

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}

¹The Hong Kong University of Science and Technology.

²Dept. of CST., BNRIst Center, Institute for AI, Tsinghua University.

³International Digital Economy Academy (IDEA).

⁴The Hong Kong University of Science and Technology (Guangzhou).

{hzhangcx, fliay, hxubr}@connect.ust.hk

{liusl20}@mails.tsinghua.edu.cn

{leizhang}@idea.edu.cn

{ni, hshum}@ust.hk

Abstract

In this paper we present Mask DINO, a unified object detection and segmentation framework. Mask DINO extends DINO (DETR with Improved Denoising Anchor Boxes) by adding a mask prediction branch which supports all image segmentation tasks (instance, panoptic, and semantic). It makes use of the query embeddings from DINO to dot-product a high-resolution pixel embedding map to predict a set of binary masks. Some key components in DINO are extended for segmentation through a shared architecture and training process. Mask DINO is simple, efficient, scalable, and benefits from joint large-scale detection and segmentation datasets. Our experiments show that Mask DINO significantly outperforms all existing specialized segmentation methods, both on a ResNet-50 backbone and a pre-trained model with SwinL backbone. Notably, Mask DINO establishes the best results to date on instance segmentation (54.5 AP on COCO), panoptic segmentation (59.4 PQ on COCO), and semantic segmentation (60.8 mIoU on ADE20K). Code will be available at <https://github.com/IDEACVR/MaskDINO>.

1 Introduction

Object detection and image segmentation are fundamental tasks in computer vision. Both tasks are concerned with localizing objects of interest in an image but have different levels of focus. Object detection is to localize objects of interest and predict their bounding boxes and category labels, whereas image segmentation focuses on pixel-level grouping of different semantics. Moreover, image segmentation encompasses various tasks including instance segmentation, panoptic segmentation, and semantic segmentation with respect to different semantics, e.g. instance or category membership, foreground or background category.

Remarkable progress has been achieved by classical convolution-based algorithms developed for these tasks with specialized architectures, such as Faster RCNN [28] for object detection, Mask RCNN [12] for instance segmentation, and FCN [25] for semantic segmentation. Although these methods are conceptually simple and effective, they are tailored for specialized tasks and lack the

*Equal contribution.

†This work was done when Feng Li and Hao Zhang were interns at IDEA.

‡Corresponding author.

generalization ability to address other tasks. The ambition to bridge different tasks gives rise to more advanced methods like HTC [3] for object detection and instance segmentation and Panoptic FPN [16], K-net [38] for instance, panoptic, and semantic segmentation. Task unification not only helps simplify the algorithm development but also brings in performance improvement in multiple tasks. As an evidence, up to now, an improved HTC (HTC++) is still a widely-used object detection and instance segmentation method used by state-of-the-art (SOTA) models on the COCO object detection and instance segmentation leaderboards [1].

As we step into the new era of Transformer-based detectors, detection and segmentation tasks diverge into different models. Transformer [32] was first introduced into object detection by DETR [2]. DETR is an end-to-end query-based object detector, which adopts a set-prediction objective with bipartite matching. Although DETR addresses both the object detection and panoptic segmentation tasks, its segmentation performance is still inferior to classical segmentation models. To improve the detection and segmentation performance of query-based models, researchers have developed specialized models for object detection [40, 22, 18, 37], image segmentation [38, 6, 4], instance segmentation [10], panoptic segmentation [27], and semantic segmentation [14].

Among the efforts to improve object detection, DINO (DETR with Improved Denoising Anchor Boxes) [37] takes advantage of the dynamic anchor box formulation from DAB-DETR [22] and query denoising training from DN-DETR [18], and further develops contrastive denoising training, mixed query selection, and look forward twice methods to accelerate training and improve the detection performance. As a result, DINO achieves the SOTA result on the COCO object detection leaderboard for the first time as a DETR-like model. Similarly, for improving image segmentation, MaskFormer [6] and Mask2Former [4] propose to unify different image segmentation tasks using query-based Transformer architectures to perform mask classification. Such methods have achieved remarkable performance improvement on multiple segmentation tasks.

However, detection and segmentation models still diverge significantly, which prevents task and data cooperation between detection and segmentation tasks. For example, the state-of-the-art query-based instance segmentation model Mask2Former still lags behind classical models based on HTC++ with Swin-V2-G [23]. One reason to account for this performance gap is that HTC-based models are pre-trained on a large-scale detection dataset (i.e Objects365 [31]) but Mask2Former can not utilize detection data for pre-training. Though we believe detection and segmentation can help each other in a unified architecture, the results of simply using DINO for segmentation and using Mask2Former for detection indicate that they can not do other tasks well. Moreover, trivial multi-task training can even hurt the performance of the original tasks. It naturally leads to two questions: 1) *why cannot detection and segmentation tasks help each other in query-based models?* and 2) *is it possible to develop a unified architecture for all detection and segmentation tasks to replace specialized ones?*

To address these problems, we propose Mask DINO, which extends DINO by adding a mask prediction branch in parallel with the box prediction branch. Inspired by other unified models [33, 6, 4] for image segmentation, we reuse content query embeddings from DINO to perform mask prediction for all segmentation tasks. As DINO lacks a high-resolution feature map for mask prediction, we follow MaskFormer and Mask2Former and construct a high-resolution pixel embedding map (1/4 of the input image resolution) obtained from the backbone and Transformer encoder features. The mask branch predicts binary masks by simply dot-producing each content query embedding with the pixel embedding map. Although DINO is not designed for pixel-level alignment, we find it quite effective to reuse its content query embeddings for mask prediction. This can largely attribute to the cross attention-based feature pooling in Transformer decoder, allowing query embeddings to only aggregate related features from the input image. Besides the mask branch, we also extend three key components for box prediction in DINO to improve the segmentation performance. First, we propose unified query selection to initialize mask queries as anchors, which selects masks from encoder dense prior. Second, we propose unified denoising training for masks to accelerate segmentation training. Third, we use a hybrid bipartite matching for more precise matching with both boxes and masks.

Mask DINO is conceptually simple and easy to implement under the DINO framework. The extension for mask prediction is designed to be as simple as possible so that Mask DINO can reuse all algorithm improvements in DINO as well as its feature representation which can be pre-trained from a much larger detection dataset. We also find that the hard-constrained and dense masked attention in Mask2Former not necessary. Mask DINO indicates simply concatenating multi-scale features from Transformer encoder and using deformable attention can result in a remarkable segmentation

performance. In addition, Mask DINO is computationally efficient as we use sparse deformable attention. As an evidence, our FPS is much higher than Mask2Former [4] (14.8 vs 8.2 in Table 2).

To summarize, our contributions are three-fold.

1. We develop a unified Transformer-based framework for both object detection and segmentation. As the framework is extended from DINO, a DETR-like model, by adding a mask prediction branch, it naturally inherits most algorithm improvements in DINO including anchor box-guided cross attention, query selection, denoising training, and even a better representation pre-trained on a large scale detection dataset.
2. We demonstrate that detection and segmentation can help each other through a shared architecture design and training method. Especially, detection can significantly help segmentation tasks, even for segmenting background "stuff" categories. Under the same setting with a ResNet-50 backbone, Mask DINO outperforms all existing models compared to DINO (+0.4 AP on COCO detection) and Mask2Former (**+2.3 AP**, **+1.1 PQ**, and **+1.5 mIoU** on COCO instance, COCO panoptic, and ADE20K semantic segmentation).
3. We also show that, via a unified framework, segmentation can benefit from detection pre-training on a large-scale detection dataset. After pre-training detection on the Objects365 [31] dataset with a SwinL [24] backbone, Mask DINO significantly improves all segmentation tasks and sets new records on instance (**54.5 AP** on COCO), panoptic (**59.4 PQ** on COCO), and semantic (**60.8 mIoU** on ADE20K) segmentation.

2 Related Work

Detection: Mainstream detection algorithms have been dominated by convolutional neural network-based frameworks, until recently Transformer-based detectors [2, 22, 18, 37] achieve great progress. DETR [2] is the first end-to-end and query-based Transformer object detector, which adopts a set-prediction objective with bipartite matching. DAB-DETR [22] improves DETR by formulating queries as 4D anchor boxes and refining predictions layer by layer. DN-DETR [18] introduces a denoising training method to accelerate convergence. It takes noised ground-truth objects as input and trains the model to reconstruct the ground-truth objects on the output side. Based on DAB-DETR and DN-DETR, DINO [37] proposes several new improvements on denoising and anchor refinement and achieves new SOTA results on COCO detection. Despite the inspiring progress, DETR-like detection models are not competitive for segmentation. Vanilla DETR incorporates a segmentation head in its architecture. However, its segmentation performance is inferior to specialized segmentation models and only shows the feasibility of DETR-like detection models to deal with detection and segmentation simultaneously.

Segmentation: Segmentation mainly includes three tasks: instance, semantic, and panoptic segmentation. The three tasks are similar but focus on different semantics. Instance segmentation is to predict a mask and its corresponding category for each object instance. Semantic segmentation requires to classify each pixel including the background into different semantic categories. Panoptic segmentation [16] unifies the instance and semantic segmentation tasks and predicts a mask for each object instance or background segment. In the past few years, researchers have developed specialized architectures for the three tasks. For example, Mask-RCNN [12] and HTC [3] can only deal with instance segmentation because they predict the mask of each instance based on its box prediction. FCN [25] and U-Net [30] can only perform semantic segmentation since they predict one segmentation map based on pixel-wise classification. Although panoptic segmentation unifies the above two tasks, models designed for panoptic segmentation [15, 35] are usually sub-optimal for instance and semantic segmentation compared with specialized models. Until recently, some image segmentation models [38, 6, 4] are developed to unify the three tasks with a universal architecture. For instance, Mask2Former [4] improves MaskFormer [6] by introducing masked-attention to Transformer and achieves SOTA performance on panoptic segmentation. Mask2Former has a similar architecture as DETR to probe image features with learnable queries but differs in using a different segmentation branch and some specialized designs for mask prediction. However, while Mask2Former shows a great success in unifying all segmentation tasks, it leaves object detection untouched and our empirical study shows that its specialized architecture design is not suitable for predicting boxes.

Unified Methods: As both object detection and segmentation are concerned with localizing objects, they naturally share common model architectures and visual representations. A unified framework

not only helps simplify the algorithm develop effort but also allows to use both detection and segmentation data to improve representation learning. There have been several previous works to unify segmentation and detection tasks, e.g., Mask RCNN [12], HTC [3], and DETR [2]. Mask RCNN extends Faster RCNN and pools image features from Region Of Interest (ROI) proposed by RPN. HTC further proposes an interleaved way of predicting boxes and masks to improve the segmentation performance. However, these two models can only perform instance segmentation. DETR predicts boxes and masks together in an end-to-end manner. However, its segmentation performance largely lags behind other models. According to Table 13, when we use DETR’s segmentation head to predict instance segmentation results, the mask AP is 14% lower than the box AP. How to attain mutual assistance between segmentation and detection has long been an important problem to solve.

3 Mask DINO

Mask DINO is an extension of DINO [37]. On top of content query embeddings, DINO has two branches for box prediction and label prediction. The boxes are dynamically updated and used to guide the deformable attention in each Transformer decoder. Mask DINO adds another branch for mask prediction and minimally extends several key components in detection to fit segmentation tasks. To better understand Mask DINO, we start by briefly reviewing DINO and then introduce Mask DINO.

3.1 Preliminaries: DINO

DINO is a typical DETR-like model, which is composed of a backbone, a Transformer encoder, and a Transformer decoder. The framework is shown in Fig. 1 (the blue-shaded part without red lines). Following DAB-DETR [22], DINO formulates each positional query in DETR as a 4D anchor box, which is dynamically updated through each decoder layer. Note that DINO uses multi-scale features with deformable attention [40]. Therefore, the updated anchor boxes are also used to constrain deformable attention in a sparse and soft way. Following DN-DETR [18], DINO adopts denoising training and further develops contrastive denoising to accelerate training convergence. Moreover, DINO proposes a mixed query selection scheme to initialize positional queries in the decoder and a look-forward-twice method to improve box gradient back-propagation.

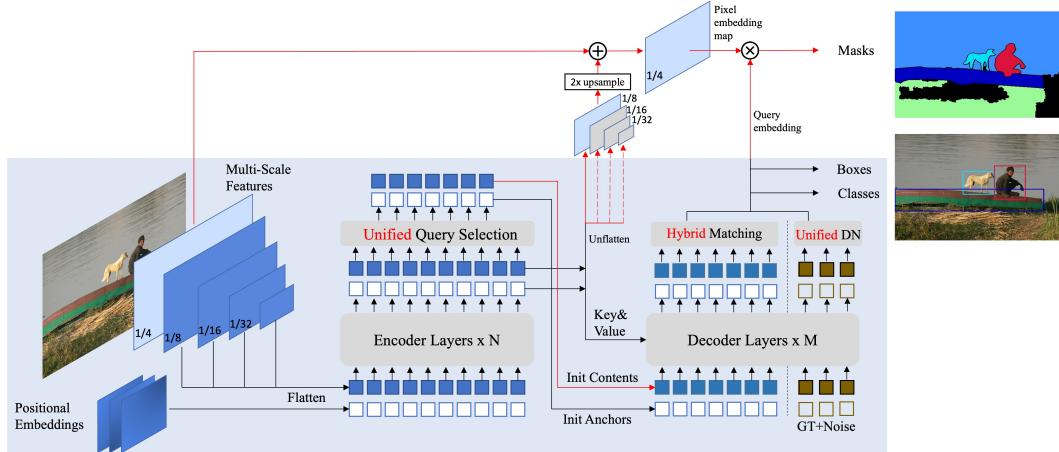


Figure 1: The framework of Mask DINO, which is based on DINO (the blue-shaded part) with minimal extensions (the red part) for segmentation tasks. The key extension is a mask prediction branch which performs dot production between each query embedding with a high resolution (1/4 of the input image resolution) pixel embedding map. In unified query selection, top K features from the last encoder layer are selected to initialize both the positional queries and content queries fed to the Transformer decoder. We also use hybrid matching with masks for more accurate matching and unified DN (denoising) for masks to accelerate segmentation training.

3.2 Why a universal model has not replaced the specialized models?

Remarkable progress has been achieved by Transformer-based detectors and segmentation models. For instance, DINO [37] and Mask2Former [4] have achieved the best result on COCO detection and panoptic segmentation, respectively. Inspired by such progress, we attempted to simply extend these specialized models for other tasks but found that the performance of other tasks lagged behind the original ones by a large margin. It seems that trivial multi-task training even hurts the performance of the original task. However, in convolution-based models, it has shown effective and mutually beneficial to combine detection and instance segmentation tasks. For example, HTC++ [3] is still ranked first on the COCO instance segmentation. We are keen to answer two questions in this work: 1) *why cannot detection and segmentation tasks help each other in Transformer-based models?* and 2) *is it possible to develop a unified architecture for all detection and segmentation tasks to replace specialized ones?* We will take DINO and Mask2Former as examples to discuss the above questions.

–**Why cannot Mask2Former do detection well?** The Transformer decoder of Mask2Former is designed for segmentation tasks and does not suit detection for three reasons. First, its **queries** follow the design in DETR [2] without being able to utilize better positional priors as studied in Conditional DETR [26], Anchor DETR [34], and DAB-DETR [22]. For example, its content queries are semantically aligned with the features from the Transformer encoder, whereas its positional queries are just learnable vectors as in vanilla DETR instead of being associated with a single-mode position⁴. If we remove its mask branch, it reduces to a variant of DETR [2], whose performance is inferior to recently improved DETR models. Second, Mask2Former adopts **masked attention** (multi-head attention with attention mask) in Transformer decoders. The attention masks predicted from a previous layer are dense and hard-constrained, which is neither efficient nor flexible for box prediction. Third, Mask2Former cannot explicitly perform **box refinement** layer by layer. Moreover, its coarse-to-fine mask refinement in decoders fails to use multi-scale features from the encoder.

–**Why cannot DETR/DINO do segmentation well?** DETR [2] has incorporated a segmentation head into its architecture to show the potential of extending to segmentation tasks. However, its performance is limited. There are three reasons. First, its **segmentation head** is not optimal. DETR lets each query embedding dot-product with the smallest feature map to compute attention maps and then upsamples them to get the mask predictions. This design lacks an interaction between queries and larger feature maps from the backbone. Second, DETR cannot use **mask auxiliary loss** in each decoder layer. DETR’s architecture is inefficient for segmentation due to its heavy segmentation head and dense mask loss. Therefore, it only computes segmentation loss in the last decoder layer. Third, It does not support **mask refinement** as the mask positional prediction from one layer cannot pass to the next layer.

–**The motivation of Mask DINO.** There has been a trend to unify detection and segmentation tasks using convolution-based models, which not only simplifies model design but also promotes mutual cooperation between detection and segmentation. There are mainly three motivations for us to propose Mask DINO. **First**, DINO [37] has achieved SOTA results on object detection. Previous works such as Mask RCNN [12], HTC [3], and DETR [2] have shown that a detection model can be extended to do segmentation and help design better segmentation models. **Second**, detection is a relatively easier task than instance segmentation. As shown in Table 2 (and other previous studies), Box AP is usually 4+ AP higher than mask AP. Therefore, box prediction can guide attention to focus on more meaningful regions and extract better features for mask prediction. **Third**, the new improvements in DINO and other DETR-like models [40, 18] such as query selection and deformable attention can also help segmentation tasks. For example, Mask2Former adopts learnable decoder queries, which cannot take advantage of the position information in the selected top K features from the encoder to guide mask predictions. Fig. 2(a)(b)(c) show that the output of Mask2Former in the 0-th decoder layer is far away from the GT mask while Mask DINO outputs much better masks as region proposals. Mask2Former also adopts specialized masked attention to guide the model to attend to regions of interest. However, masked attention is a hard constraint which ignores features outside a provided mask and may overlook important information for following decoder layers. In addition, deformable attention is also a better substitute for its high efficiency allowing attention to be applied to multi-scale features without too much computational overhead. Fig. 2(d)(e) show a predicted mask of Mask2Former in its 1-st decoder layer and the corresponding output of Mask DINO. The prediction of Mask2Former only covers less than half of the GT mask, which means that the attention can not see the whole instance in the next decoder layer. Moreover, a box can also guide deformable attention to a proper region for background stuff, as shown in Fig. 2(f)(g).

⁴We refer the interested readers to discussions in Sec. 3 in DAB-DETR [22]

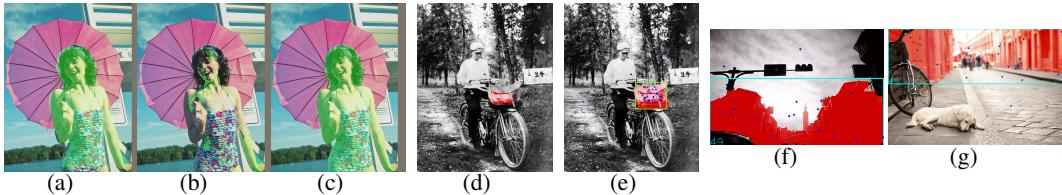


Figure 2: (a) The green transparent region is the ground truth mask for the girl. (b)(c) The predicted masks of the 0-th decoder layer in Mask2Former and Mask DINO, respectively. Note that we attain the predicted masks by first choosing the query which is finally assigned to the ground truth mask in the last decoder layer. Then we visualize the predicted mask of this query by performing dot production with the pixel embedding map. (d)(e) The outputs of the 1-st layer in Mask2Former and Mask DINO. The red masks are predicted masks and the green box is the predicted box by Mask DINO. The blue points are sampled points by deformable attention. Since the 0-th layer of Mask2Former usually outputs unfavorable masks, we avoid using its 0-th layer here. (f)(g) show that Mask DINO can predict correct sampled points, boxes, and masks for background stuffs.

Models	Detection	Segmentation	End-to-End	Feature Extraction	Denoising	Box Helps Mask	Mask Helps Box
Mask-RCNN [12]	✓	Instance		RoI pooling	No	✓	
DETR [2]	✓	Panoptic and instance	✓	Standard attention	No	✓	
DINO [37]	✓	No	✓	Deformable attention	Contrastive DN for box		
HTC [3]	✓	Instance		RoI pooling	No	✓	
Mask2Former [4]		Panoptic, instance, and semantic	✓	Masked attention	No		
Mask DINO (ours)	✓	Panoptic, instance, and semantic	✓	Deformable attention	DN for both mask and box	✓	✓

Table 1: Comparison of representative related models and Mask DINO. “End-to-end” shows whether the model is end-to-end optimized without handcrafted components such as NMS. “Feature extraction” shows how each model pools features. “Denoising” indicates what kind of denoising scheme is used in DINO and Mask DINO, where DN stands for query denoising [18]. “Box helps mask” and “Mask helps box” show mutual task cooperation between the two tasks.

3.3 Our Method: Mask DINO

Mask DINO adopts the same architecture design for detection as in DINO with minimal modifications. In the Transformer decoder, Mask DINO adds a mask branch for segmentation and extends several key components in DINO for segmentation tasks. As shown in Fig. 1, the framework in the blue-shaded part is the original DINO model and the additional design for segmentation is marked with red lines. **Segmentation branch:** Following other unified models [33, 6, 4] for image segmentation, we perform mask classification for all segmentation tasks. Note that DINO is not designed for pixel-level alignment as its positional queries are formulated as anchor boxes and its content queries are used to predict box offset and class membership. To perform mask classification, we adopt a key idea from Mask2Former [4] to construct a pixel embedding map which is obtained from the backbone and Transformer encoder features. As shown in Fig. 1, the pixel embedding map is obtained by fusing the $1/4$ resolution feature map C_b from the backbone with an upsampled $1/8$ resolution feature map C_e from the Transformer encoder. Then we dot-product each content query embedding q_c from the decoder with the pixel embedding map to obtain an output mask m .

$$m = q_c \otimes \mathcal{M}(\mathcal{T}(C_b) + \mathcal{F}(C_e)), \quad (1)$$

where \mathcal{M} is the segmentation head, \mathcal{T} is a convolutional layer to map the channel dimension to the Transformer hidden dimension, and \mathcal{F} is a simple interpolation function to perform $2\times$ upsampling of C_e . This segmentation branch is conceptually simple and easy to implement in the DINO framework, as shown in Fig. 1.

Unified query selection for mask: We extend the box query selection scheme in DINO to Mask DINO. We predict both boxes and masks in the encoder and select the top-ranked ones to initialize decoder queries. The selected masks and boxes serve as better initial anchors for the decoder. Note that we initialize both the content and anchor box queries in Mask DINO whereas DINO only initializes anchor box queries.

Unified denoising for mask: Query denoising in object detection has shown effective [37, 18] to accelerate convergence and improve performance. We also extend this technique to Mask DINO, where we feed the noised ground-truth (GT) boxes and their labels to the decoder and train the model to reconstruct both the GT boxes and masks.

Model	Epochs	Query type	Mask AP	Box AP	AP_{50}^{mask}	AP_{75}^{mask}	AP_S^{mask}	AP_M^{mask}	AP_L^{mask}	GFLOPS	Params	FPS
Mask-RCNN [12, 8, 11], HTC [3]	400 36 36	Dense anchors Dense anchors 300 queries	42.5 39.7 40.6	48.2 44.9 45.6	— 61.4 63.0	— 43.1 44.0	23.8 22.6 23.4	45.0 42.2 42.5	60.0 50.6 52.8	207 441 —	40M 80M —	10.3 5 7.0
DINO-4scale [37]	36	900 queries	—	50.5	—	—	—	—	—	245	47M	19.6
DINO-4scale [37]	36	300 queries	—	49.6	—	—	—	—	—	236	47M	21.0
Mask2Former* [4]	50	100 queries	43.7	46.2 [†]	66.0	46.9	23.4	47.2	64.8	226	44M	8.2
Mask DINO (ours)	50	100 queries	45.4	49.8	67.9	49.3	25.2	48.3	65.8	227	50M	15.2
Mask DINO (ours)	50	300 queries	46.0_(+2.3)	50.5	68.9	50.3	26.0_(+2.6)	49.3 _(+2.1)	65.5 _(+0.7)	234	50M	14.8
Mask DINO (ours)	24	300 queries	44.2_(+0.5)	48.4	66.6	47.9	23.9	47.0	64.0	234	44M	14.8
Mask2Former [4]	12	100 queries	38.7	—	59.8	41.2	18.2	41.5	59.8	226	44M	8.2
Mask DINO (ours)	12	300 queries	41.4_(+2.7)	44.5	62.9	44.6	21.1	44.2	61.4	234	50M	14.8

Table 2: Results for Mask DINO and other object detection and instance segmentation models with a ResNet-50 backbone on COCO val2017. We follow the common practice in DETR-like models to use 300 queries. * Mask2Former using 300 queries is not listed as its performance will degenerate when using 300 queries. [†] indicates the box AP is derived from mask prediction. We test the FPS and GFLOPS of Mask2Former and Mask DINO on the A100 GPU using detectron2.

Model	Epochs	Query type	PQ	PQ^{Th}	PQ^{St}	$Box AP_{pan}^{Th}$	$Mask AP_{pan}^{Th}$
DETR [2]	500 + 25	100 queries	43.4	48.2	36	—	31.1
Panoptic Segformer [19]	24	353 queries	49.6	54.4	42.4	—	41.7
Mask2Former* [4]	50	100 queries	51.9 _{/51.5[†]}	57.7	43.0	—	41.7
Mask DINO (ours)	50	100 queries	52.3	58.3	43.2	47.7	43.7
Mask DINO (ours)	50	300 queries	53.0_(+1.1)	59.1_(+1.4)	43.9_(+0.9)	48.8	44.3_(+2.6)
Mask DINO (ours)	24	300 queries	51.5	57.3	42.6	46.4	42.8
Mask2Former [4]	12	100 queries	46.9	52.5	38.4	—	37.2
Panoptic Segformer [19]	12	353 queries	48.0	52.3	41.5	—	—
Mask DINO (ours)	12	300 queries	49.0_(+1.0)	54.8	40.2	43.2	40.4_(+3.2)

Table 3: Results for Mask DINO and other panoptic segmentation models with a ResNet-50 backbone on COCO val2017. * Mask2Former using 300 queries is not listed as its performance will degenerate when using 300 queries. [†] Our reproduced result.

Hybrid matching: Mask DINO, as in some traditional models [3, 12], predicts boxes and masks with two parallel heads in a loosely coupled manner. Hence the two heads can predict a pair of box and mask that are inconsistent with each other. To address this issue, we consider 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. For example, many "stuff" categories are background like "sky", whose GT mask-derived boxes are highly irregular and often cover the whole image. Therefore, box prediction for these categories can mislead the instance-level ("thing") detection and segmentation. To address this problem, we remove box loss and box matching for "stuff" categories. More specifically, the box prediction pipeline remains the same for "stuff" to locate meaningful regions and extract features with deformable attention. However, we do not count their box prediction loss. In our hybrid matching, the box loss for "stuff" is set to the mean of "thing" categories. This decoupled design can accelerate training and yield additional gains for panoptic segmentation.

We show a comparison between our model and other mainstream methods in Table 1.

4 Experiments

We conduct extensive experiments and compare with several specialized models for four popular tasks including object detection, instance, panoptic, and semantic segmentation on COCO [21], ADE20K [39], and Cityscapes [7]. For all experiments, we use batch size 16 and A100 GPUs with 40GB memory. We use a ResNet-50 [13] and a SwinL [24] backbone for our main results and SOTA model. Under ResNet-50, we use 4 A100 GPUs for all tasks. The detailed setting is in Appendix A.

4.1 Main Results

Instance segmentation and object detection. In Table 2, we compare Mask DINO with other instance segmentation and object detection models. Mask DINO outperforms both the specialized models such as Mask2Former [4] and DINO [37] and hybrid models such as HTC [3] under the same setting. Especially, the instance segmentation results surpass the strong baseline Mask2Former by a large margin (+2.7 AP and +2.3 AP) on the 12-epoch and 50-epoch settings. In addition, Mask DINO significantly improves the convergence speed, outperforming Mask2Former with less than half training epochs (44.2 AP in 24 epochs). We also observe that the performance of Mask2Former

Model	Iterations	Crop size	mIoU (mean)	mIoU (high)	mIoU (reported)
Mask2Former [4]	160k	512	46.1	46.5	47.2
Mask DINO (ours)	160k	512	47.7_(+1.6)	48.7_(+2.2)	48.7_(+1.6)

Table 4: Results for Mask DINO and Mask2Former with 100 queries using a ResNet-50 backbone on ADE20K val. We found the performance variance on this dataset is high and run three times to report both the mean and highest results for both models.

Model	Iterations	mIoU (mean)	mIoU (high)	mIoU (reported)
Mask2Former [4]	90k	78.7	79.0	79.4
Mask DINO (ours)	90k	79.8_(+1.1)	80.0_(+1.0)	80.0_(+0.6)

Table 5: Results for Mask DINO and Mask2Former with 100 queries using a ResNet-50 backbone on Cityscapes val. We found the performance variance on this dataset is high and run three times to report both the mean and highest results for both models.

Method	Params	Backbone	Backbone Pre-training Dataset		Detection Pre-training Dataset	val	
						w/o TTA	w/ TTA
Instance segmentation on COCO							
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)	223M	SwinL	IN-22K-14M	O365	54.5_(+1.1)	—	—
Panoptic segmentation on COCO							
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.

degenerates when using 300 queries, which restricts its scalability. As we use the sparse and soft-constrained deformable attention in the decoder, our model is more computationally efficient. For example, the GFLOPS of Mask DINO is comparable with that of Mask2Former when using more queries, multi-scale features, and an additional box prediction branch.

Panoptic segmentation. We compare Mask DINO with other models in Table 3. Mask DINO outperforms all previous best models on both the 12-epoch and 50-epoch settings by **1.0** PQ and **1.1** PQ, respectively. This indicates Mask DINO has the advantages of both faster convergence and superior performance. One interesting observation is that we outperform Mask2Former [4] in terms of both PQ^{Th} and PQ^{St} . However, instead of using dense and hard-constrained masked attention, we predict boxes and then use them in deformable attention to extract query features. Therefore, our box-oriented deformable attention also works well with "stuff" categories, which makes our unified model simple and efficient. In addition, we improve the mask AP_{pan}^{Th} by **2.6** to 44.3 AP, which is 0.6 higher than the specialized instance segmentation model Mask2Former (43.7 AP).

Semantic segmentation. In Table 4 and 5, we show the performance of semantic segmentation with a ResNet-50 backbone. We use 100 queries for these small datasets. We outperform Mask2Former on both ADE20K and Cityscapes by 1.6 and 0.6 mIoU on the reported performance.

4.2 Comparison with SOTA Models

In Table 6, we compare Mask DINO with SOTA models on three image segmentation tasks to show its scalability. We use the SwinL [24] backbone and pre-train DINO on the Objects365 [31] detection dataset. As Mask DINO is an extension of DINO, the pre-trained DINO model can be used to fine-tune Mask DINO for segmentation tasks. After fine-tuning Mask DINO on the corresponding tasks, we achieve the best results on instance (**54.5** AP), panoptic (**59.4** PQ), and semantic (**60.8** mIoU) segmentation. Compared to SwinV2-G [23], we significantly reduce the model size to 1/15 and backbone pre-training dataset to 1/5. Our detection pre-training also significantly helps all segmentation tasks including panoptic and semantic with "stuff" categories. However, previous specialized segmentation models such as Mask2Former can not use detection datasets, which severely limits their data scalability. By unifying four tasks in one model, we only need to pre-train one model on a large-scale dataset and finetune on all tasks for 10 to 20 epochs (Mask2Former needs 100 epochs), which is more computationally efficient and simpler in model design.

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.

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.

Matching	Box AP			Mask AP
	Box	Mask	12ep	
✓			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.

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.

	Epochs	PQ	PQ ^{thing}	PQ ^{stf}	Box AP _{pan} Th	Mask AP _{pan} Th
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.

	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.

4.3 Ablation Studies

We conduct ablation studies using a ResNet-50 backbone to analyze Mask DINO on COCO val2017. Unless otherwise stated, our experiments are based on object detection and instance segmentation.

Query selection for masks. Table 7 shows the results of our query selection for instance segmentation, where we additionally provide the performance of different decoder layers in one single model. Mask2Former also predicts the masks of learnable queries as initial region proposals. However, their performance lags behind Mask DINO by a large margin (**-38.5AP**). With our effective query selection scheme, the mask performance achieves **39.6 AP** without using the decoder. In addition, our mask performance at layer six is already comparable to the final results with 9 layers.

Feature scales. Mask2Former [4] shows that concatenating multi-scale features as input to Transformer decoder layers does not improve the segmentation performance. However, in Table 8, Mask DINO shows that using more feature scales in the decoder consistently improves the performance.

Decoder layer number. In DINO, increasing the decoder layer number to nine will decrease the performance of box. In Table 9, the result indicates that increasing the number of decoder layers will contribute to both detection and segmentation in Mask DINO. We hypothesize that the multi-task training become more complex and require more decoders to learn the needed mapping function.

Object detection and segmentation help each other. To validate task cooperation in Mask DINO, we use the same model but train different tasks and report the 12 epoch and 50 epoch results. As shown in Table 10, only training one task will lead to a performance drop. Although only training object detection results in faster convergence in the early stage for box prediction, the final performance is still inferior to training both tasks together.

Decoupled box prediction. In Table 11, we show the effectiveness of our decoupled box prediction for panoptic segmentation. This decoupled design of "thing" and "stuff" accelerates training in the early stage (12-epoch setting) and improves the final performance (50-epoch setting).

Matching. In Table 12, we show that only using boxes or masks to perform bipartite matching is not optimal in Mask DINO. A unified matching objective makes the optimization more consistent.

Effectiveness of the algorithm components. In Table 13, we remove each algorithm component at a time and show that each component contributes to the final performance. When we follow DETR [2] to fine-tune segmentation after finishing training detection with its segmentation head, the mask performance drops by a large margin (-5.7 AP). This result also advocates our analysis in Section 3.2 that using the naive DETR segmentation head can not lead to a favorable performance.

5 Conclusion

In this paper, we have presented Mask DINO as a unified Transformer-based framework for both object detection and image segmentation. Conceptually, Mask DINO is a natural extension of DINO from detection to segmentation with minimal modifications on some key components. Mask DINO outperforms previous specialized models and achieves the best results on all three segmentation tasks (instance, panoptic, and semantic). Moreover, Mask DINO shows that detection and segmentation can help each other in query-based models. In particular, Mask DINO enables semantic and panoptic segmentation to benefit from a better visual representation pre-trained on a large-scale detection dataset. We hope Mask DINO can provide insights for enabling task cooperation and data cooperation towards designing a universal model for more vision tasks.

Limitations: Different segmentation tasks fail to achieve mutual assistance in Mask DINO. For example, in COCO panoptic segmentation, the mask AP still lags behind the model only trained with instances. In addition, under the large-scale setting, we have not achieved a new SOTA detection performance as the segmentation head requires additional GPU memory. To accommodate this memory limitation, for the large-scale setting, we have to use smaller image size and less number of queries compared with DINO, which impacts the final performance of object detection. In the future, we will further optimize the implementation to develop a more universal and efficient model to promote task cooperation.

References

- [1] Papers with code - coco test-dev benchmark (instance segmentation).
- [2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.
- [3] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019.
- [4] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2022.
- [5] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. *arXiv preprint arXiv:2104.06404*, 2021.
- [6] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021.
- [7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.
- [8] Xianzhi Du, Barret Zoph, Wei-Chih Hung, and Tsung-Yi Lin. Simple training strategies and model scaling for object detection. *arXiv preprint arXiv:2107.00057*, 2021.
- [9] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.
- [10] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6910–6919, 2021.
- [11] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2918–2928, 2021.

- [12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [14] Jitesh Jain, Anukriti Singh, Nikita Orlov, Zilong Huang, Jiachen Li, Steven Walton, and Humphrey Shi. Semask: Semantically masked transformers for semantic segmentation. *arXiv preprint arXiv:2112.12782*, 2021.
- [15] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6399–6408, 2019.
- [16] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019.
- [17] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020.
- [18] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. *arXiv preprint arXiv:2203.01305*, 2022.
- [19] Zhiqi Li, Wenhui Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, Tong Lu, and Ping Luo. Panoptic segformer. *arXiv preprint arXiv:2109.03814*, 2021.
- [20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020.
- [21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [22] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022.
- [23] Ze Liu, Han Hu, Yutong Lin, Zhiliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. *arXiv preprint arXiv:2111.09883*, 2021.
- [24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.
- [25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [26] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. *arXiv preprint arXiv:2108.06152*, 2021.
- [27] Zipeng Qin, Jianbo Liu, Xiaolin Zhang, Maoqing Tian, Aojun Zhou, Shuai Yi, and Hongsheng Li. Pyramid fusion transformer for semantic segmentation. *arXiv preprint arXiv:2201.04019*, 2022.
- [28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

- [29] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.
- [30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [31] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019.
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [33] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5463–5474, 2021.
- [34] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. *arXiv preprint arXiv:2109.07107*, 2021.
- [35] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8818–8826, 2019.
- [36] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3060–3069, 2021.
- [37] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.
- [38] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34, 2021.
- [39] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.
- [40] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR 2021: The Ninth International Conference on Learning Representations*, 2021.

A Implementation details

A.1 General settings

Dataset and metrics: We evaluate Mask DINO on two challenging datasets: COCO 2017 [21] for object detection, instance segmentation, and panoptic segmentation; ADE20K [39] for semantic segmentation. They both have "thing" and "stuff" categories, therefore we follow the common practice to evaluate object detection and instance segmentation on the "thing" categories and evaluate panoptic and semantic segmentation on the union of the "thing" and "stuff" categories. Unless otherwise stated, all results are trained on the `train` split and evaluated on the `validation` split. For object detection and instance segmentation, the results are evaluated with the standard average precision (AP) and mask AP [21] result. For panoptic segmentation, we evaluate the results with the panoptic quality (PQ) metric [16]. We also report AP_{pan}^{Th} (AP on the "thing" categories) and AP_{pan}^{St} (AP on the "stuff" categories). For semantic segmentation, the results are evaluated with the mean

Intersection-over-Union (mIoU) metric [9].

Backbone: We report results with two public backbones: ResNet-50 [13] and SwinL [24]. To achieve SOTA performance using a large model with the SwinL backbone, we use Objects365 [31] to pre-train an object detection model and then fine-tune the model on the corresponding datasets for all tasks. Though we only pre-train for object detection, our model generalizes well to improve the performance of all segmentation tasks.

Loss function: As we train detection and segmentation tasks jointly, there are totally three kinds of losses, including classification loss \mathcal{L}_{cls} , box loss \mathcal{L}_{box} , and mask loss \mathcal{L}_{mask} . Among them, box loss (L1 loss \mathcal{L}_{L1} and GIOU loss [29] \mathcal{L}_{giou}) and classification loss (focal loss [20]) are the same as DINO [37]. For mask loss, we adopt cross-entropy \mathcal{L}_{ce} and dice loss \mathcal{L}_{dice} . We also follow [17, 5, 4] to use point loss in mask loss for efficiency. Therefore, the total loss is a linear combination of three kinds of losses: $\lambda_{cls}\mathcal{L}_{cls} + \lambda_{L1}\mathcal{L}_{L1} + \lambda_{giou}\mathcal{L}_{giou} + \lambda_{ce}\mathcal{L}_{ce} + \lambda_{dice}\mathcal{L}_{dice}$, where we set $\lambda_{cls} = 4$, $\lambda_{L1} = 5$, $\lambda_{giou} = 2$, $\lambda_{ce} = 5$, and $\lambda_{dice} = 5$.

Basic hyper-parameters: Mask DINO has the same architecture as DINO [37], which is composed of a backbone, a Transformer encoder, and a Transformer decoder. Compared to DINO, we increase the number of decoder layers from six to nine and use 300 queries. We follow Mask-RCNN [12] and Mask2Former [4] to setup the training and inference settings for segmentation tasks. We use batch size 16 and train 50 epoch for COCO segmentation tasks (instance and panoptic), 160K iteration for ADE20K semantic segmentation, and 90K iterations for Cityscapes semantic segmentation. We set the initial learning rate (lr) as 1×10^{-4} and adopt a simple lr scheduler, which drops lr by multiplying 0.1 at the 11-th epoch for the 12-epoch setting and the 20-th epoch for the 24-epoch setting. For the other segmentation settings, we drop the lr at 0.9 and 0.95 fractions of the total number of training steps by multiplying 0.1. Under the ResNet-50 backbone, we use 4 A100 GPUs each with 40GB memory for all tasks. We report the frames-per-second (fps) tested on the same A100 NVIDIA GPU for Mask2Former and Mask DINO by taking the average computing time with batch size 1 on the entire validation set.

Augmentations and Multi-scale setting: We use the same training augmentations as in Mask2Former [4], where the major difference from DINO [37] on COCO is that we use large-scale jittering (LSJ) augmentation [8, 11] and a fixed size crop to 1024×1024 , which also works well for detection tasks. We use the same multi-scale setting as in DINO [37] to use 4 scales in ResNet-50-based models and 5 scales in SwinL-based models.

A.2 Large models setting

For large models with the SwinL backbone, we follow the same setting of DINO [37] to pre-train a model on the Objects365 [31] dataset for object detection. Then we finetune the pre-trained model on COCO instance and panoptic segmentation for 24 epochs and on ADE20K semantic segmentation for 160k iterations. For training settings on instance and panoptic segmentation on COCO, we use $1.2 \times$ larger scale (1280×1280) and 16 A100 GPUs. For training settings on ADE20K semantic, we use $3 \times$ more queries (900) and 8 A100 GPUs. We also use Exponential Moving Average (EMA) in this setting, which helps in ADE20K semantic segmentation.

B SOTA Results on COCO test-dev

Method	Params	Backbone	Backbone Pre-training Dataset	Detection Pre-training Dataset	test	
					w/o TTA	w/ TTA
Instance segmentation on COCO						
Mask2Former [4]	216M	SwinL	IN-22K-14M	—	50.5	—
Soft Teacher [36]	284M	SwinL	IN-22K-14M	O365	-	53.0
SwinV2-G-HTC++ [23]	3.0B	SwinV2-G	IN-22K-ext-70M [23]	O365	-	54.4
MasK DINO(Ours)	223M	SwinL	IN-22K-14M	O365	54.7	—
Panoptic segmentation on COCO						
PQ						
Panoptic SegFormer [19]	—M	SwinL	IN-22K-14M	—	56.2	—
Mask2Former [4]	216M	SwinL	IN-22K-14M	—	58.3	—
MasK DINO (ours)	223M	SwinL	IN-22K-14M	O365	59.5	—

Table 14: Comparison of SOTA models on COCO test-dev. Mask DINO outperforms all existing models. "TTA" means test-time-augmentation. "O365" denotes the Objects365 [31] dataset.