—Supplementary Materials—
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).

{fliay,hzhangcx,hxubr}@connect.ust.hk {liusl20}@mails.tsinghua.edu.cn {leizhang}@idea.edu.cn {ni,hshum}@ust.hk

A. Motivation and Visualization analysis

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. However, in Transformer-based models, specialized models perform much better than unified models. Though detection is a twin task of segmentation, our experiments in Sec. 3 of the paper indicate that trivially extending Transformer-based detection and segmentation models to other tasks achieve inferior performance compared to the original tasks. Therefore, instead of achieving unification by sacrificing performance on each task, Mask DINO aims to develop a unified framework that promotes mutual cooperation between detection and segmentation tasks.

There are mainly three motivations for us to propose Mask DINO. **First**, DINO [21] has achieved remarkable results in object detection. Previous works such as Mask RCNN [8], HTC [2], and DETR [1] 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 3 in the paper (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 [12, 23] 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. 1(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 that 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. 1(d)(e) shows 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. 1(f)(g).

B. Implementation details

The code is available in the supplementary materials. We also provide some detailed descriptions of our implementation here.

B.1. General settings

Dataset and metrics: We evaluate Mask DINO on two challenging datasets: COCO 2017 [15] for object detection, instance segmentation, and panoptic segmentation; ADE20K [22] for semantic segmentation. They both have "thing" and "stuff" categories, therefore, we follow the com-

*Equal contribution.
†Work done when Feng Li and Hao Zhang were interns at IDEA.
‡Corresponding author.

(a) (b) (c) (d) (e) (f) (g)

Figure 1. (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.

mon practice of evaluating 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 [15] result. For panoptic segmentation, we evaluate the results with the panoptic quality (PQ) metric [10]. We also report AP^{Th}_{pan} (AP on the "thing" categories) and AP^{St}_{pan} (AP on the "stuff" categories). For semantic segmentation, the results are evaluated with the mean Intersection-over-Union (mIoU) metric [6].

Backbone: We report results with two public backbones: ResNet-50 [9] and SwinL [17]. To achieve SOTA performance using a large model with the SwinL backbone, we use Objects365 [19] 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 [18] $\mathcal{L}_{giou}$) and classification loss (focal loss [14]) are the same as DINO [21]. For mask loss, we adopt cross-entropy $\mathcal{L}_{ce}$ and dice loss $\mathcal{L}_{dice}$. We also follow [3, 4, 11] 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 [21], 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 [8] and Mask2Former [3] 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 [3], where the major difference from DINO [21] on COCO is that we use large-scale jittering (LSJ) augmentation [5, 7] 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 [21] to use 4 scales in ResNet-50-based models and 5 scales in SwinL-based models.

B.2. Denoising training

Following DN-DETR [12], we train the model to reconstruct the ground-truth objects given the noised ones. These noised objects will be concatenated with the original decoder queries during training, but will be removed during inference. We add noise to both the bounding box and labels, which will serve as positional embedding and content embedding input to decoder queries. As a box can be viewed as a noised version of a segmentation mask, our unified denoising training will reconstruct the masks given the noised boxes, which improves segmentation training.

Label noise: For label noise, we use *label flip*, which ran-

Method	Params	Backbone	Backbone Pre-training Dataset	Detection Pre-training Dataset	test w/o TTA	test w/ TTA
Instance segmentation on COCO					AP	
Mask2Former [3]	216M	SwinL	IN-22K-14M	–	50.5	–
Soft Teacher [20]	284M	SwinL	IN-22K-14M	O365	-	53.0
SwinV2-G-HTC++ [16]	3.0B	SwinV2-G	IN-22K-ext-70M [16]	O365	-	54.4
MasK DINO(Ours)	**223**M	SwinL	IN-22K-14M	O365	**54.7**	–
Panoptic segmentation on COCO					PQ	
Panoptic SegFormer [13]	–M	SwinL	IN-22K-14M	–	56.2	–
Mask2Former [3]	216M	SwinL	IN-22K-14M	–	58.3	–
MasK DINO (ours)	223M	SwinL	IN-22K-14M	O365	**59.5**	–

Table 1. Comparison of SOTA models on COCO `test-dev`. Mask DINO outperforms all existing models. "TTA" means test time augmentation. "O365" denotes the Objects365 [19] dataset.

domly flips a ground-truth label into another possible label in the dataset with probability p. After adding noise, all the labels will go through a label embedding to construct high-dimensional vectors, which will be the content queries of the decoder. p is set to 0.2 in our model.

Box noise: A box can be formulated as (x, y, w, h), which is also the positional query of DINO [21]. We add two kinds of noise to the box, including *center shifting* and *box scaling*. For center shifting, we sample a random perturbation $(\Delta x, \Delta y)$ to the box center. The sampled noise is constrained to $|\Delta x| < \frac{\lambda_1 w}{2}$ and $|\Delta y| < \frac{\lambda_1 h}{2}$, where $\lambda_1 \in (0, 1)$ is a hyperparameter to control the maximum shifting. For box scaling, the width and height of the box are randomly scaled to $[(1 - \lambda_2), (1 + \lambda_2)]$ of the original ones, where λ_2 is also a hyperparameter to control the scaling. In our model, we set $\lambda_1 = \lambda_2 = 0.4$.

C. Large models setting

For large models with the SwinL backbone, we follow the same setting of DINO [21] to pre-train a model on the Objects365 [19] dataset for object detection. Then we fine-tune 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 the 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.

D. SOTA Results on COCO test-dev

We show the COCO test-dev results in Table 1. We achieve **54.7** AP on COCO instance segmentation and **59.5** PQ on COCO panoptic segmentation.

References

[1] 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. 1

[2] 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. 1

[3] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention Mask Transformer for Universal Image Segmentation. 2022. 2, 3

[4] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. *arXiv preprint arXiv:2104.06404*, 2021. 2

[5] 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. 2

[6] 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. 2

[7] 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. 2

[8] 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. 1, 2

[9] 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. 2

[10] 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. 2

[11] 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. 2

[12] 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. 1, 2

[13] Zhiqi Li, Wenhai Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, Tong Lu, and Ping Luo. Panoptic SegFormer. *arXiv preprint arXiv:2109.03814*, 2021. 3

[14] 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. 2

[15] 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. 1, 2

[16] Ze Liu, Han Hu, Yutong Lin, Zhuliang 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. 3

[17] 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. 2

[18] 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. 2

[19] 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. 2, 3

[20] 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. 3

[21] 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. 1, 2, 3

[22] 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. 1

[23] 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. 1