MaX-DeepLab: End-to-End Panoptic Segmentation with Mask Transformers

Huiyu Wang1* Yukun Zhu2 Hartwig Adam2 Alan Yuille1 Liang-Chieh Chen2
1Johns Hopkins University 2Google Research
Code: https://github.com/google-research/deeplab2

Abstract

We present MaX-DeepLab, the first end-to-end model for panoptic segmentation. Our approach simplifies the current pipeline that depends heavily on surrogate sub-tasks and hand-designed components, such as box detection, non-maximum suppression, thing-stuff merging, etc. Although these sub-tasks are tackled by area experts, they fail to comprehensively solve the target task. By contrast, our MaX-DeepLab directly predicts class-labeled masks with a mask transformer, and is trained with a panoptic quality inspired loss via bipartite matching. Our mask transformer employs a dual-path architecture that introduces a global memory path in addition to a CNN path, allowing direct communication with any CNN layers. As a result, MaX-DeepLab shows a significant 7.1% PQ gain in the box-free regime on the challenging COCO dataset, closing the gap between box-based and box-free methods for the first time. A small variant of MaX-DeepLab improves 3.0% PQ over DETR with similar parameters and M-Adds. Furthermore, MaX-DeepLab, without test time augmentation, achieves new state-of-the-art 51.3% PQ on COCO test-dev set.

1. Introduction

The goal of panoptic segmentation [48] is to predict a set of non-overlapping masks along with their corresponding class labels. Modern panoptic segmentation methods address this mask prediction problem by approximating the target task with multiple surrogate sub-tasks. For example, Panoptic-FPN [47] adopts a ‘box-based pipeline’ with three levels of surrogate sub-tasks, as demonstrated in a tree structure in Fig. 1. Each level of this proxy tree involves manually-designed modules, such as anchors [77], box assignment rules [107], non-maximum suppression (NMS) [7], thing-stuff merging [100], etc. Although there are good solutions [77, 12, 33] to individual surrogate sub-tasks and modules, undesired artifacts are introduced when these sub-tasks fit into a pipeline for panoptic segmentation, especially in the challenging conditions (Fig. 2).

Recent work on panoptic segmentation attempted to simplify this box-based pipeline. For example, UPSNet [100] proposes a parameter-free panoptic head, permitting backpropagation to both semantic and instance segmentation...
MaX-DeepLab with a necessary in existing box-based and box-free methods (Tab. 1). In companion with direct training and inference, we remove those hand-coded priors that are necessary for existing methods.

Another line of work made efforts to completely remove boxes from the pipeline, which aligns better with the mask-based definition of panoptic segmentation. The state-of-the-art method in this regime, Axial-DeepLab [91], along with other box-free methods [102, 21, 11], predicts pixel-wise offsets to pre-defined instance centers. But this center-based surrogate sub-task makes it challenging to deal with highly deformable objects, or near-by objects with close centers. As a result, box-free methods do not perform as well as box-based methods on the challenging COCO dataset [60].

In this paper, we streamline the panoptic segmentation pipeline with an end-to-end approach. Inspired by DETR [10], our model directly predicts a set of non-overlapping masks and their corresponding semantic labels with a mask transformer. The output masks and classes are optimized with a panoptic quality (PQ) style objective. Specifically, inspired by the definition of PQ [48], we define a similarity metric between two class-labeled masks as the multiplication of their mask similarity and their class similarity. Our model is trained by maximizing this similarity over a large number of parameters and M-Adds of DETR. We observe that MaX-DeepLab-S outperforms DETR [10] by 7.1% PQ in the box-free regime, and outperforms DetectoRS [95] (with TTA) by 1.7% PQ, bridging the gap between box-based and box-free methods for the first time. For a fair comparison with DETR [10], we also evaluate a lightweight model, MaX-DeepLab-S, that matches the number of parameters and M-Adds of DETR. We observe that MaX-DeepLab-S outperforms DETR by 3.3% PQ on the val set and 3.0% PQ on the test-dev set. In addition, we perform extensive ablation studies and analyses on our end-to-end formulation, model scaling, dual-path architectures, and our loss functions. We also notice that the extra-long training schedule of DETR [10] is not necessary for MaX-DeepLab.

To summarize, our contributions are four-fold:

- MaX-DeepLab is the first end-to-end model for panoptic segmentation, inferring masks and classes directly without hand-coded priors like object centers or boxes.
- We propose a training objective that optimizes a PQ-style loss function via a PQ-style bipartite matching between predicted masks and ground truth masks.
- Our dual-path transformer enables CNNs to read and write a global memory at any layer, providing a new way of combining transformers with CNNs.
- MaX-DeepLab closes the gap between box-based and box-free methods and sets a new state-of-the-art on COCO, even without using test time augmentation.

## 2. Related Work

**Transformers.** Transformers [89], first introduced for neural machine translation, have advanced the state-of-the-art in many natural language processing tasks [27, 79, 26]. Attention [2], as the core component of Transformers, was developed to capture both correspondence of tokens across modalities [2] and long-range interactions in a single context (self-attention) [22, 89]. Later, the complexity of transformer attention has been reduced [49, 92], by introducing local [68] or sparse attention [23], together with a global memory [6, 105, 31, 1]. The global memory, which inspires
our dual-path transformer, recovers long-range context by propagating information globally.

Transformer and attention have been applied to computer vision as well, by combining non-local modules [93, 9] with CNNs or by applying self-attention only [76, 37, 91]. Both classes of methods have boosted various vision tasks such as image classification [18, 5, 76, 37, 57, 91, 28], object detection [93, 80, 76, 36, 10, 110], semantic segmentation [15, 108, 39, 29, 111, 109], video recognition [93, 18], image generation [72, 35], and panoptic segmentation [91]. It is worth mentioning that DETR [10] stacked a transformer on top of a CNN for end-to-end object detection.

Box-based panoptic segmentation. Most panoptic segmentation models, such as Panoptic FPN [47], follow a box-based approach that detects object bounding boxes and predicts a mask for each box, usually with a Mask R-CNN [33] and FPN [58]. Then, the instance segments (‘thing’) and semantic segments (‘stuff’) [13] are fused by merging modules [54, 56, 73, 67, 103] to generate panoptic segmentation. For example, UPSNet [100] developed a parameter-free panoptic head, which facilitates unified training and inference [55]. Recently, DETR [10] extended box-based methods with its transformer-based end-to-end detector.

Box-free panoptic segmentation. Contrary to box-based approaches, box-free methods typically start with semantic segments [12, 14, 16]. Then, instance segments are obtained by grouping ‘thing’ pixels with various methods, such as instance center regression [44, 88, 70, 102, 20], Watershed transform [90, 3, 8], Hough-voting [4, 53, 8], or pixel affinity [45, 66, 81, 30, 8]. Recently, Axial-DeepLab [91] advanced the state-of-the-art by equipping Panoptic-DeepLab [21] with a fully axial-attention [35] backbone. In this work, we extend Axial-DeepLab with a mask transformer for end-to-end panoptic segmentation.

3. Method

In this section, we describe how MaX-DeepLab directly predicts class-labeled masks for panoptic segmentation, followed by the PQ-style loss used to train the model. Then, we introduce our dual-path transformer architecture as well as the auxiliary losses that are helpful in training.

3.1. MaX-DeepLab formulation

The goal of panoptic segmentation is to segment the image \( I \in \mathbb{R}^{H \times W \times 3} \) into a set of class-labeled masks:

\[
\{ y_i \}_{i=1}^K = \{ (m_i, c_i) \}_{i=1}^K . \tag{1}
\]

The \( K \) ground truth masks \( m_i \in \{0, 1\}^{H \times W} \) do not overlap with each other, \( i.e., \sum_{i=1}^K m_i \leq 1^{H \times W} \), and \( c_i \) denotes the ground truth class label of mask \( m_i \).

Our MaX-DeepLab directly predicts outputs in the exact same form as the ground truth. MaX-DeepLab segments the image \( I \) into a fixed-size set of class-labeled masks:

\[
\{ \hat{y}_i \}_{i=1}^N = \{ (\hat{m}_i, \hat{p}_i(c)) \}_{i=1}^N . \tag{2}
\]

The constant size \( N \) of the set is much larger than the typical number of masks in an image [10]. The predicted masks \( \hat{m}_i \in \{0, 1\}^{H \times W} \) are softly exclusive to each other, \( i.e., \sum_{i=1}^N \hat{m}_i = 1^{H \times W} \), and \( \hat{p}_i(c) \) denotes the probability of assigning class \( c \) to mask \( \hat{m}_i \). Possible classes \( \mathbb{C} \ni c \) include thing classes, stuff classes, and a \( \emptyset \) class (no object). In this way, MaX-DeepLab deals with thing and stuff classes in a unified manner, removing the need for merging operators.

Simple inference. End-to-end inference of MaX-DeepLab is enabled by adopting the same formulation for both ground truth definition and model prediction. As a result, the final panoptic segmentation prediction is obtained by simply performing argmax twice. Specifically, the first argmax predicts a class label for each mask:

\[
\hat{c}_i = \arg\max_c \hat{p}_i(c) . \tag{3}
\]

And the other argmax assigns a mask-ID \( \hat{z}_{h,w} \) to each pixel:

\[
\hat{z}_{h,w} = \arg\max_i \hat{m}_{i,h,w} , \quad \forall h \in \{1, 2, \ldots, H\}, \quad \forall w \in \{1, 2, \ldots, W\} . \tag{4}
\]

In practice, we filter each argmax with a confidence threshold—Masks or pixels with a low confidence are removed as described in Sec. 4. In this way, MaX-DeepLab infers panoptic segmentation directly, dispensing with common manually-designed post-processing, \( e.g., \) NMS and thing-stuff merging in almost all previous methods [47, 100]. Besides, MaX-DeepLab does not rely on hand-crafted priors such as anchors, object boxes, or instance mass centers, \( e.t.c \).

3.2. PQ-style loss

In addition to simple inference, MaX-DeepLab enables end-to-end training as well. In this section, we introduce how we train MaX-DeepLab with our PQ-style loss, which draws inspiration from the definition of panoptic quality (PQ) [48]. This evaluation metric of panoptic segmentation, PQ, is defined as the multiplication of a recognition quality (RQ) term and a segmentation quality (SQ) term:

\[
PQ = RQ \times SQ . \tag{5}
\]

Based on this decomposition of PQ, we design our objective in the same manner: First, we define a PQ-style similarity metric between a class-labeled ground truth mask and a predicted mask. Next, we show how we match a predicted mask to each ground truth mask with this metric, and finally how to optimize our model with the same metric.
Mask similarity metric. Our mask similarity metric \( \text{sim}(\cdot, \cdot) \) between a class-labeled ground truth mask \( y_i = (m_i, c_i) \) and a prediction \( \hat{y}_j = (\hat{m}_j, \hat{p}_j(c)) \) is defined as
\[
\text{sim}(y_i, \hat{y}_j) = \hat{p}_j(c_i) \times \text{Dice}(m_i, \hat{m}_j),
\]
where \( \hat{p}_j(c_i) \in [0, 1] \) is the probability of predicting the correct class (recognition quality) and \( \text{Dice}(m_i, \hat{m}_j) \in [0, 1] \) is the Dice coefficient between a predicted mask \( \hat{m}_j \) and a ground truth \( m_i \) (segmentation quality). The two terms are multiplied together, analogous to the decomposition of PQ.

This mask similarity metric has a lower bound of 0, which means either the class prediction is incorrect, OR the two masks do not overlap with each other. The upper bound, 1, however, is only achieved when the class prediction is correct AND the mask is perfect. The AND gating enables this metric to serve as a good optimization objective for both model training and mask matching.

Mask matching. In order to assign a predicted mask to each ground truth, we solve a one-to-one bipartite matching problem between the prediction set \( \{\hat{y}_j\}_{j=1}^{K} \) and the ground truth set \( \{y_i\}_{i=1}^{N} \). Formally, we search for a permutation of \( N \) elements \( \sigma \in \mathcal{S}_N \) that best assigns the predictions to achieve the maximum total similarity to the ground truth:
\[
\hat{\sigma} = \arg \max_{\sigma \in \mathcal{S}_N} \sum_{i=1}^{K} \text{sim}(y_i, \hat{y}_{\sigma(i)}).
\]
The optimal assignment is computed efficiently with the Hungarian algorithm [51], following prior work [10, 82]. We refer to the \( K \) matched predictions as positive masks which will be optimized to predict the corresponding ground truth masks and classes. The \( (N - K) \) masks left are negatives, which should predict the \( \emptyset \) class (no object).

Our one-to-one matching is similar to DETR [10], but with a different purpose: DETR allows only one positive match in order to remove duplicated boxes in the absence of NMS, while in our case, duplicated or overlapping masks are precluded by design. But in our case, assigning multiple predicted masks to one ground truth mask is problematic too, because multiple masks cannot possibly be optimized to fit a single ground truth mask at the same time. In addition, our one-to-one matching is consistent with the PQ metric, where only one predicted mask can theoretically match \( \text{i.e.}, \) have an IoU over 0.5 with each ground truth mask.

PQ-style loss. Given our mask similarity metric and the mask matching process based on this metric, it is straightforward to optimize model parameters \( \theta \) by maximizing this same similarity metric over matched \( \text{i.e.}, \) positive masks:
\[
\max_{\theta} \sum_{i=1}^{K} \text{sim}(y_i, \hat{y}_{\sigma(i)}) \Leftrightarrow \max_{\theta, \sigma \in \mathcal{S}_N} \sum_{i=1}^{K} \text{sim}(y_i, \hat{y}_{\sigma(i)}).
\]
Substituting the similarity metric (Eq. (6)) gives our PQ-style objective \( \mathcal{L}_{\text{PQ}}^{\text{pos}} \) to be maximized for positive masks:
\[
\max_{\theta} \mathcal{L}_{\text{PQ}}^{\text{pos}} = \sum_{i=1}^{K} \hat{p}_{\sigma(i)}(c_i) \times \text{Dice}(m_i, \hat{m}_{\sigma(i)}).
\]
In practice, we rewrite \( \mathcal{L}_{\text{PQ}}^{\text{pos}} \) into two common loss terms by applying the product rule of gradient and then changing a probability \( \hat{p} \) to a log probability \( \log \hat{p} \). The change from \( \hat{p} \) to \( \log \hat{p} \) aligns with the common cross-entropy loss and scales gradients better in practice for optimization:
\[
\mathcal{L}_{\text{PQ}}^{\text{pos}} = \sum_{i=1}^{K} \frac{\hat{p}_{\sigma(i)}(c_i) \times \text{Dice}(m_i, \hat{m}_{\sigma(i)})}{\text{Dice} loss} + \sum_{i=1}^{K} \frac{\text{Dice}(m_i, \hat{m}_{\sigma(i)}) \times [-\log \hat{p}_{\sigma(i)}(c_i)]}{\text{Cross-entropy loss}},
\]
where the loss weights are constants \( \text{i.e.}, \) no gradient is passed to them. This reformulation provides insights by bridging our objective with common loss functions: Our PQ-style loss is equivalent to optimizing a dice loss weighted by the class correctness and optimizing a cross-entropy loss weighted by the mask correctness. The logic behind this loss is intuitive: we want both of the mask and class to be correct at the same time. For example, if a mask is far off the target, it is a false negative anyway, so we disregard its class. This intuition aligns with the down-weighting of class losses for wrong masks, and vice versa.

Apart from the \( \mathcal{L}_{\text{PQ}}^{\text{pos}} \) for positive masks, we define a cross-entropy term \( \mathcal{L}_{\text{PQ}}^{\text{neg}} \) for negative (unmatched) masks:
\[
\mathcal{L}_{\text{PQ}}^{\text{neg}} = \sum_{i=K+1}^{N} \left[ -\log \hat{p}_{\sigma(i)}(\emptyset) \right].
\]
This term trains the model to predict \( \emptyset \) for negative masks. We balance the two terms by \( \alpha \), as a common practice to weight positive and negative samples [59]:
\[
\mathcal{L}_{\text{PQ}} = \alpha \mathcal{L}_{\text{PQ}}^{\text{pos}} + (1 - \alpha) \mathcal{L}_{\text{PQ}}^{\text{neg}},
\]
where \( \mathcal{L}_{\text{PQ}} \) denotes our final PQ-style loss.

3.3. MaX-DeepLab Architecture

As shown in Fig. 3, MaX-DeepLab architecture includes a dual-path transformer, a stacked decoder, and output heads that predict the masks and classes.

Dual-path transformer. Instead of stacking a transformer on top of a CNN [10], we integrate the transformer and the CNN in a dual-path fashion, with bidirectional communication between the two paths. Specifically, we augment a 2D pixel-based CNN with a 1D global memory of size \( N \) \( \text{i.e.}, \) the total number of predictions \) and propose a transformer block as a drop-in replacement for any CNN.
block or an add-on for a pretrained CNN block. Our transformer block enables all four possible types of communication between the 2D pixel-path CNN and the 1D memory-path: (1) the traditional memory-to-pixel (M2P) attention, (2) memory-to-memory (M2M) self-attention, (3) pixel-to-memory (P2M) feedback attention that allows pixels to read from the memory, and (4) pixel-to-pixel (P2P) self-attention, implemented as axial-attention blocks [39, 35, 91]. We select axial-attention [91] rather than global 2D attention [10, 93, 5] for efficiency on high resolution feature maps. One could optionally approximate the pixel-to-pixel self-attention with a convolutional block that only allows local communication. This transformer design with a memory path besides the main CNN path is termed dual-path transformer. Unlike previous work [10], it allows transformer blocks to be inserted anywhere in the backbone at any resolution. In addition, the P2M feedback attention enables the pixel-path CNN to refine its feature given the memory-path features that encode mask information.

Formally, given a 2D input feature $x^p \in \mathbb{R}^{H \times W \times d_{in}}$ with height $H$, width $W$, channels $d_{in}$, and a 1D global memory feature $x^m \in \mathbb{R}^{N \times d_{in}}$ with length $N$ (i.e., the size of the prediction set). We compute pixel-path queries $q^p$, keys $k^m$, and values $v^m$, by learnable linear projections of the pixel-path feature map $x^p$ at each pixel. Similarly, $q^m$, $k^m$, $v^m$ are computed from $x^m$ with another set of projection matrices. The query (key) and value channels are $d_q$ and $d_v$, for both paths. Then, the output of feedback attention $(P2M)$, $y_a^p \in \mathbb{R}^{d_{out}}$, at pixel position $a$, is computed as

$$y_a^p = \sum_{n=1}^{N} \text{softmax}_n (q_a^p \cdot k_n^m) v_n^m,$$

where the $\text{softmax}_n$ denotes a softmax function applied to the whole memory of length $N$. Similarly, the output of memory-to-pixel $(M2P)$ and memory-to-memory $(M2M)$ attention $y_b^m \in \mathbb{R}^{d_{out}}$, at memory position $b$, is

$$y_b^m = \sum_{n=1}^{N} \text{softmax}_n (q_b^m \cdot k_n^m) v_n^m,$$

where a single softmax is performed over the concatenated dimension of size $(H \cdot W + N)$, inspired by ETC [1].

**Stacked decoder.** Unlike previous work [21, 91] that uses a light-weight decoder, we explore stronger hourglass-style stacked decoders [78, 71, 19]. As shown in Fig. 3, our decoder is stacked $L$ times, traversing output strides $(4, 8, \text{and } 16)$ multiple times. At each decoding resolution, features are fused by simple summation after bilinear resizing. Then, convolutional blocks or transformer blocks are applied, before the decoder feature is sent to the next resolution. This stacked decoder is similar to feature pyramid networks [58, 63, 86, 75] designed for pyramidal anchor predictions [64], but our purpose here is only to aggregate multi-scale features, i.e., intermediate pyramidal features are not directly used for prediction.

**Output heads.** From the memory feature of length $N$, we predict mask classes $\tilde{p}(c) \in \mathbb{R}^{N \times |C|}$ with two fully-connected layers (2FC) and a softmax. Another 2FC head predicts mask feature $f \in \mathbb{R}^{N \times D}$. Similarly, we employ two convolutions (2Conv) to produce a normalized feature $g \in \mathbb{R}^{D \times \frac{H}{4} \times \frac{W}{4}}$ from the decoder output at stride 4. Then, our mask prediction $\hat{m}$ is simply the multiplication of transformer feature $f$ and decoder feature $g$:

$$\hat{m} = \text{softmax}_N (f \cdot g) \in \mathbb{R}^{N \times \frac{H}{4} \times \frac{W}{4}}.$$  

In practice, we use batch norm [41] on $f$ and $(f \cdot g)$ to avoid deliberate initialization, and we bilinear upsample the mask prediction $\hat{m}$ to the original image resolution. Finally, the combination $\{\hat{m}_i, \tilde{p}(c)_i\}_{i=1}^N$ is our mask transformer output to generate panoptic results as introduced in Sec. 3.1.

Our mask prediction head is inspired by CondInst [87] and SOLOv2 [94], which extend dynamic convolution [43, 101] to instance segmentation. However, unlike our end-to-end method, these methods require hand-designed object centers and assignment rules for instance segmentation, and a thing-stuff merging module for panoptic segmentation.
3.4. Auxiliary losses

In addition to the PQ-style loss (Sec. 3.2), we find it beneficial to incorporate auxiliary losses in training. Specifically, we propose a pixel-wise instance discrimination loss that helps cluster decoder features into instances. We also use a per-pixel mask-ID cross-entropy loss that classifies each pixel into N masks, and a semantic segmentation loss. Our total loss function thus consists of the PQ-style loss $L_{PQ}$ and these three auxiliary losses.

**Instance discrimination.** We use a per-pixel instance discrimination loss to help the learning of the feature map $g \in \mathbb{R}^{D \times \frac{H}{4} \times \frac{W}{4}}$. Given a downsampled ground truth mask $m_i \in \{0, 1\}^{\frac{H}{4} \times \frac{W}{4}}$, we first compute a normalized feature embedding $t_{i,:} \in \mathbb{R}^{D}$ for each annotated mask by averaging the feature vectors $g_{i,h,w}$ inside the mask $m_i$:

$$t_{i,:} = \frac{\sum_{h,w} m_i \cdot g_{i,h,w}}{\sum_{h,w} m_i \cdot g_{i,h,w}}, \quad i = 1, 2, \ldots, K. \quad (16)$$

This gives us K instance embeddings $\{t_{i,:}\}_{i=1}^K$ representing K ground truth masks. Then, we let each pixel feature $g_{i,h,w}$ perform an instance discrimination task, i.e., each pixel should correctly identify which mask embedding (out of K) it belongs to, as annotated by the ground truth masks. The contrastive loss at a pixel $(h, w)$ is written as:

$$L_{\text{InstDis}}^{h,w} = - \log \frac{\sum_{i=1}^{K} m_{i,h,w} \exp (t_{i,:} \cdot g_{i,h,w}/\tau)}{\sum_{i=1}^{K} \exp (t_{i,:} \cdot g_{i,h,w}/\tau)}, \quad (17)$$

where $\tau$ denotes the temperature, and note that $m_{i,h,w}$ is non-zero only when pixel $(h, w)$ belongs to the ground truth mask $m_i$. In practice, this per-pixel loss is applied to all instance pixels in an image, encouraging features from the same instance to be similar and features from different instances to be distinct, in a contrastive fashion, which is exactly the property required for instance segmentation.

Our instance discrimination loss is inspired by previous works [98, 96, 40, 17, 32, 46]. However, they discriminate instances either unsupervisedly or with image classes [46], whereas we perform a pixel-wise instance discrimination task, as annotated by panoptic segmentation ground truth.

**Mask-ID cross-entropy.** In Equ. (4), we describe how we infer the mask-ID map given our mask prediction. In fact, we can train this per-pixel classification task by applying a cross-entropy loss on it. This is consistent with the literature [42, 85, 10] that uses a cross-entropy loss together with a dice loss [69] to learn better segmentation masks.

**Semantic segmentation.** We also use an auxiliary semantic segmentation loss to help capture per pixel semantic feature. Specifically, we apply a semantic head [21] on top of the backbone if no stacked decoder is used (i.e., $L = 0$). Otherwise, we connect the semantic head to the first decoder output at stride 4, because we find it helpful to separate the final mask feature $g$ with semantic segmentation.

4. Experiments

We report our main results on COCO, comparing with state-of-the-art methods. Then, we provide a detailed ablation study on the architecture variants and losses. Finally, we analyze how MaX-DeepLab works with visualizations.

**Technical details.** Most of our default settings follow Axial-DeepLab [91]. Specifically, we train our models with 32 TPU cores for 100k (400k for main results) iterations (54 epochs), a batch size of 64, Radam [62] Lookahead [106], a ‘poly’ schedule learning rate of $10^{-3}$ ($3 \times 10^{-4}$ for MaX-DeepLab-L), a backbone learning rate multiplier of 0.1, a weight decay of $10^{-4}$, and a drop path rate [38] of 0.2. We resize and pad images to 641 $\times$ 641 [21, 91] (1025 $\times$ 1025 for main results) for inference and M-Adds calculation. During inference, we set masks with class confidence below 0.7 to void and filter pixels with mask-ID confidence below 0.4. Finally, following previous work [100, 21, 91], we filter stuff masks with an area limit of 4096 pixels, and instance masks with a limit of 256 pixels. In training, we set our PQ-style loss weight (Equ. (12), normalized by $N$) to 3.0, with $\alpha = 0.75$. Our instance discrimination uses $\tau = 0.3$, and a weight of 1.0. We set the mask-ID cross-entropy weight to 0.3, and semantic segmentation weight to 1.0. We use an output size $N = 128$ and $D = 128$ channels. We fill the initial memory with learnable weights [10] (more details and architectures in Sec. A.6).

4.1. Main results

We present our main results on COCO val set and test-dev set [60], with a small model, MaX-DeepLab-S, and a large model, MaX-DeepLab-L.

MaX-DeepLab-S augments ResNet-50 [34] with axial-attention blocks [91] in the last two stages. After pretraining, we replace the last stage with dual-path transformer blocks and use an $L = 0$ (not stacked) decoder. We match parameters and M-Adds to DETR-R101 [10], for fair comparison.

MaX-DeepLab-L stacks an $L = 2$ decoder on top of Wide-ResNet-41 [104, 97, 11]. And we replace all stride 16 residual blocks by our dual-path transformer blocks with wide axial-attention blocks [91]. This large variant is meant to be compared with state-of-the-art results.

**Val set.** In Tab. 2, we report our validation set results and compare with both box-based and box-free panoptic segmentation methods. As shown in the table, our single-scale MaX-DeepLab-S already outperforms all other box-free methods by a large margin of more than 4.5 % PQ, no matter whether other methods use test time augmentation (TTA, usually flipping and multi-scale) or not. Specifically, it surpasses single-scale Panoptic-DeepLab by 8.7% PQ, and
This scaling further improves the single-scale performance to 51.1% PQ, outperforming Axial-DeepLab by 5.4% PQ, with similar M-Adds. We also compare MaX-DeepLab-L with DETR [10] by 7.2% PQ with similar inference M-Adds.

Test-dev set. Our improvements on the val set transfers well to the test-dev set, as shown in Tab. 3. On the test-dev set, we are able to compare with more competitive methods and stronger backbones equipped with group convolution [50, 99], deformable convolution [25], or recursive backbone [65, 75], while we do not use these improvements in our model. In the regime of no TTA, our MaX-DeepLab-L outperforms Axial-DeepLab [91] by 5.4% PQ, and DETR [10] by 3.0% PQ. Our MaX-DeepLab-L, without TTA further attains 51.3% PQ, surpassing Axial-DeepLab with TTA by 7.1% PQ. This result also outperforms the best box-based method DetectoRS [75] with TTA by 1.7% PQ, closing the large gap between box-based and box-free methods on COCO for the first time. Our MaX-DeepLab sets a new state-of-the-art on COCO, even without using TTA.

4.2. Ablation study

In this subsection, we provide more insights by teasing apart the effects of MaX-DeepLab components on the val set. We first define a default baseline setting and then vary each component of it: We augment Wide-ResNet-41 [104, 97, 11] by applying dual-path transformer to all blocks at stride 16, enabling all four types of attention. For faster wall-clock training, we use an $L = 0$ (not stacked) decoder and approximate P2P attention with convolutional blocks.

### Scaling

We first study the scaling of MaX-DeepLab in Tab. 4. We notice that replacing convolutional blocks with Axial-attention blocks, stacking decoder $L$ times, and training with more Iterations.

### Dual-path transformer

Next, we vary attention types of our dual-path transformer and the strides (strides) where we apply transformer blocks. Note that we always apply M2P attention that attaches the transformer to the CNN. And P2P

---

1https://github.com/facebookresearch/detr
attention is already ablated above. As shown in Tab. 5, removing our P2M feedback attention causes a drop of 0.7% PQ. On the other hand, we find MaX-DeepLab robust (-0.6% PQ) to the removal of M2M self-attention. We attribute this robustness to our non-overlapping mask formulation. Note that DETR [10] relies on M2M self-attention to remove duplicated boxes. In addition, it is helpful (+1.0% PQ) to apply transformer blocks to stride 8 also, which is impossible for DETR without our dual-path design. Pushing it further to stride 4 does not show more improvements.

**Loss ablation.** Finally, we ablate our PQ-style loss and auxiliary losses in Tab. 6. We first switch our PQ-style similarity in Equ. (6) from RQ × SQ to RQ + SQ, which differs in the hungarian matching (Equ. (7)) and removes dynamic loss weights in Equ. (10). We observe that RQ + SQ works reasonably well, but RQ × SQ improves 0.8% PQ on top of it, confirming the effect of our PQ-style loss in practice, besides its conceptual soundness. Next, we vary auxiliary losses applied to MaX-DeepLab, without tuning loss weights for remaining losses. Our PQ-style loss alone achieves a reasonable performance of 39.5% PQ. Adding instance discrimination significantly improves PQ\(^{ih}\), showing the importance of a clustered feature embedding. Mask-ID prediction shares the same target with the Dice term in Equ. (10), but helps focus on large masks when the Dice term is overwhelmed by small objects. Combining both of the auxiliary losses leads to a large 5.6% PQ gain. Further multi-tasking with semantic segmentation improves 0.6% PQ, because its class-level supervision helps stuff classes but not instance-level discrimination for thing classes.

![Figure 4. Training curves for (a) validation PQ, (b) average class confidence, \(\hat{p}_{\theta(i)}(c_i)\), of matched masks, (c) average mask dice, \(\text{Dice}(m_i, \hat{m}_{\theta(i)})\), of matched masks, (d) per-pixel instance discrimination accuracy, and (e) per-pixel mask-ID prediction accuracy.](image)

![Figure 5. (b) Pixels of the same instance have similar colors (features), while pixels of different instances have distinct colors. (c) The transformer predicts mask colors (features) and classes.](image)

4.3. Analysis

We provide more insights of MaX-DeepLab by plotting our training curves and visualizing the mask output head.

**Training curves.** We first report the validation PQ curve in Fig. 4(a), with our default ablation model. MaX-DeepLab converges quickly to around 46% PQ within 100k iterations (54 epochs), 1/10 of DETR [10]. In Fig. 4(b) and Fig. 4(c), we plot the characteristics of all matched masks in an image. The matched masks tend to have a better class correctness than mask correctness. Besides, we report per-pixel accuracies for instance discrimination (Fig. 4(d)) and mask-ID prediction (Fig. 4(e)). We see that most pixels learn quickly to find their own instances (out of \(K\)) and predict their own mask-IDs (out of \(N\)). Only 10% of all pixels predict wrong mask-IDs, but they contribute to most of the PQ error.

**Visualization.** In order to intuitively understand the normalized decoder output \(g\), the transformer mask feature \(f\), and how they are multiplied to generate our mask output \(\hat{m}\), we train a MaX-DeepLab with \(D = 3\) and directly visualize the normalized features as RGB colors. As shown in Fig. 5, the decoder feature \(g\) assigns similar colors (or feature vectors) to pixels of the same mask, no matter the mask is a thing or stuff, while different masks are colored differently. Such effective instance discrimination (as colorization) facilitates our simple mask extraction with an inner product.

5. Conclusion

In this work, we have shown for the first time that panoptic segmentation can be trained end-to-end. Our MaX-DeepLab directly predicts masks and classes with a mask
transformer, removing the needs for many hand-designed priors such as object bounding boxes, thing-stuff merging, etc. Equipped with a PQ-style loss and a dual-path transformer, MaX-DeepLab achieves the state-of-the-art result on the challenging COCO dataset, closing the gap between box-based and box-free methods for the first time.

Acknowledgments. We would like to thank Maxwell Collins and Sergey Ioffe for their feedbacks on the paper, Jiquan Ngiam for Hungarian Matching implementation, Siyuan Qiao for DetectoRS segmentation results, Chen Wei for instance discrimination insights, Jieneng Chen for dice loss comments and the support from Google Mobile Vision. This work is supported by Google Research Faculty Award.

A. Appendix

A.1. Panoptic Segmentation Results

Similar to the case study in Fig. 2, we provide more panoptic segmentation results of our MaX-DeepLab-L and compare them to the state-of-the-art box-free method, Axial-DeepLab [91], the state-of-the-art box-based method, DetectoRS [75], and the first Detection Transformer, DETR [10] in Fig. A.1 and Fig. A.2. MaX-DeepLab demonstrates robustness to the challenging cases of similar object bounding boxes and nearby objects with close centers, while other methods make systematic mistakes because of their individual surrogate sub-task design. MaX-DeepLab also shows exceptional mask quality, and performs well in the cases of many small objects. Similar to DETR [10], MaX-DeepLab fails typically when there are too many object masks.

A.2. Runtime

In Tab. A.1, we report the end-to-end runtime (i.e., inference time from an input image to final panoptic segmentation) of MaX-DeepLab on a V100 GPU. All results are obtained by (1) a single-scale input without flipping, and (2) built-in TensorFlow library without extra inference optimization. In the fast regime, MaX-DeepLab-S takes 67 ms with a typical 641×641 input. This runtime includes 5 ms of postprocessing and 15 ms of batch normalization that can be easily optimized. This fast MaX-DeepLab-S does not only outperform DETR-R101 [10], but is also around 2x faster. In the slow regime, the standard MaX-DeepLab-S takes 131 ms with a 1025×1025 input, similar to Panoptic-DeepLab-X71 [21]. This runtime is also similar to our run of the official DETR-R101 which takes 128 ms on a V100, including 63 ms for box detection and 65 ms for the heavy mask decoding.

A.3. Mask Output Slot Analysis

In this subsection, we analyze the statistics of all $N = 128$ mask prediction slots using MaX-DeepLab-L. In Fig. A.3, we visualize the joint distribution of mask slot firings and the classes they predict. We observe that the mask slots have imbalanced numbers of predictions and they specialize on ‘thing’ classes and ‘stuff’ classes. Similar to this Mask-Class joint distribution, we visualize the Mask-Pixel joint distribution by extracting an average mask for each mask slot, as shown in Fig. A.4. Specifically, we resize all COCO [60] validation set panoptic segmentation results to a unit square and take an average of masks that are predicted by each mask slot. We split all mask slots into three categories according to their total firings and visualize mask slots in each category. We observe that besides the class-level specialization, our mask slots also specialize on certain regions of an input image. This observation is similar to DETR [10], but we do not see the pattern that almost all slots have a mode of predicting large image-wide masks.

A.4. Mask Head Visualization

In Fig. 5, we visualize how the mask head works by training a MaX-DeepLab with only $D = 3$ decoder feature channels (for visualization purpose only). Although this extreme setting degrades the performance from 45.7% PQ to 37.8% PQ, it enables us to directly visualize the decoder features as RGB colors. Here in Fig. A.5 we show more examples using this model, together with the corresponding panoptic segmentation results. We see a similar clustering effect of instance colors, which enables our simple mask extraction with just a matrix multiplication (a.k.a. dynamic convolution [87, 94, 43, 101]).

A.5. Transformer Attention Visualization

We also visualize the M2P attention that connects the transformer to the CNN. Specifically, given an input image from COCO validation set, we first select four output masks of interest from the MaX-DeepLab-L panoptic prediction. Then, we probe the attention weights between the four masks and all the pixels, in the last dual-path transformer block. Finally, we colorize the four attention maps with four colors and visualize them in one figure. This process is repeated for two images and all eight attention heads as shown in Fig. A.6. We omit our results for the first transformer block since it is mostly flat. This is expected because the memory feature in the first transformer block is unaware of the pixel-path input image at all. Unlike DETR [10] which focuses on object extreme points for detecting bounding boxes, our MaX-DeepLab attends to individual object (or stuff) masks. This mask-attending property makes MaX-DeepLab relatively robust to nearby objects with similar bounding boxes or close object centers.

---

2https://github.com/facebookresearch/detr
MaX-DeepLab segments the baby with its occluded leg correctly. DetectoRS and DETR merge the two people into one instance, probably because the two people have similar bounding boxes. In addition, DETR introduces artifacts around the head of the horse.

MaX-DeepLab correctly segments all the boards, the zebras, and the people. All other methods fail in these challenging cases of similar bounding boxes and nearby object centers.

MaX-DeepLab generates a high quality mask for the cat, arguably better than the ground truth. Axial-DeepLab predicts cat pixels on the right of the image, as the center of the cat is close to the center of the bike. And DETR misses the cat and introduces artifacts.

MaX-DeepLab also performs well in the presence of many small instances.

Figure A.1. Comparing MaX-DeepLab with other representative methods on the COCO val set. (Colors modified for better visualization).
| Method                  | Backbone      | Input Size | Runtime (ms) | PQ [val] | PQ [test] |
|------------------------|---------------|------------|--------------|----------|----------|
| Fast Regime             |               |            |              |          |          |
| Panoptic-DeepLab [21]   | X-71 [24]     | 641×641    | 74           | 38.9     | 38.8     |
| MaX-DeepLab-S           | MaX-S         | 641×641    | 67           | 46.4     | 46.7     |
| Slow Regime             |               |            |              |          |          |
| DETR [10]               | RN-101        | 1333×800   | 128²         | 45.1     | 46.0     |
| Panoptic-DeepLab [21]   | X-71 [24]     | 1025×1025  | 132          | 39.7     | 39.6     |
| MaX-DeepLab-S           | MaX-S         | 1025×1025  | 131          | 48.4     | 49.0     |

Table A.1. End-to-end runtime. **PQ [val]**: PQ (%) on COCO val set. **PQ [test]**: PQ (%) on COCO test-dev set.

### A.6. More Technical Details

In Fig. A.7, Fig. A.8, and Fig. A.9, we include more details of our MaX-DeepLab architectures. As marked in the figure, we pretrain our model on ImageNet [50]. The pretraining model uses only $P2P$ attention (could be a convolutional residual block or an axial-attention block), without the other three types of attention, the feed-forward network (FFN), or the memory. We directly pretrain with an average pooling followed by a linear layer. This pretrained model is used as a backbone for panoptic segmentation, and it uses the backbone learning rate multiplier we mentioned in Sec. 4. After pretraining the CNN path, we apply (with random initialization) our proposed memory path, including the memory, the three types of attention, the FFNs, the decoding layers, and the output heads for panoptic segmentation. In addition, we employ multi-head attention with 8 heads for all attention operations. In MaX-DeepLab-L, we use shortcuts in the stacked decoder. Specifically, each decoding stage (resolution) is connected to the nearest two previous decoding stage outputs of the same resolution.
Similar to DETR [10], MaX-DeepLab fails typically when there are too many masks to segment in an image. This example contains more than 200 masks that should be predicted, mostly people and ties.

In this failure case, MaX-DeepLab mistakes the birds for kites in the sky, probably because the birds are too small.

Figure A.2. Failure cases of MaX-DeepLab on the COCO val set.

Figure A.3. The joint distribution for our $N = 128$ mask slots and 133 classes with 80 ‘thing’ classes on the left and 53 ‘stuff’ classes on the right. We observe that a few mask slots predict a lot of the masks. Some mask slots are used less frequently, probably only when there are a lot of objects in one image. Some other slots do not fire at all. In addition, we see automatic functional segregation between ‘thing’ mask slots and ‘stuff’ mask slots, with a few exceptions that can predict both thing and stuff masks.
Figure A.4. The average masks that each mask slot predicts, normalized by image shape. Mask slots are categorized by their total number of firings and sorted from most firings to few firings. We observe spatial clustered patterns, meaning that the mask slots specialize on certain regions of an input image. For example, the most firing mask slot 71, focusing on the center of an image, predicts almost all 80 ‘thing’ classes but ignores ‘stuff’ classes (Fig. A.3). The top three categories are tennis rackets, cats, and dogs. The second firing mask slot 106 segments 14 classes of masks on the bottom of an image, such as road, floor, or dining-tables. The third firing mask slot 125 concentrates 99.9% on walls or trees that are usually on the top of an image. The fourth firing mask slot 69 focuses entirely on the person class and predicts 2663 people in the 5000 validation images.

| Most Firings (sorted) | Mask Slot 71 | Mask Slot 106 | Mask Slot 125 | Mask Slot 69 | Mask Slot 116 | Mask Slot 4 | Mask Slot 27 | Mask Slot 103 |
|---|---|---|---|---|---|---|---|---|
| Medium Firings (sorted) | Mask Slot 84 | Mask Slot 67 | Mask Slot 23 | Mask Slot 101 | Mask Slot 127 | Mask Slot 28 | Mask Slot 105 | Mask Slot 122 |
| Few Firings (sorted) | Mask Slot 25 | Mask Slot 66 | Mask Slot 98 | Mask Slot 110 | Mask Slot 63 | Mask Slot 95 | Mask Slot 40 | Mask Slot 79 |

Figure A.5. More visualizations of the decoder feature $g$ with $D = 3$. Similar to Fig. 5, we observe a clustering effect of instance colors, i.e., pixels of the same instance have similar colors (features) while pixels of different instances have distinct colors. Note that in this extreme case of $D = 3$ (that achieves 37.8% PQ), there are not enough colors for all masks, which causes missing objects or artifacts at object boundaries, but these artifacts do not present in our normal setting of $D = 128$ (that achieves 45.7% PQ).
Figure A.6. Visualizing the transformer $M2P$ attention maps for selected predicted masks. We observe that head 2, together with head 5, 7, and 8, mainly attends to the output mask regions. Head 1, 3, and 4 gather more context from broader regions, such as semantically-similar instances (scene 1 head 1) or mask boundaries (scene 2 head 4). In addition, we see that head 6 does not pay much attention to the pixel-path, except for some minor firings on the playing field and on the table. Instead, it focuses more on $M2M$ self-attention which shares the same softmax with $M2P$ attention (Equ. (14)).
Figure A.7. An example Axial-Block from Axial-DeepLab [91]. This axial-attention bottleneck block consists of two axial-attention layers operating along height- and width-axis sequentially.

Figure A.8. Building blocks for our MaX-DeepLab architectures.
Figure A.9. More detailed MaX-DeepLab architectures. **Pretrain** labels where we use a classification head to pretrain our models on ImageNet [50]. (a) A dual-path transformer block with $C$ intermediate bottleneck channels. (b) The baseline architecture for our ablation studies in Sec. 4.2. (c) MaX-DeepLab-S that matches the number of parameters and M-Adds of DETR-R101-Panoptic [10]. **Axial-Block** (Fig. A.7) is an axial-attention bottleneck block borrowed from Axial-DeepLab-L [91]. (d) MaX-DeepLab-L that achieves the state-of-the-art performance on COCO [60]. **Wide-Axial** is a wide version of Axial-Block with doubled intermediate bottleneck channels, similar to the one used in Axial-DeepLab-XL [91]. (The residual connections are dropped for neatness).
References

[1] Joshua Ainslie, Santiago Ontanon, Chris Alberti, Philip Pham, Aniruddh Ravula, and Sumit Sanghvi. Etc: Encoding long and structured data in transformers. In EMNLP, 2020. 2, 5

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In ICLR, 2015. 2

[3] Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. In CVPR, 2017. 3

[4] Dana H Ballard. Generalizing the hough transform to detect arbitrary shapes. Pattern Recognition, 1981. 3

[5] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In ICCV, 2017. 1

[6] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150, 2020. 2

[7] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In CVPR, 2005. 3

[8] UjwalBonde, Pablo F Alcantarilla, and Stefan Leutenegger. Towards bounding-box free panoptic segmentation. arXiv:2002.07705, 2020. 3

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In ICCV, 2017. 3

[10] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab. In ICCV COCO + Mapillary Joint Recognition Challenge Workshop, 2019. 3

[11] 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. 1, 3, 5, 6, 7, 9, 11

[12] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D Collins, Ekin D Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-Student: Leveraging Longformer for instance segmentation. In CVPR, 2020. 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 16

[13] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In ICLR, 2015. 1, 3

[14] 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. IEEE TPAMI, 2017. 3

[15] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587, 2017. 3

[16] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In CVPR, 2016. 3

[17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In ICML, 2020. 6

[18] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. A 2-nets: Double attention networks. In NeurIPS, 2018. 3

[19] Bowen Cheng, Liang-Chieh Chen, Yunchao Wei, Yukun Zhu, Zilong Huang, Jinjun Xiong, Thomas S Huang, Wenhui Hwu, and Honghui Shi. Spgnet: Semantic prediction guidance for scene parsing. In ICCV, 2019. 2, 5

[20] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab. In ICCV COCO + Mapillary Joint Recognition Challenge Workshop, 2019. 3

[21] 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. 1, 3, 5, 6, 7, 9, 11

[22] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. In EMNLP, 2016. 2

[23] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. arXiv:1904.10509, 2019. 2

[24] François Chollet. Xception: Deep learning with depthwise separable convolutions. In CVPR, 2017. 7, 11

[25] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In ICCV, 2017. 7

[26] Jifeng Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-xl: Attention language models beyond a fixed-length context. In ACL, 2019. 2

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL, 2019. 2

[28] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929, 2020. 3

[29] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In CVPR, 2019. 3

[30] Moshe Fine, Yanhu Shan, Yupei Wang, Xin Zhao, Yinan Yu, Ming Yang, and Kaiqi Huang. Ssap: Single-shot instance segmentation with affinity pyramid. In ICCV, 2019. 3

[31] Ankit Gupta and Jonathan Berant. Gmat: Global memory augmentation for transformers. arXiv:2006.03274, 2020. 2

[32] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In CVPR, 2020. 6

[33] Kaiming He, Georgi Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In ICCV, 2017. 1, 3

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016. 6
[35] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. arXiv:1912.12180, 2019. 3, 5
[36] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In CVPR, 2018. 3
[37] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In ICCV, 2019. 3
[38] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In ECCV, 2016. 6
[39] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunhao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In ICCV, 2019. 3, 5
[40] Jyh-Jing Hwang, Stella X Yu, Jianbo Shi, Maxwell D Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chan. SegSort: Segmentation by discriminative sorting of segments. In ICCV, 2019. 6
[41] Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In ICML, 2015. 5
[42] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Waserthal, Gregori Kochler, Tobias Norajitra, Sebastian Wirkert, et al. nnunet: Self-adapting framework for u-net-based medical image segmentation. arXiv:1809.10486, 2018. 6
[43] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In NeurIPS, 2016. 5, 9
[44] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In CVPR, 2018. 3
[45] Margret Keuper, Evgeny Levinkov, Nicolas Bonneel, Guillaume Lavoué, Thomas Brox, and Bjorn Andres. Efficient decomposition of image and mesh graphs by lifted multicut. In ICCV, 2015. 3
[46] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In NeurIPS, 2020. 6
[47] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In CVPR, 2019. 1, 2, 3, 7
[48] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In CVPR, 2019. 1, 2, 3
[49] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reform: The efficient transformer. In ICLR, 2020. 2
[50] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In NeurIPS, 2012. 7, 11, 16
[51] Harold W Kuhn. The hungarian method for the assignment problem. Naval research logistics quarterly, 2(1-2):83–97, 1955. 2, 4
[52] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998. 2
[53] Bastian Leibe, Alex Leonardis, and Bernt Schiele. Combined object categorization and segmentation with an implicit shape model. In Workshop on statistical learning in computer vision, ECCV, 2004. 3
[54] Jie Li, Allan Raventos, Arjun Bhargava, Takaaki Tagawa, and Adrien Gaidon. Learning to fuse things and stuff. arXiv:1812.01192, 2018. 3
[55] Qizhu Li, Xiaojuan Qi, and Philip HS Torr. Unifying training and inference for panoptic segmentation. In CVPR, 2020. 3
[56] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In CVPR, 2019. 3
[57] Yingwei Li, Xiaojie Jin, Jieru Mei, Xiaochen Lian, Linjie Yang, Cihang Xie, Qihang Yu, Yuyin Zhou, Song Bai, and Alan Yuille. Neural architecture search for lightweight non-local networks. In CVPR, 2020. 3
[58] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In CVPR, 2017. 3, 5
[59] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In ICCV, 2017. 4
[60] 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. 2, 6, 9, 16
[61] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In CVPR, 2019. 5
[62] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In ICLR, 2020. 6
[63] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In CVPR, 2018. 5
[64] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In ECCV, 2016. 5
[65] Yudong Liu, Yongtao Wang, Siwei Wang, TingTing Liang, QiJie Zhao, Zhi Tang, and Haibin Ling. Cbnet: A novel composite backbone network architecture for object detection. In AAAI, 2020. 7
[66] Yiding Liu, Siyu Yang, Bin Li, Wengang Zhou, Jizheng Xu, Houqiang Li, and Yan Lu. Affinity derivation and graph merge for instance segmentation. In ECCV, 2018. 3
[67] Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, and Wei Jiang. An end-to-end network for panoptic segmentation. In CVPR, 2019. 3
[68] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In EMNLP, 2015. 2
[102] Tien-Ju Yang, Maxwell D Collins, Yukun Zhu, Jyh-Jing Hwang, Ting Liu, Xiao Zhang, Vivienne Sze, George Papandreou, and Liang-Chieh Chen. Deeperlab: Single-shot image parser. *arXiv:1902.05093*, 2019. 2, 3

[103] Yibo Yang, Hongyang Li, Xia Li, Qijie Zhao, Jianlong Wu, and Zhouchen Lin. Sognet: Scene overlap graph network for panoptic segmentation. In *AAAI*, 2020. 3

[104] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 6, 7

[105] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. In *NeurIPS*, 2020. 2

[106] Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. Lookahead optimizer: k steps forward, 1 step back. In *NeurIPS*, 2019. 6

[107] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *CVPR*, 2020. 1

[108] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Pspnet: Point-wise spatial attention network for scene parsing. In *ECCV*, 2018. 3

[109] Xizhou Zhu, Dazhi Cheng, Zheng Zhang, Stephen Lin, and Jifeng Dai. An empirical study of spatial attention mechanisms in deep networks. In *ICCV*, pages 6688–6697, 2019. 3

[110] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv:2010.04159*, 2020. 3

[111] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *CVPR*, 2019. 3