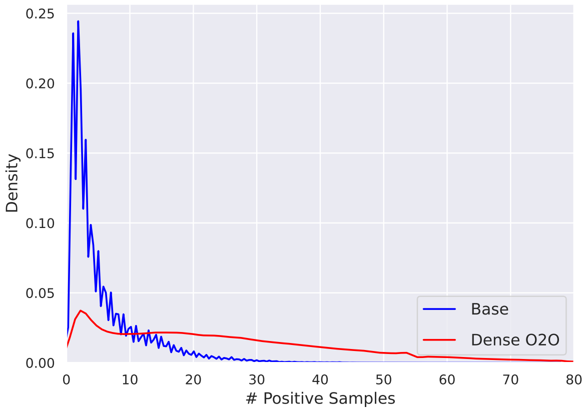

Figure 6. # **Positive Samples with and without Dense O2O in One Epoch of Training.** *Base* indicates without Dense O2O.

with and without using Dense O2O, as shown in Fig. 6. After incorporating Dense O2O, the number of positive samples significantly increases. This further supports our claim that Dense O2O effectively enhances supervision.

Studies of the number of positive samples. We adjust the average number of objects per image during training by

Table 12. **Varying the number of objects per training image.**

Avg # objects	AP	AP_{50}	AP_{75}
Training 24 Epochs			
$\sim$ 10	51.7	69.5	55.8
$\sim$ 25	52.5	70.6	56.7
$\sim$ 50	52.2	70.1	56.4

Table 13. **Training and validation accuracy.**

Model	AP$_{train}$	AP$_{val}$
RT-DETRv2-R50	**65.1**	53.4
w/ DEIM	64.8	**54.3**

modifying Dense O2O. As shown in Tab. 12, performance improves significantly when the number increases from 10 (without Dense O2O) to 25 (Default Dense O2O) but drops at 50 (Max Dense O2O). This decline is likely due to an imbalance in the positive-to-negative ratio and a data distribution shift caused by too many objects. Notably, an average of 25 objects aligns with the default experimental setting

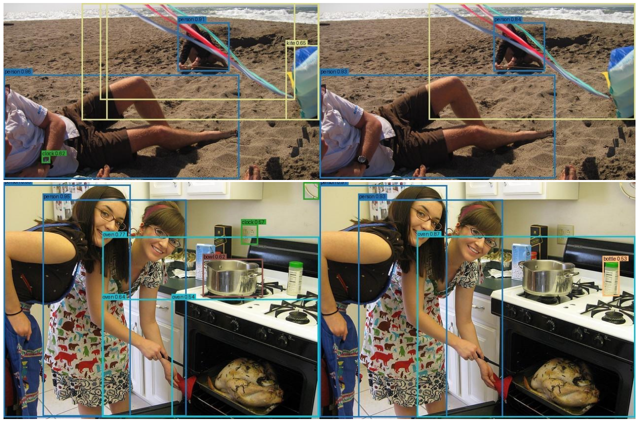

Figure 7. **Qualitative Comparison between D-FINE-L and DEIM.** In each paired image, the left is from D-FINE-L while the right is predicted by DEIM-D-FINE-L (Score threshold = 0.5).

used in this study, corresponding to the default Dense O2O configuration.

Training vs. validation accuracy. As shown in Tab. 13, DEIM achieves higher validation accuracy and slightly lower training accuracy, indicating reduced overfitting on the training set and improved adaptability to new samples.

4. Visualizations

We present the qualitative comparison results in Fig. 7. These results demonstrate that DEIM effectively addresses two critical issues faced by D-FINE-L: high-confidence duplicated predictions and false positives. For example, in the top row, a single kite is erroneously assigned four highly overlapping bounding boxes, each with high confidence scores. Furthermore, as shown in the bottom row, D-FINE-L misclassifies a socket and a wall-mounted object as a clock while failing to detect the bottle. By incorporating DEIM during training, the detector successfully resolves these challenges. This visualization highlights the significant advancements enabled by DEIM, underscoring its potential for improving detection accuracy.