# Supplementary Materials: Masked-attention Mask Transformer for Universal Image Segmentation

Bowen Cheng[1,2*] Ishan Misra[1] Alexander G. Schwing[2] Alexander Kirillov[1] Rohit Girdhar[1]
[1]Facebook AI Research (FAIR) [2]University of Illinois at Urbana-Champaign (UIUC)
https://bowenc0221.github.io/mask2former

## Appendix

We first provide more results for Mask2Former with different backbones as well as test-set performance on standard benchmarks (Appendix A): We use COCO panoptic [16] for panoptic, COCO [19] for instance, and ADE20K [34] for semantic segmentation. Then, we provide more detailed results on additional datasets (Appendix B). Finally, we provide additional ablation studies (Appendix C) and visualization of Mask2Former predictions for all three segmentation tasks (Appendix D).

## A. Additional results

Here, we provide more results of Mask2Former with different backbones on COCO panoptic [16] for panoptic segmentation, COCO [19] for instance segmentation and ADE20K [34] for semantic segmentation. More specifically, for each benckmark, we evaluate Mask2Former with ResNet [15] with 50 and 101 layers, as well as Swin [20] Tiny, Small, Base and Large variants as backbones. We use ImageNet [23] pre-trained checkpoints to initialize backbones.

### A.1. Panoptic segmentation.

In Table I, we report Mask2Former with various backbones on COCO panoptic `val2017`. Mask2Former outperforms *all* existing panoptic segmentation models with various backbones. Our best model sets a new state-of-the-art of 57.8 PQ.

In Table II, we further report the best Mask2Former model on the `test-dev` set. Note that Mask2Former **trained only with the standard `train2017` data**, achieves the *absolute* new state-of-the-art performance on both validation and test set. Mask2Former even outperforms the best COCO competition entry which uses extra training data and test-time augmentation.

*Work done during an internship at Facebook AI Research.

### A.2. Instance segmentation.

In Table III, we report Mask2Former results obtained with various backbones on COCO `val2017`. Mask2Former outperforms the best single-scale model, HTC++ [3, 20]. Note that it is non-trivial to do multi-scale inference for instance-level segmentation tasks without introducing complex post-processing like non-maximum suppression. Thus, we only compare Mask2Former with other single-scale inference models. We believe multi-scale inference can further improve Mask2Former performance and it remains an interesting future work.

In Table IV, we further report the best Mask2Former model on the `test-dev` set. Mask2Former achieves the *absolute* new state-of-the-art performance on both validation and test set. On the one hand, Mask2Former is extremely good at segmenting large objects: we can even outperform the challenge winner (which uses extra training data, model ensemble, *etc*.) on $AP^L$ by a large margin without any bells-and-whistles. On the other hand, the poor performance on small objects leaves room for further improvement in the future.

### A.3. Semantic segmentation.

In Table V, we report Mask2Former results obtained with various backbones on ADE20K `val`. Mask2Former outperforms *all* existing semantic segmentation models with various backbones. Our best model sets a new state-of-the-art of 57.7 mIoU.

In Table VI, we further report the best Mask2Former model on the `test` set. Following [8], we train Mask2Former on the union of ADE20K `train` and `val` set with ImageNet-22K pre-trained checkpoint and use multi-scale inference. Mask2Former is able to outperform previous state-of-the-art methods on all metrics.

## B. Additional datasets

We study Mask2Former on three image segmentation tasks (panoptic, instance and semantic segmentation) using four datasets. Here we report additional results on

| | method | backbone | search space | epochs | PQ | $PQ^{Th}$ | $PQ^{St}$ | $AP^{Th}_{pan}$ | $mIoU_{pan}$ | #params. | FLOPs |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CNN backbones | DETR [2] | R50 | 100 queries | 500+25 | 43.4 | 48.2 | 36.3 | 31.1 | - | - | - |
| | | R101 | 100 queries | 500+25 | 45.1 | 50.5 | 37.0 | 33.0 | - | - | - |
| | K-Net [32] | R50 | 100 queries | 36 | 47.1 | 51.7 | 40.3 | - | - | - | - |
| | Panoptic SegFormer [18] | R50 | 400 queries | 50 | 50.0 | 56.1 | 40.8 | - | - | 47M | 246G |
| | MaskFormer [8] | R50 | 100 queries | 300 | 46.5 | 51.0 | 39.8 | 33.0 | 57.8 | 45M | 181G |
| | | R101 | 100 queries | 300 | 47.6 | 52.5 | 40.3 | 34.1 | 59.3 | 64M | 248G |
| | **Mask2Former** (ours) | R50 | 100 queries | 50 | 51.9 | 57.7 | 43.0 | 41.7 | 61.7 | 44M | 226G |
| | | R101 | 100 queries | 50 | **52.6** | **58.5** | **43.7** | **42.6** | **62.4** | 63M | 293G |
| Transformer backbones | Max-DeepLab [26] | Max-S | 128 queries | 216 | 48.4 | 53.0 | 41.5 | - | - | 62M | 324G |
| | | Max-L | 128 queries | 216 | 51.1 | 57.0 | 42.2 | - | - | 451M | 3692G |
| | Panoptic SegFormer [18] | PVTv2-B5 [28] | 400 queries | 50 | 54.1 | 60.4 | 44.6 | - | - | 101M | 391G |
| | K-Net [32] | Swin-L† | 100 queries | 36 | 54.6 | 60.2 | 46.0 | - | - | - | - |
| | MaskFormer [8] | Swin-T | 100 queries | 300 | 47.7 | 51.7 | 41.7 | 33.6 | 60.4 | 42M | 179G |
| | | Swin-S | 100 queries | 300 | 49.7 | 54.4 | 42.6 | 36.1 | 61.3 | 63M | 259G |
| | | Swin-B | 100 queries | 300 | 51.1 | 56.3 | 43.2 | 37.8 | 62.6 | 102M | 411G |
| | | Swin-B† | 100 queries | 300 | 51.8 | 56.9 | 44.1 | 38.5 | 63.6 | 102M | 411G |
| | | Swin-L† | 100 queries | 300 | 52.7 | 58.5 | 44.0 | 40.1 | 64.8 | 212M | 792G |
| | **Mask2Former** (ours) | Swin-T | 100 queries | 50 | 53.2 | 59.3 | 44.0 | 43.3 | 63.2 | 47M | 232G |
| | | Swin-S | 100 queries | 50 | 54.6 | 60.6 | 45.7 | 44.7 | 64.2 | 69M | 313G |
| | | Swin-B | 100 queries | 50 | 55.1 | 61.0 | 46.1 | 45.2 | 65.1 | 107M | 466G |
| | | Swin-B† | 100 queries | 50 | 56.4 | 62.4 | 47.3 | 46.3 | 67.1 | 107M | 466G |
| | | Swin-L† | *200 queries* | *100* | **57.8** | **64.2** | **48.1** | **48.6** | **67.4** | 216M | 868G |

Table I. **Panoptic segmentation on COCO panoptic `val2017` with 133 categories.** Mask2Former outperforms *all* existing panoptic segmentation models by a large margin with different backbones on all metrics. Our best model sets a new state-of-the-art of 57.8 PQ. Besides PQ for panoptic segmentation, we also report $AP^{Th}_{pan}$ (the AP evaluated on the 80 "thing" categories using *instance segmentation annotation*) and $mIoU_{pan}$ (the mIoU evaluated on the 133 categories for semantic segmentation converted from panoptic segmentation annotation) of the same model trained for panoptic segmentation (**note: we train all our models with panoptic segmentation annotation only**). Backbones pre-trained on ImageNet-22K are marked with †.

| method | backbone | PQ | $PQ^{Th}$ | $PQ^{St}$ | SQ | RQ |
|---|---|---|---|---|---|---|
| Max-DeepLab [26] | Max-L | 51.3 | 57.2 | 42.4 | 82.5 | 61.3 |
| Panoptic FCN [17] | Swin-L | 52.7 | 59.4 | 42.5 | - | - |
| MaskFormer [8] | Swin-L | 53.3 | 59.1 | 44.5 | 82.0 | 64.1 |
| Panoptic SegFormer [18] | PVTv2-B5 [28] | 54.4 | 61.1 | 44.3 | 83.3 | 64.6 |
| K-Net [32] | Swin-L | 55.2 | 61.2 | 46.2 | - | - |
| Megvii (challenge winner) | - | 54.7 | 64.6 | 39.8 | 83.6 | 64.3 |
| **Mask2Former** (ours) | Swin-L | **58.3** | **65.1** | **48.1** | **84.1** | **68.6** |

Table II. **Panoptic segmentation on COCO panoptic `test-dev` with 133 categories.** Mask2Former, without any bells-and-whistles, outperforms the challenge winner (which uses extra training data, model ensemble, *etc*.) on the `test-dev` set. We only train our model on the COCO `train2017` set with ImageNet-22K pre-trained checkpoint.

Cityscapes [10], ADE20K [34] and Mapillary Vistas [22] as well as more detailed training settings.

## B.1. Cityscapes

Cityscapes is an urban egocentric street-view dataset with high-resolution images ($1024 \times 2048$ pixels). It contains 2975 images for training, 500 images for validation and 1525 images for testing with a total of 19 classes.

**Training settings.** For all three segmentation tasks: we use a crop size of $512 \times 1024$, a batch size of 16 and train all models for 90k iterations. During inference, we operate on the whole image ($1024 \times 2048$). Other implementation details largely follow Section 4.1 (panoptic and instance segmentation follow semantic segmentation training settings), except that we use 200 queries for panoptic and instance segmentation models with Swin-L backbone. All other backbones or semantic segmentation models use 100 queries.

**Results.** In Table VII, we report Mask2Former results obtained with various backbones on Cityscapes for three segmentation tasks and compare it with other state-of-the-art methods *without using extra data*. For panoptic segmentation, Mask2Former with Swin-L backbone outperforms the state-of-the-art Panoptic-DeepLab [6] with SWideRnet [5] using single-scale inference. For semantic segmentation, Mask2Former with Swin-B backbone outperforms the state-of-the-art SegFormer [31].

| | method | backbone | search space | epochs | AP | $AP^S$ | $AP^M$ | $AP^L$ | $AP^{boundary}$ | #params. | FLOPs |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CNN backbones | Mask R-CNN [14] | R50 | dense anchors | 36 | 37.2 | 18.6 | 39.5 | 53.3 | 23.1 | 44M | 201G |
| | | R50 | dense anchors | 400 | 42.5 | 23.8 | 45.0 | 60.0 | 28.0 | 46M | 358G |
| | | R101 | dense anchors | 36 | 38.6 | 19.5 | 41.3 | 55.3 | 24.5 | 63M | 266G |
| | | R101 | dense anchors | 400 | 43.7 | **24.6** | 46.4 | 61.8 | 29.1 | 65M | 423G |
| | **Mask2Former** (ours) | R50 | 100 queries | 50 | 43.7 | 23.4 | 47.2 | 64.8 | 30.6 | 44M | 226G |
| | | R101 | 100 queries | 50 | **44.2** | 23.8 | **47.7** | **66.7** | **31.1** | 63M | 293G |
| Transformer backbones | QueryInst [12] | Swin-L† | 300 queries | 50 | 48.9 | 30.8 | 52.6 | 68.3 | 33.5 | - | - |
| | Swin-HTC++ [3, 20] | Swin-B† | dense anchors | 36 | 49.1 | - | - | - | - | 160M | 1043G |
| | | Swin-L† | dense anchors | 72 | 49.5 | **31.0** | 52.4 | 67.2 | 34.1 | 284M | 1470G |
| | **Mask2Former** (ours) | Swin-T | 100 queries | 50 | 45.0 | 24.5 | 48.3 | 67.4 | 31.8 | 47M | 232G |
| | | Swin-S | 100 queries | 50 | 46.3 | 25.3 | 50.3 | 68.4 | 32.9 | 69M | 313G |
| | | Swin-B | 100 queries | 50 | 46.7 | 26.1 | 50.5 | 68.8 | 33.2 | 107M | 466G |
| | | Swin-B† | 100 queries | 50 | 48.1 | 27.8 | 52.0 | 71.1 | 34.4 | 107M | 466G |
| | | Swin-L† | *200 queries* | *100* | **50.1** | 29.9 | **53.9** | **72.1** | **36.2** | 216M | 868G |

Table III. **Instance segmentation on COCO `val2017` with 80 categories.** Mask2Former outperforms strong Mask R-CNN [14] baselines with $8\times$ fewer training epochs for both AP and $AP^{boundary}$ [7] metrics. Our best model is also competitive to the state-of-the-art specialized instance segmentation model on COCO and has higher boundary quality. For a fair comparison, we only consider single-scale inference and models trained using only COCO `train2017` set data. Backbones pre-trained on ImageNet-22K are marked with †.

| method | backbone | AP | AP50 | AP75 | $AP^S$ | $AP^M$ | $AP^L$ |
|---|---|---|---|---|---|---|---|
| QueryInst [12] | Swin-L | 49.1 | 74.2 | 53.8 | **31.5** | 51.8 | 63.2 |
| Swin-HTC++ [3, 20] | Swin-L | 50.2 | - | - | - | - | - |
| Swin-HTC++ [3, 20] (multi-scale) | Swin-L | 51.1 | - | - | - | - | - |
| Megvii (challenge winner) | - | 53.1 | 76.8 | 58.6 | 36.6 | 56.5 | 67.7 |
| **Mask2Former** (ours) | Swin-L | **50.5** | **74.9** | **54.9** | 29.1 | **53.8** | **71.2** |

Table IV. **Instance segmentation on COCO `test-dev` with 80 categories.** Mask2Former is extremely good at segmenting large objects: we can even outperform the challenge winner (which uses extra training data, model ensemble, *etc*.) on $AP^L$ by a large margin without any bells-and-whistles. We only train our model on the COCO `train2017` set with ImageNet-22K pre-trained checkpoint.

### B.2. ADE20K

**Training settings.** For panoptic and instance segmentation, we use the exact same training parameters as we used for semantic segmentation, except that we always use a crop size of $640 \times 640$ for all backbones. Other implementation details largely follow Section 4.1 , except that we use 200 queries for panoptic and instance segmentation models with Swin-L backbone. All other backbones or semantic segmentation models use 100 queries.

**Results.** In Table VIII, we report the results of Mask2Former obtained with various backbones on ADE20K for three segmentation tasks and compare it with other state-of-the-art methods. Mask2Former with Swin-L backbone sets a new state-of-the-art performance on ADE20K for panoptic segmentation. As there are few papers reporting results on ADE20K, we hope this experiment could set up a useful benchmark for future research.

### B.3. Mapillary Vistas

Mapillary Vistas is a large-scale urban street-view dataset with 18k, 2k and 5k images for training, validation and testing. It contains images with a variety of resolutions, ranging from $1024 \times 768$ to $4000 \times 6000$. We only report panoptic and semantic segmentation results for this dataset.

**Training settings.** For both panoptic and semantic segmentation, we follow the same data augmentation of [8]: standard random scale jittering between 0.5 and 2.0, random horizontal flipping, random cropping with a crop size of $1024 \times 1024$ as well as random color jittering. We train our model for 300k iterations with a batch size of 16 using the "poly" learning rate schedule [4]. During inference, we resize the longer side to 2048 pixels. Our panoptic segmentation model with a Swin-L backbone uses 200 queries. All other backbones or semantic segmentation models use 100 queries.

**Results.** In Table IX, we report Mask2Former results obtained with various backbones on Mapillary Vistas for panoptic and semantic segmentation tasks and compare it with other state-of-the-art methods. Our Mask2Former is very competitive compared to state-of-the art specialized models even if it is not designed for Mapillary Vistas.

## C. Additional ablation studies

We perform additional ablation studies of Mask2Former using the same settings that we used in the main paper: a single ResNet-50 backbone [15].

| | method | backbone | crop size | mIoU (s.s.) | mIoU (m.s.) | #params. | FLOPs |
|---|---|---|---|---|---|---|---|
| CNN | MaskFormer [8] | R50 | 512 × 512 | 44.5 | 46.7 | 41M | 53G |
| | | R101 | 512 × 512 | 45.5 | 47.2 | 60M | 73G |
| | **Mask2Former** (ours) | R50 | 512 × 512 | 47.2 | 49.2 | 44M | 71G |
| | | R101 | 512 × 512 | **47.8** | **50.1** | 63M | 90G |
| Transformer backbones | Swin-UperNet [20, 30] | Swin-L† | 640 × 640 | - | 53.5 | 234M | 647G |
| | FaPN-MaskFormer [8, 21] | Swin-L† | 640 × 640 | 55.2 | 56.7 | - | - |
| | BEiT-UperNet [1, 30] | BEiT-L† | 640 × 640 | - | 57.0 | 502M | - |
| | MaskFormer [8] | Swin-T | 512 × 512 | 46.7 | 48.8 | 42M | 55G |
| | | Swin-S | 512 × 512 | 49.8 | 51.0 | 63M | 79G |
| | | Swin-B | 640 × 640 | 51.1 | 52.3 | 102M | 195G |
| | | Swin-B† | 640 × 640 | 52.7 | 53.9 | 102M | 195G |
| | | Swin-L† | 640 × 640 | 54.1 | 55.6 | 212M | 375G |
| | **Mask2Former** (ours) | Swin-T | 512 × 512 | 47.7 | 49.6 | 47M | 74G |
| | | Swin-S | 512 × 512 | 51.3 | 52.4 | 69M | 98G |
| | | Swin-B | 640 × 640 | 52.4 | 53.7 | 107M | 223G |
| | | Swin-B† | 640 × 640 | 53.9 | 55.1 | 107M | 223G |
| | | Swin-L† | 640 × 640 | 56.1 | 57.3 | 215M | 403G |
| | | Swin-L-FaPN† | 640 × 640 | **56.4** | **57.7** | 217M | - |

Table V. **Semantic segmentation on ADE20K `val` with 150 categories.** Mask2Former consistently outperforms MaskFormer [8] by a large margin with different backbones (all Mask2Former models use MSDeformAttn [35] as pixel decoder, except Swin-L-FaPN uses FaPN [21]). Our best model outperforms the best specialized model, BEiT [1], with less than half of the parameters. We report both single-scale (s.s.) and multi-scale (m.s.) inference results. Backbones pre-trained on ImageNet-22K are marked with †.

| method | backbone | P.A. | mIoU | score |
|---|---|---|---|---|
| SETR [33] | ViT-L | 78.35 | 45.03 | 61.69 |
| Swin-UperNet [20, 30] | Swin-L | 78.42 | 47.07 | 62.75 |
| MaskFormer [8] | Swin-L | 79.36 | 49.67 | 64.51 |
| **Mask2Former** (ours) | Swin-L-FaPN | **79.80** | **49.72** | **64.76** |

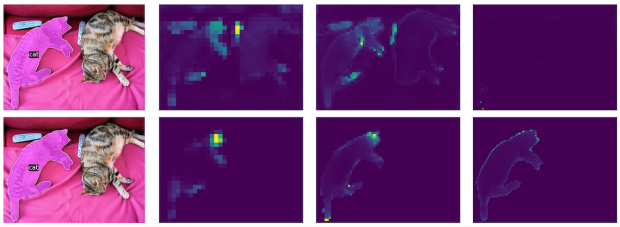
Table VI. **Semantic segmentation on ADE20K `test` with 150 categories.** Mask2Former outperforms previous state-of-the-art methods on all three metrics: pixel accuracy (P.A.), mIoU, as well as the final test score (average of P.A. and mIoU). We train our model on the union of ADE20K `train` and `val` set with ImageNet-22K pre-trained checkpoint following [8] and use multi-scale inference.

## C.1. Convergence analysis

We train Mask2Former with 12, 25, 50 and 100 epochs with either standard scale augmentation (Standard Aug.) [29] or the more recent large-scale jittering augmentation (LSJ Aug.) [11, 13]. As shown in Figure IV, Mask2Former converges in 25 epochs using standard augmentation and almost converges in 50 epochs using large-scale jittering augmentation. This shows that Mask2Former with our proposed Transformer decoder converges faster than models using the standard Transformer decoder: *e.g.*, DETR [2] and MaskFormer [8] require 500 epochs and 300 epochs respectively.

## C.2. Masked attention analysis

We quantitatively and qualitatively analyzed the COCO panoptic model with the R50 backbone. First, we visualize the last three attention maps of our model using cross-attention (Figure Ia top) and masked attention (Figure Ia bottom) of a single query that predicts the "cat." With cross-attention, the attention map spreads over the entire image and the region with highest response is outside the object of interest. We believe this is because the softmax used in cross-attention never attains zero, and small attention weights on large background regions start to dominate. Instead, masked attention limits the attention weights to focus on the object. We validate this hypothesis in Table Ib: we compute the cumulative attention weights on foreground (defined by the matching ground truth to each prediction) and background for all queries on the entire COCO `val`



resolution: 1/32 resolution: 1/16 resolution: 1/8

(a) Visualization of cross-attention (top) and masked attention (bottom) for different resolutions.

| | 1/32 | | 1/16 | | 1/8 | | average | |
|---|---|---|---|---|---|---|---|---|
| | fg | bg | fg | bg | fg | bg | fg | bg |
| cross-attention | 0.23 | 0.77 | 0.23 | 0.77 | 0.15 | 0.85 | 0.20 | 0.80 |
| masked attention | 0.53 | 0.47 | 0.61 | 0.39 | 0.64 | 0.36 | 0.59 | 0.41 |

(b) Cumulative attention weights on foreground (fg) and background (bg) regions for different resolutions.

Figure I. Masked attention analysis.

| method | backbone | panoptic model PQ (s.s.) | PQ (m.s.) | $AP^{Th}_{pan}$ | $mIoU_{pan}$ | instance model AP | AP50 | semantic model mIoU (s.s.) | mIoU (m.s.) |
|---|---|---|---|---|---|---|---|---|---|
| Panoptic-DeepLab [6] | R50 | 60.3 | - | 32.1 | 78.7 | - | - | - | - |
| | X71 [9] | 63.0 | 64.1 | 35.3 | 80.5 | - | - | - | - |
| | SWideRNet [5] | 66.4 | 67.5 | 40.1 | 82.2 | - | - | - | - |
| Panoptic FCN [17] | Swin-L† | 65.9 | - | - | - | - | - | - | - |
| Segmenter [24] | ViT-L† | - | - | - | - | - | - | - | 81.3 |
| SETR [33] | ViT-L† | - | - | - | - | - | - | - | 82.2 |
| SegFormer [31] | MiT-B5 | - | - | - | - | - | - | - | 84.0 |
| **Mask2Former** (ours) | R50 | 62.1 | - | 37.3 | 77.5 | 37.4 | 61.9 | 79.4 | 82.2 |
| | R101 | 62.4 | - | 37.7 | 78.6 | 38.5 | 63.9 | 80.1 | 81.9 |
| | Swin-T | 63.9 | - | 39.1 | 80.5 | 39.7 | 66.9 | 82.1 | 83.0 |
| | Swin-S | 64.8 | - | 40.7 | 81.8 | 41.8 | 70.4 | 82.6 | 83.6 |
| | Swin-B† | 66.1 | - | 42.8 | 82.7 | 42.0 | 68.8 | **83.3** | **84.5** |
| | Swin-L† | **66.6** | - | **43.6** | **82.9** | **43.7** | **71.4** | **83.3** | 84.3 |

Table VII. **Image segmentation results on Cityscapes `val`.** We report both single-scale (s.s.) and multi-scale (m.s.) inference results for PQ and mIoU. All other metrics are evaluated with *single-scale* inference. Since Mask2Former is an end-to-end model, we only use single-scale inference for instance-level segmentation tasks to avoid the need for further post-processing (*e.g.*, NMS).

| method | backbone | panoptic model PQ | $AP^{Th}_{pan}$ | $mIoU_{pan}$ | instance model AP | $AP^S$ | $AP^M$ | $AP^L$ | semantic model mIoU (s.s.) | mIoU (m.s.) |
|---|---|---|---|---|---|---|---|---|---|---|
| MaskFormer [8] | R50 | 34.7 | - | - | - | - | - | - | - | - |
| Panoptic-DeepLab [6] | SWideRNet [5] | 37.9* | - | 50.0* | - | - | - | - | - | - |
| Swin-UperNet [20,30] | Swin-L† | - | - | - | - | - | - | - | - | 53.5 |
| MaskFormer [8] | Swin-L† | - | - | - | - | - | - | - | 54.1 | 55.6 |
| FaPN-MaskFormer [8,21] | Swin-L† | - | - | - | - | - | - | - | 55.2 | 56.7 |
| BEiT-UperNet [1,30] | BEiT-L† | - | - | - | - | - | - | - | - | 57.0 |
| **Mask2Former** (ours) | R50 | 39.7 | 26.5 | 46.1 | 26.4 | 10.4 | 28.9 | 43.1 | 47.2 | 49.2 |
| | Swin-L† | **48.1** | **34.2** | 54.5 | **34.9** | **16.3** | **40.0** | **54.7** | 56.1 | 57.3 |
| | Swin-L-FaPN† | 46.2 | 33.2 | **55.4** | 33.4 | 14.6 | 37.6 | 54.6 | **56.4** | **57.7** |

Table VIII. **Image segmentation results on ADE20K `val`.** Mask2Former is competitive to specialized models on ADE20K. Panoptic segmentation models use single-scale inference by default, multi-scale numbers are marked with *. For semantic segmentation, we report both single-scale (s.s.) and multi-scale (m.s.) inference results.

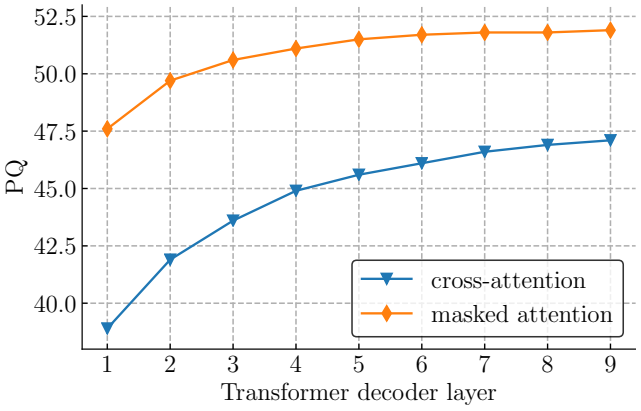



Figure II. Panoptic segmentation performance of each Transformer decoder layer.

set. On average, only 20% of the attention weights in cross-attention focus on the foreground while masked attention increases this ratio to almost 60%. Second, we plot the panoptic segmentation performance using output from each Transformer decoder layer (Figure II). We find masked attention with a single Transformer decoder layer already outperforms cross-attention with 9 layers. We hope the effectiveness of masked attention, together with this analysis, leads to better attention design.

## C.3. Object query analysis

Object queries play an important role in Mask2Former. We ablate different design choices of object queries including the number of queries and making queries learnable.

**Number of queries.** We study the effect of different number of queries for three image segmentation tasks in Table Xa. For instance and semantic segmentation, using 100 queries achieves the best performance, while using 200 queries can further improve panoptic segmentation results. As panoptic segmentation is a combination of instance and semantic segmentation, it has more segments per image than the other two tasks. This ablation suggests that picking the number of queries for Mask2Former may depend on the number of segments per image for a particular task or dataset.

**Learnable queries.** An object query consists of two parts: object query features and object query positional embed-

| method | backbone | panoptic model PQ | panoptic model $\text{mIoU}_{\text{pan}}$ | semantic model mIoU (s.s.) | semantic model mIoU (m.s.) |
|---|---|---|---|---|---|
| Panoptic-DeepLab [6] | ensemble | 42.2* | 58.7* | - | - |
| | SWideRNet [5] | 43.7 | 59.4 | - | - |
| | SWideRNet [5] | 44.8* | 60.0* | - | - |
| Panoptic FCN [17] | Swin-L† | **45.7** | - | - | - |
| MaskFormer [8] | R50 | - | - | 53.1 | 55.4 |
| HMSANet [25] | HRNet [27] | - | - | - | 61.1 |
| **Mask2Former** (ours) | R50 | 36.3 | 50.7 | 57.4 | 59.0 |
| | Swin-L† | **45.5** | **60.8** | **63.2** | **64.7** |

Table IX. **Image segmentation results on Mapillary Vistas `val`.** Mask2Former is competitive to specialized models on Mapillary Vistas. Panoptic segmentation models use single-scale inference by default, multi-scale numbers are marked with *. For semantic segmentation, we report both single-scale (s.s.) and multi-scale (m.s.) inference results.

dings. Object query features are only used as the initial input to the Transformer decoder and are updated through decoder layers; whereas query positional embeddings are added to query features in every Transformer decoder layer when computing the attention weights. In DETR [2], query features are zero-initialized and query positional embeddings are learnable. Furthermore, there is no direct supervision on these query features before feeding them into the Transformer (since they are zero vectors). In our Mask2Former, we still make query positional embeddings learnable. In addition, we make query features learnable as well and directly apply losses on these learnable query features before feeding them into the Transformer decoder.

In Table Xb, we compare our learnable query features with zero-initialized query features in DETR. We find it is important to directly supervise object queries even before feeding them into the Transformer decoder. Learnable queries *without* supervision perform similarly well as zero-initialized queries in DETR.

### C.4. MaskFormer *vs*. Mask2Former

Mask2Former builds upon the same meta architecture as MaskFormer [8] with two major differences: 1) We use more advanced training parameters summarized in Table XIa; and 2) we propose a new Transformer decoder with masked attention, instead of using the standard Transformer decoder, as well as some optimization improvements summarized in Table XIb. To better understand Mask2Former's improvements over MaskFormer, we perform ablation studies on training parameter improvements and Transformer decoder improvements in isolation.

In Table XIc, we study our new training parameters. We train the MaskFormer model with either its original training parameters in [8] or our new training parameters. We observe significant improvements of using our new training parameters for MaskFormer as well. This shows the new training parameters are also generally applicable to other models.

In Table XId, we study our new Transformer decoder.

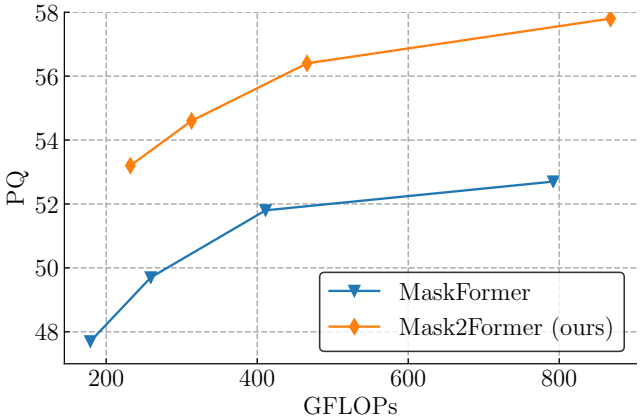



Figure III. MaskFormer [8] *vs*. Mask2Former (ours) with different Swin Transformer backbones.

We train a MaskFormer model and a Mask2Former model with the exact same backbone, *i.e*., a ResNet-50; pixel decoder, *i.e*., a FPN; and training parameters. That is, the only difference is in the Transformer decoder, summarized in Table XIb. We observe improvements for all three tasks, suggesting that the new Transformer decoder itself is indeed better than the standard Transformer decoder.

While computational efficiency was not our primary goal, we find that Mask2Former actually has a better compute-performance trade-off compared to MaskFormer (Figure III). Even the lightest instantiation of Mask2Former outperforms the heaviest MaskFormer instantiation, using $\frac{1}{4}^{\text{th}}$ the FLOPs.

## D. Visualization

We visualize sample predictions of the Mask2Former model with Swin-L [20] backbone on three tasks: COCO panoptic `val2017` set for panoptic segmentation (57.8 PQ) in Figure V, COCO `val2017` set for instance segmentation (50.1 AP) in Figure VI and ADE20K validation set for semantic segmentation (57.7 mIoU, multi-scale inference) in Figure VII.

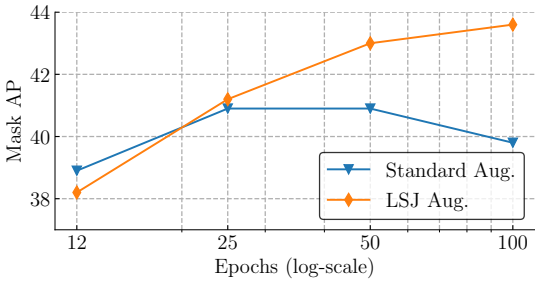



Figure IV. **Convergence analysis.** We train Mask2Former with different epochs using either standard scale augmentation (Standard Aug.) [29] or the more recent large-scale jittering augmentation (LSJ Aug.) [11, 13]. Mask2Former converges in 25 epochs using standard augmentation and almost converges in 50 epochs using large-scale jittering augmentation. Using LSJ also improves performance with longer training epochs (*i.e.*, with more than 25 epochs).

| | AP (COCO) | PQ (COCO) | mIoU (ADE20K) | FLOPs (COCO) |
|---|---|---|---|---|
| 50 | 42.4 | 50.5 | 46.2 | 217G |
| 100 | **43.7** | 51.9 | **47.2** | 226G |
| 200 | 43.5 | **52.2** | 47.0 | 246G |
| 300 | 43.5 | 52.1 | 46.5 | 265G |
| 1000 | 40.3 | 50.7 | 44.8 | 405G |

(a) **Number of queries ablation.** For instance and semantic segmentation, using 100 queries achieves the best performance while using 200 queries can further improve panoptic segmentation results.

| | AP (COCO) | PQ (COCO) | mIoU (ADE20K) | FLOPs (COCO) |
|---|---|---|---|---|
| zero-initialized (DETR [2]) | 42.9 | 51.2 | 45.5 | 226G |
| learnable *w/o* supervision | 42.9 | 51.2 | 47.0 | 226G |
| learnable *w/* supervision | **43.7** | **51.9** | **47.2** | 226G |

(b) **Learnable queries ablation.** It is important to supervise object queries before feeding them into the Transformer decoder. Learnable queries *without* supervision perform similarly well as zero-initialized queries in DETR.

Table X. **Analysis of object queries.** Table Xa: ablation on number of queries. Table Xb: ablation on using learnable queries.

## References


[1] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. *arXiv*, 2021.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 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 *CVPR*, 2019.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *PAMI*, 2018.

[5] Liang-Chieh Chen, Huiyu Wang, and Siyuan Qiao. Scaling wide residual networks for panoptic segmentation. *arXiv:2011.11675*, 2020.

[6] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *CVPR*, 2020.

[7] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *CVPR*, 2021.

[8] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021.

[9] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017.

[10] 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 *CVPR*, 2016.

[11] 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.

[12] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *ICCV*, 2021.

[13] 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 *CVPR*, 2021.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[16] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019.

[17] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Yukang Chen, Lu Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation with point-based supervision. *arXiv preprint arXiv:2108.07682*, 2021.

| training parameters | MaskFormer | Mask2Former (ours) |
|---|---|---|
| learning rate | 0.0001 | 0.0001 |
| weight decay | 0.0001 | 0.05 |
| batch size | 16* | 16 |
| epochs | 75* | 50 |
| data augmentation | standard scale aug. w/ crop | LSJ aug. |
| $\lambda_{\text{cls}}$ | 1.0 | 2.0 |
| $\lambda_{\text{focal}}$ / $\lambda_{\text{ce}}$ | 20.0 / - | - / 5.0 |
| $\lambda_{\text{dice}}$ | 1.0 | 5.0 |
| mask loss | mask | 12544 sampled points |

(a) Comparison of training parameters for MaskFormer [8] and our Mask2Former on the COCO dataset. *: in the original MaskFormer implementation, the model is trained with a batch size of 64 for 300 epochs. We find MaskFormer achieves similar performance when trained with a batch size of 16 for 75 epochs, *i.e.*, the same number of iterations with a smaller batch size.

| Transformer decoder | MaskFormer | Mask2Former (ours) |
|---|---|---|
| # of layers | 6 | 9 |
| single layer | SA-CA-FFN | MA-SA-FFN |
| dropout | 0.1 | 0.0 |
| feature resolution | $\{1/32\} \times 6$ | $\{1/32, 1/16, 1/8\} \times 3$ |
| input query features | zero init. | learnable |
| query p.e. | learnable | learnable |

(b) Comparison of Transformer decoder in MaskFormer [8] and our Mask2Former. SA: self-attention, CA: cross-attention, FFN: feed-forward network, MA: masked attention, p.e.: positional embedding.

| model | training params. | AP (COCO) | PQ (COCO) | mIoU (ADE20K) |
|---|---|---|---|---|
| MaskFormer | MaskFormer | 34.0 | 46.5 | 44.5 |
| MaskFormer | Mask2Former | 37.8 (+3.8) | 48.2 (+1.7) | 45.3 (+0.8) |

(c) Improvements from better **training parameters**.

| Transformer decoder | pixel decoder | AP (COCO) | PQ (COCO) | mIoU (ADE20K) |
|---|---|---|---|---|
| MaskFormer | FPN | 37.8 | 48.2 | 45.3 |
| Mask2Former | FPN | 41.5 (+3.7) | 50.7 (+2.5) | 45.6 (+0.3) |

(d) Improvements from better **Transformer decoder**.

Table XI. **MaskFormer *vs*. Mask2Former.** Table XIa and Table XIb provide an in-depth comparison between MaskFormer and our Mask2Former settings. Table XIc: MaskFormer benefits from our new training parameters as well. Table XId: Comparison between MaskFormer and our Mask2Former with the exact same backbone, pixel decoder and training parameters. The improvements solely come from a better Transformer decoder.


[18] 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.

[19] 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 *ECCV*, 2014.

[20] 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. *arXiv:2103.14030*, 2021.

[21] Shihua Huang Zhichao Lu, Ran Cheng, and Cheng He. Fapn: Feature-aligned pyramid network for dense image prediction. *arXiv*, 2021.

[22] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *CVPR*, 2017.

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[24] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021.

[25] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv:2005.10821*, 2020.

[26] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. MaX-DeepLab: End-to-end panoptic segmentation with mask transformers. In *CVPR*, 2021.

[27] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *PAMI*, 2019.

[28] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer. *arXiv preprint arXiv:2106.13797*, 2021.

[29] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[30] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018.

[31] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021.

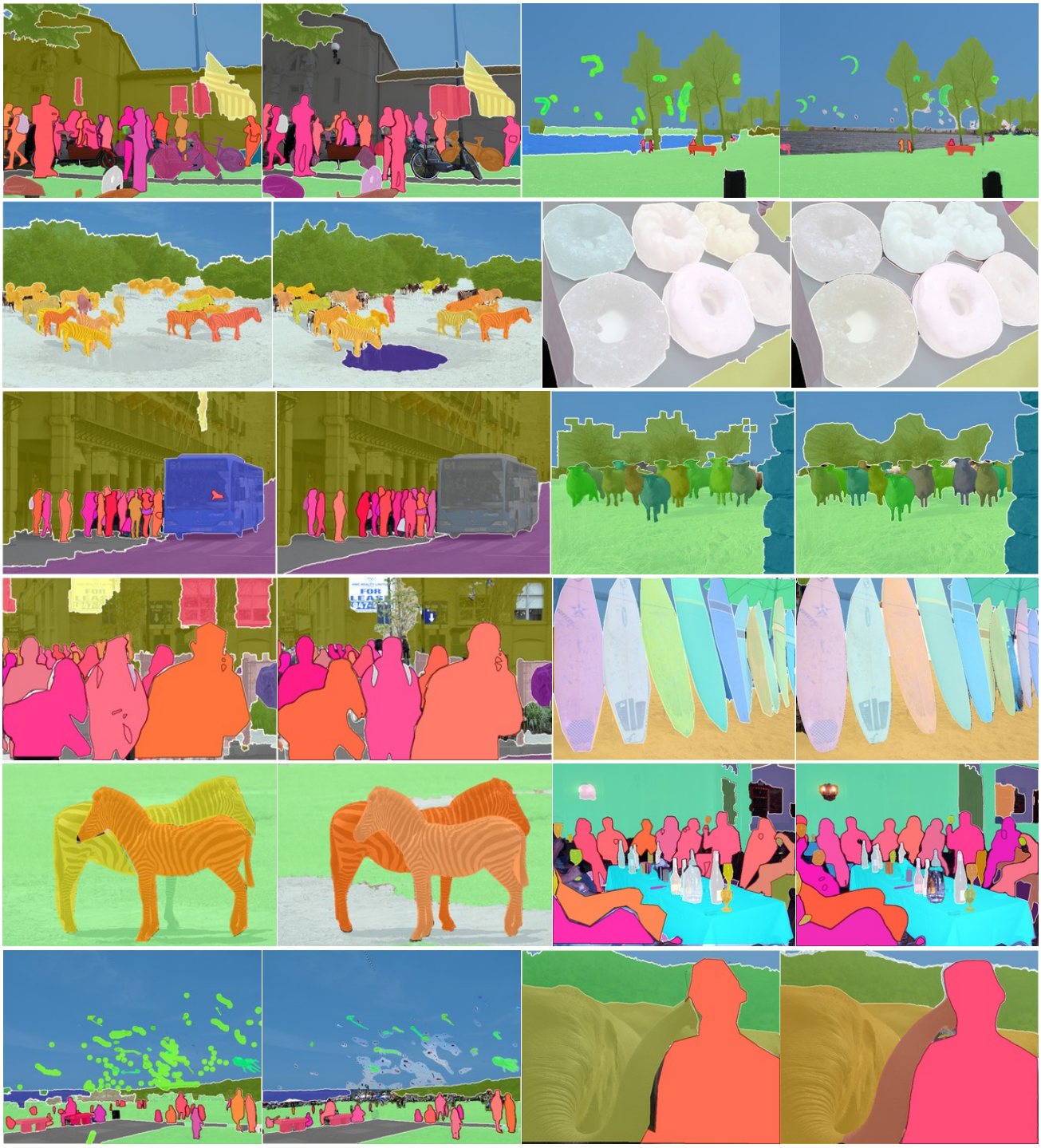


Figure V. Visualization of **panoptic segmentation** predictions on the COCO panoptic dataset: Mask2Former with Swin-L backbone which achieves 57.8 PQ on the validation set. First and third columns: ground truth. Second and fourth columns: prediction. **Last row shows failure cases.**


[32] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. In *NeurIPS*, 2021.

[33] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmen-

Figure VI. Visualization of **instance segmentation** predictions on the COCO dataset: Mask2Former with Swin-L backbone which achieves 50.1 AP on the validation set. First and third columns: ground truth. Second and fourth columns: prediction. **Last row shows failure cases.** We show predictions with confidence scores greater than 0.5.

tation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021.

[34] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017.

[35] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang,

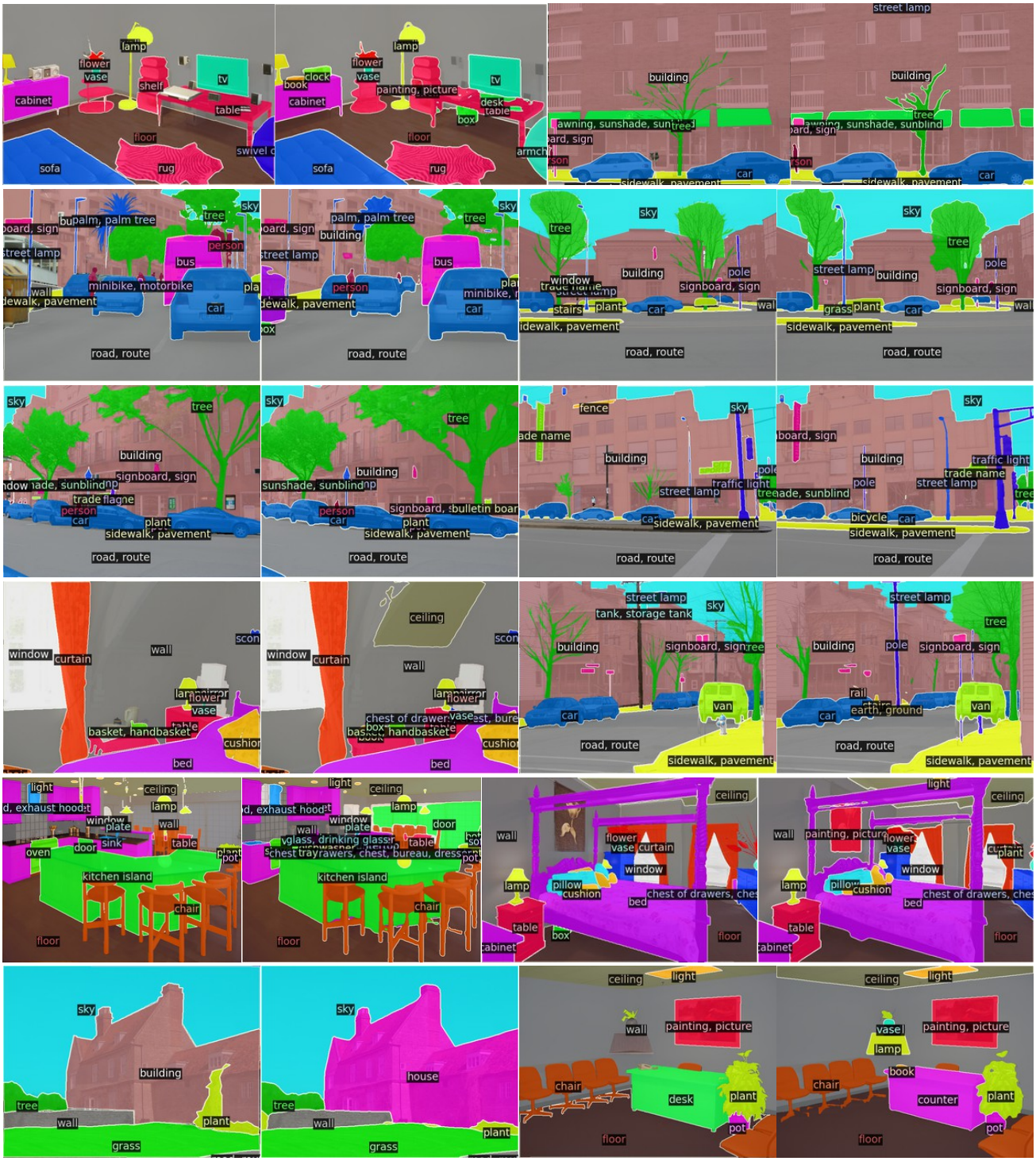


Figure VII. Visualization of **semantic segmentation** predictions on the ADE20K dataset: Mask2Former with Swin-L backbone which achieves 57.7 mIoU (multi-scale) on the validation set. First and third columns: ground truth. Second and fourth columns: prediction. **Last row shows failure cases.**

and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021.