Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation

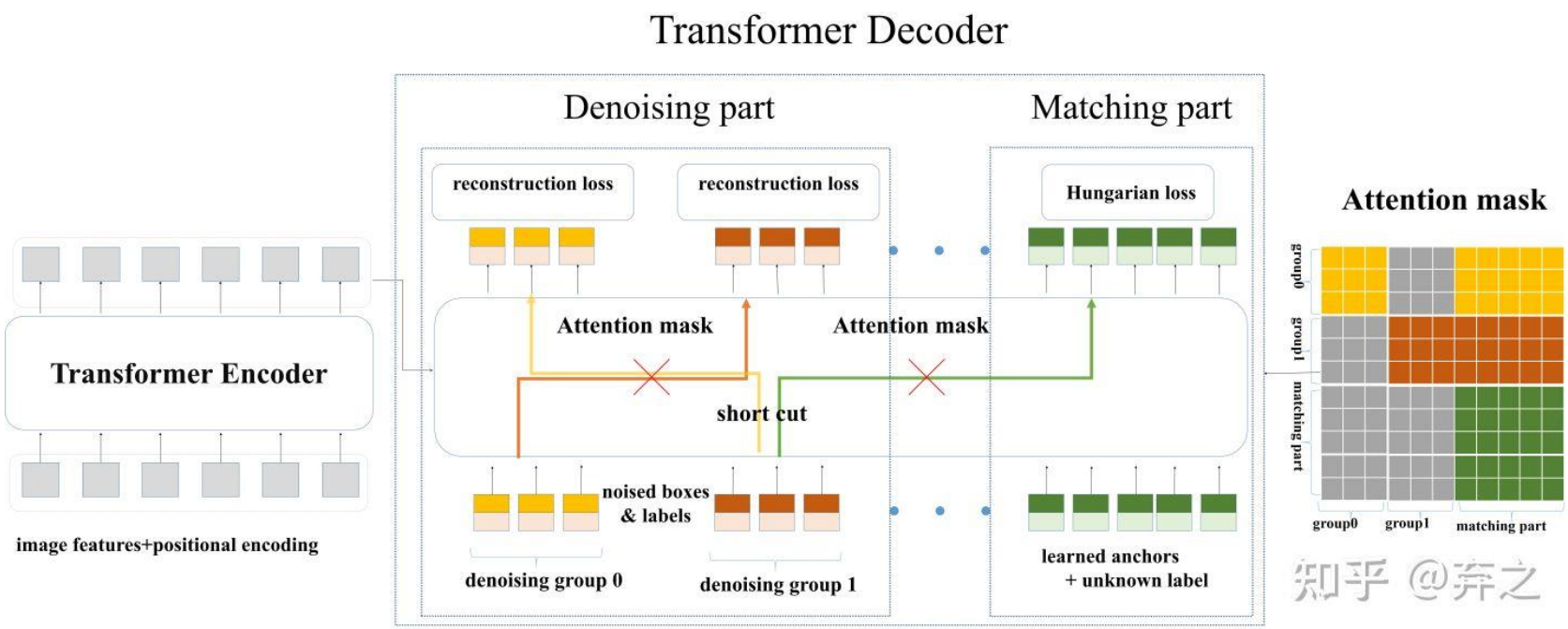
Feng Li[1,3*†], **Hao Zhang**[1,3*†], **Huaizhe Xu**[1,3], **Shilong Liu**[2,3],
Lei Zhang[3‡], **Lionel M. Ni**[1,4], **Heung-Yeung Shum**[1,3]
[1]The Hong Kong University of Science and Technology.
[2]Dept. of CST., BNRist Center, Institute for AI, Tsinghua University.
[3]International Digital Economy Academy (IDEA).
[4]The Hong Kong University of Science and Technology (Guangzhou).

DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection

Hao Zhang[1,3*†], Feng Li[1,3*†], Shilong Liu[2,3*†], Lei Zhang[3‡],
Hang Su[2], Jun Zhu[2], Lionel M. Ni[1,4], Heung-Yeung Shum[1,3]

[1]The Hong Kong University of Science and Technology.
[2]Dept. of CST., BNRist Center, Institute for AI, Tsinghua University.
[3]International Digital Economy Academy (IDEA).
[4]The Hong Kong University of Science and Technology (Guangzhou).

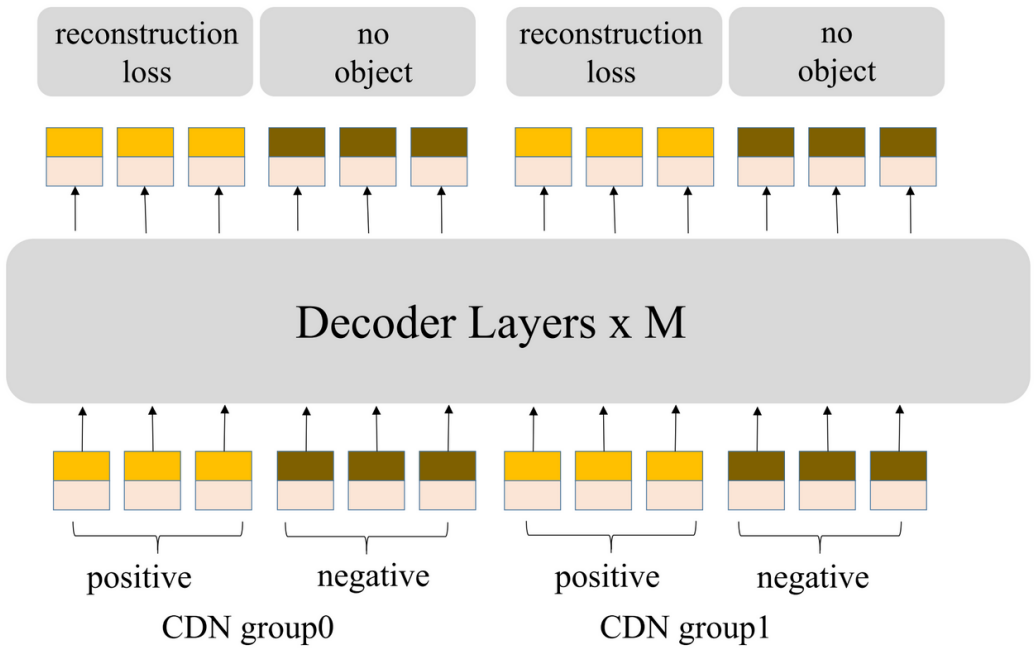
DN-DETR

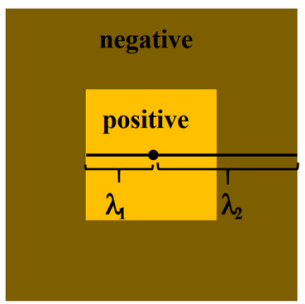


知乎 @弃之

DINO

- Lacks a capability of predicting “no object” for anchors with no object nearby



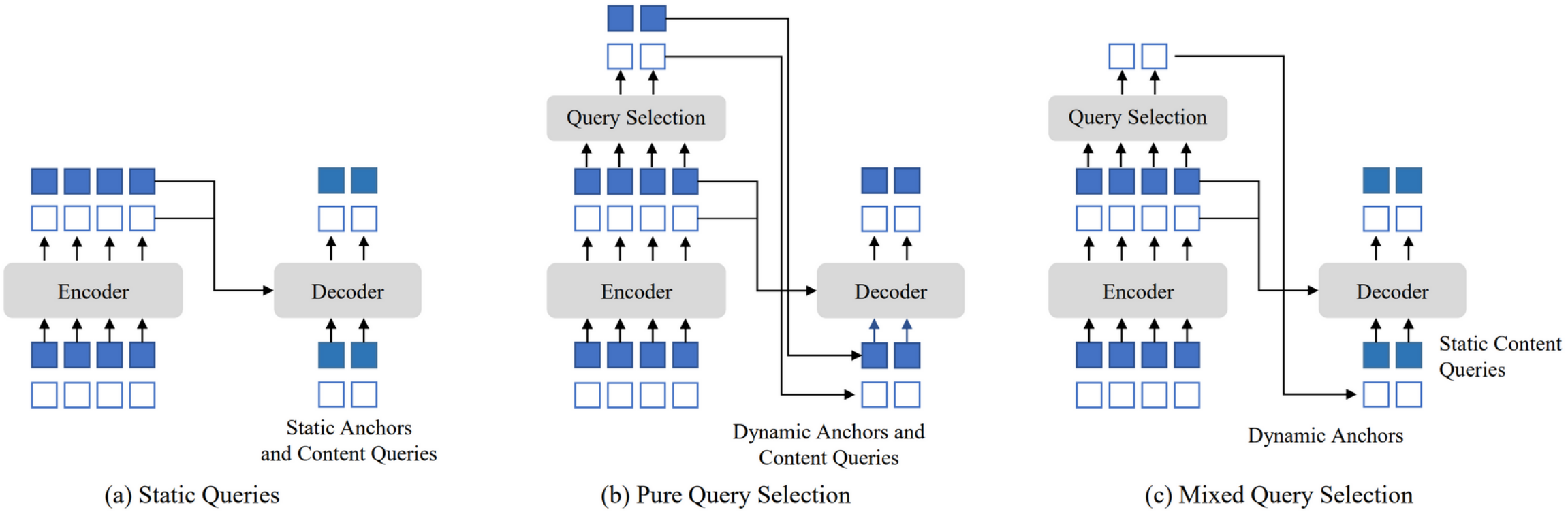


Mixed Query Selection

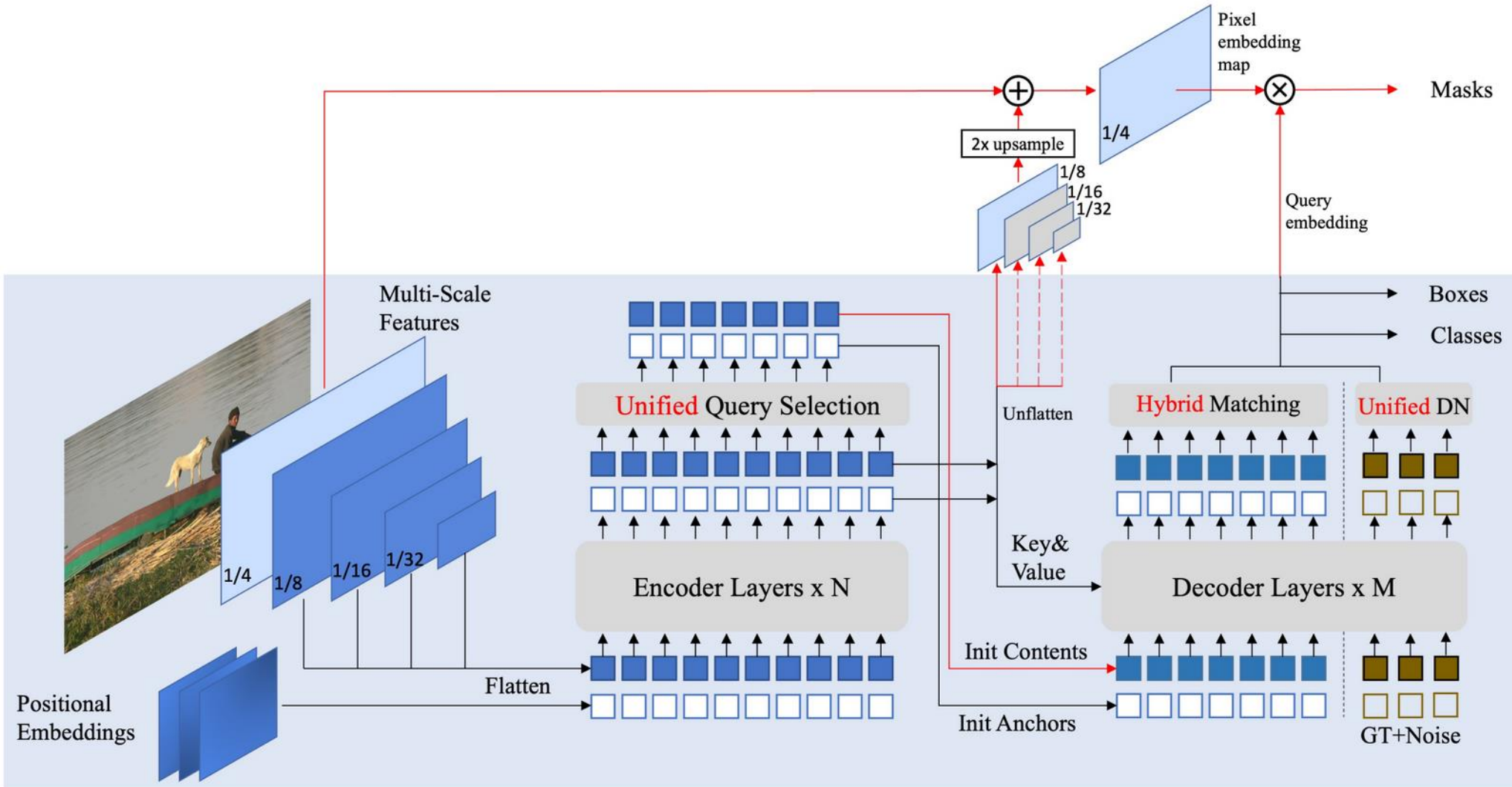
◆ Deformable-DETR select top K encoder features from the last encoder layer as priors to enhance decoder queries.

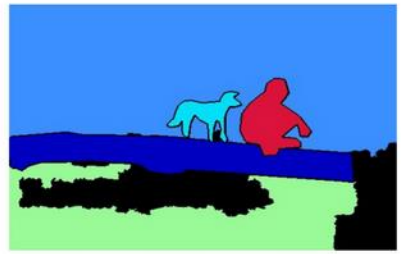
In it, each pixel is assigned as an object query, which directly predicts a bounding box.

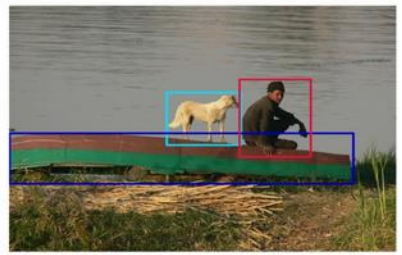
Top scoring bounding boxes are picked as region proposals.



Mask-DINO：a unified object detection and segmentation framework

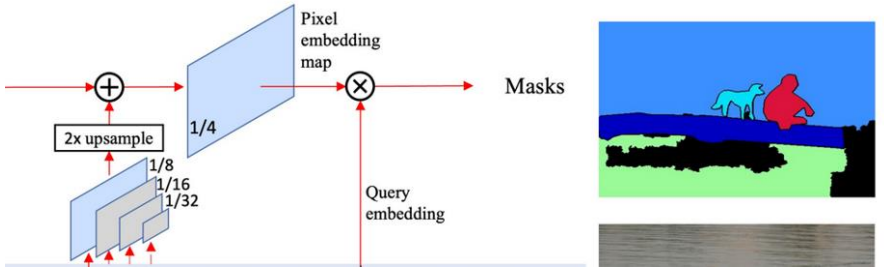






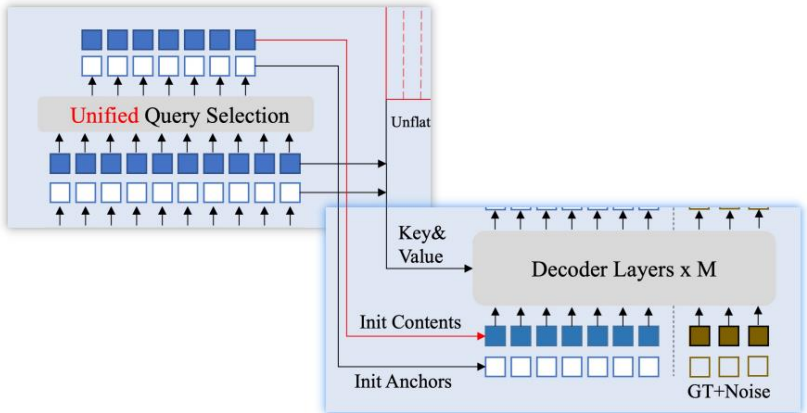
Segmentation branch

$$m = q_c \otimes \mathcal{M}(\mathcal{T}(C_b) + \mathcal{F}(C_e)),$$



Unified query selection for mask

- Predict **both boxes and masks** in the encoder and select the top-ranked ones to initialize decoder queries



Unified denoising for mask

- Noised GT **boxes** and their **labels**

Hybrid matching

- the two heads can predict a pair of box and mask that are **inconsistent** with each other

- both box and mask in bipartite matching to encourage more accurate matching results

Decoupled box prediction

- For the panoptic segmentation task, box prediction for "stuff" categories is **unnecessary** and intuitively **inefficient**

- Remove box loss and box matching for “stuff” categories. The box loss for “stuff” is set to the mean of “thing” categories.

test layer#	Mask DINO	Mask2Former
layer 0	**39.6**(+38.5)	1.1
layer 3	**44.0**	42.3
layer 6	**45.9**	43.3
layer 9	**46.0**	43.7

Table 7: Effectiveness of our query selection for mask initialization. We evaluate the instance segmentation performance from different decoder layers in the same model after training for 50 epochs.

Feature scale	box AP	mask AP
single scale(1/8)	45.8	45.1
3 scales	50.5	45.8
4 scales	50.5	**46.0**

Table 8: Comparison of multi-scale features for Transformer decoder under the 50-epoch setting. Both detection and segmentation benefit from more feature scales.

Decoder layer#	Box AP	Mask AP
3	43.1	40.7
6	44.3	41.1
9	44.5	**41.4**
12	44.8	41.1

Table 9: Decoder layer number comparison under the 12-epoch setting. Mask DINO benefits from more decoders, while DINO's performance will decrease with 9 decoders.

Tasks		Box AP		Mask AP
Box	Mask	12ep	50ep	
✓		**45.1**	50.1	-
	✓	-		43.3
✓	✓	44.5	**50.5**(+0.4)	**46.0**(+2.7)

Table 10: Task comparison under the 50-epoch setting. We train the same Mask DINO with different tasks and validate that box and mask can achieve mutual cooperation.

	Epochs	PQ	PQ^{thing}	PQ^{stf}	Box AP^{Th}_{pan}	Mask AP^{Th}_{pan}
w/o decouple	12	47.9	54.0	38.8	42.8	39.6
w/ decouple	12	**49.0**(+1.1)	**54.8**	**40.2**	**43.2**	**40.4**
w/o decouple	50	52.7	58.8	43.5	48.7	44.1
w/ decouple	50	**53.0**(+0.3)	**59.1**	**43.9**	**48.8**	**44.3**

Table 11: Effectiveness of decoupled box prediction for panoptic segmentation under the 12-epoch and 50-epoch settings.

Matching		Box AP	Mask AP
Box	Mask		
✓		44.4	40.5
	✓	40.2	38.4
✓	✓	**44.5**	**41.4**

Table 12: Matching method comparison under the 12-epoch setting. We train both tasks together but use different matching methods to verify the effectiveness of hybrid matching.

	Box AP	Mask AP
Mask DINO (ours)	44.5	**41.4**
– DINO Mask branch*	49.5†	35.7 (-5.7)
– Unified query selection for masks	43.6	40.3 (-1.1)
– Unified denoising for masks	44.6	40.7 (-0.7)
– Hybrid matching	44.4	40.5 (-0.9)

Table 13: Comparison of the proposed components under the 12-epoch setting. * indicates that we use the original DETR [2] segmentation branch in Mask DINO, where we follow DETR to fine-tune segmentation after finishing training detection. † the performance of detection drops (49.6 AP as shown in Table 2 when only training detection) after training segmentation.

Method	Params	Backbone	Backbone Pre-training Dataset	Detection Pre-training Dataset	val	
					w/o TTA	w/ TTA
Instance segmentation on COCO					AP	
Mask2Former [4]	216M	SwinL	IN-22K-14M	–	50.1	–
Soft Teacher [36]	284M	SwinL	IN-22K-14M	O365	51.9	52.5
SwinV2-G-HTC++ [23]	3.0B	SwinV2-G	IN-22K-ext-70M [23]	O365	53.4	53.7
MasK DINO(Ours)	**223**M	SwinL	IN-22K-14M	O365	**54.5**(+1.1)	–
Panoptic segmentation on COCO					PQ	
Panoptic SegFormer [19]	–M	SwinL	IN-22K-14M	–	55.8	–
Mask2Former [4]	216M	SwinL	IN-22K-14M	–	57.8	–
MasK DINO (ours)	223M	SwinL	IN-22K-14M	O365	**59.4**(+1.6)	–
Semantic segmentation on ADE20K					mIoU	
Mask2Former [4]	215M	SwinL	IN-22K-14M	–	56.1	57.3
Mask2Former [4]	217M	SwinL-FaPN	IN-22K-14M	–	56.4	57.7
SeMask-L MSFaPN-Mask2Former [14]	–M	SwinL-FaPN	IN-22K-14M	–	–	58.2
SwinV2-G-UperNet [23]	3.0B	SwinV2-G	IN-22K-ext-70M [23]	–	59.3	59.9
MasK DINO (ours)	**223M**	SwinL	IN-22K-14M	O365	**59.5**	**60.8**(+0.9)

Table 6: Comparison of the SOTA models on three segmentation tasks. Mask DINO outperforms all existing models. "TTA" means test-time-augmentation. “O365” denotes the Objects365 [31] dataset.