Supplementary Materials:
Boundary IoU: Improving Object-Centric Image Segmentation Evaluation

Bowen Cheng[1*] Ross Girshick[2] Piotr Dollár[2] Alexander C. Berg[2] Alexander Kirillov[2]

[1]UIUC [2]Facebook AI Research (FAIR)

1. Additional Measure Analysis

Trimap IoU computes IoU for a band around the ground truth boundary and, therefore, it ignores errors away from the ground truth boundary (*e.g.* inner mask prediction errors). We generate pseudo-predictions with such errors by adding holes of random shapes to ground truth masks. In Figure 1 we show that Trimap IoU penalizes inner mask prediction errors less than Mask IoU.

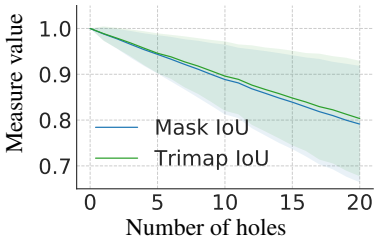

Figure 1: **Sensitivity analysis for Trimap IoU.** The measure penalizes inner mask errors less than Mask IoU.

F-measure matches the pixels of the predicted and ground truth contours if they are within the pixels distance threshold d. In the experiments presented in the main text we observe that this strategy makes F-measure ignore scale type errors for smaller objects. The Mean F-measure (mF-measure) modification ameliorates this limitation by averaging several F-measures with different threshold parameters d. Figure 2 demonstrates the sensitivity curves of this measure for the scale (dilation) error type. For mF-measure we use d from 0.1% to 2.1% image diagonal with 0.4% increment (from 1 pixel to 17 pixels on average) to compare it with Boundary IoU that uses single d set to 2% image diagonal. We observe that mF-measure behaves similarly to Boundary IoU for large objects, however it under-penalizes errors in small objects where Boundary IoU matches Mask IoU behavior. Furthermore, mF-measure is substantially slower than Boundary IoU as it requires to perform the matching of prediction/ground truth pairs several times for different thresholds d.

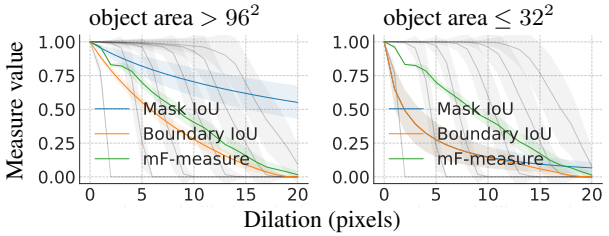

Figure 2: **Sensitivity analysis for mean F-measure (mF-measure).** The measure under-penalizes scale type (dilation) errors in small objects in comparison with Boundary IoU that matches Mask IoU behavior for such objects. F-measure curves for different threshold parameters d are shown in gray.

Boundary IoU can award a perfect score for two non-identical masks (see Figure 3). As discussed in the main text, we observe that Boundary IoU is smaller or equal to Mask IoU in the absolute majority of cases and the inequality is violated when prediction misses interior part of an object (similar to the toy example in Figure 3). To mitigate this limitation, we propose a simple combination of Mask IoU and Boundary IoU by taking their minimum for real-world segmentation evaluation metrics.

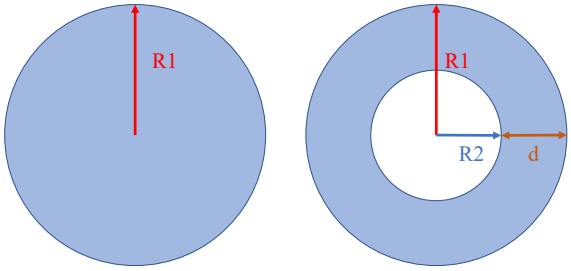

Figure 3: Boundary IoU gives a perfect score for two non-identical masks: a disc mask and a ring mask that has the same center and outer radius as the disk, plus the inner radius that is exactly d pixels smaller than the outer one.

*Work done during an internship at Facebook AI Research.

Mask resolution	Evaluation metric	COCO [14] AP	COCO [14] AP_S	COCO [14] AP_M	COCO [14] AP_L	LVIS*v0.5 [8] AP	LVIS*v0.5 [8] AP_S	LVIS*v0.5 [8] AP_M	LVIS*v0.5 [8] AP_L	Cityscapes [7] AP	Cityscapes [7] AP_S	Cityscapes [7] AP_M	Cityscapes [7] AP_L
28×28	Mask AP	96.5	98.9	95.7	95.1	94.3	96.7	93.7	93.9	93.5	98.4	91.6	91.4
	Boundary AP	85.9	98.9	93.0	73.0	85.5	96.7	91.4	73.1	75.9	98.4	86.4	55.9
56×56	Mask AP	99.5	99.8	99.4	99.3	98.2	98.4	99.0	98.9	98.9	99.7	99.0	98.7
	Boundary AP	95.2	99.8	99.3	89.5	94.6	98.4	98.8	89.9	90.9	99.7	97.7	80.5
112×112	Mask AP	99.9	100.0	99.9	99.9	98.8	98.5	99.7	99.9	99.8	99.9	99.9	99.9
	Boundary AP	99.0	100.0	99.9	97.9	98.1	98.5	99.7	98.3	97.2	99.9	99.9	94.9

Table 1: Boundary AP and Mask AP on COCO `val` set, LVIS*v0.5 `val` set and Cityscapes `val` set for synthetic 28×28, 56×56, and 112×112 predictions generated from the ground truth. Unlike Mask AP, Boundary AP_L successfully captures the lack of fidelity in the synthetic prediction with lower effective resolution for large objects that have area $> 96^2$.

2. Application

2.1. Instance Segmentation

Datasets. We evaluate instance segmentation on three datasets: COCO [14], LVIS [8] and Cityscapes [7].

COCO [14] is the most popular instance segmentation benchmark for common objects. It contains 80 categories. There are 118k images for training, 5k images for validation and 20k images for testing.

LVIS [8] is a federated dataset with more than 1000 categories. It shares the same set of images as COCO but the dataset has higher quality ground truth masks. We use LVISv0.5 version of the dataset. Following [13], we construct the LVIS*v0.5 dataset which keeps only the 80 COCO categories from LVISv0.5. LVIS*v0.5 allows us to compare models trained on COCO using higher quality mask annotations from LVIS (*i.e.* AP* in [13]).

Cityscapes [7] is a street-scene high-resolution dataset. There are 5k images annotated with high quality pixel-level annotations and 8 classes with instance-level segmentation.

Evaluation on synthetic predictions. We simulate predictions by capping the effective resolution of each mask. First, we downscale cropped ground truth masks to a fixed resolution mask with continuous values, we then upscale it back using bilinear interpolation, and finally binarize it. Figure 4 show visualization of the synthetic predictions with different effective resolutions. In Table 1 we compare Mask AP and Boundary AP for the synthetic predictions with different synthetic scales across different datasets.

Evaluation on real predictions. In addition to the experiments with COCO in the main text, we evaluate Mask R-CNN [9], PointRend [13], and Boundary-preserving Mask R-CNN (BMask R-CNN) [5] on LVIS*v0.5 and Cityscapes in Table 2. For each method we feed ground truth boxes to isolate the segmentation quality aspect of the instance segmentation task. On all datasets we observe that Boundary AP better captures improvements in the mask quality.

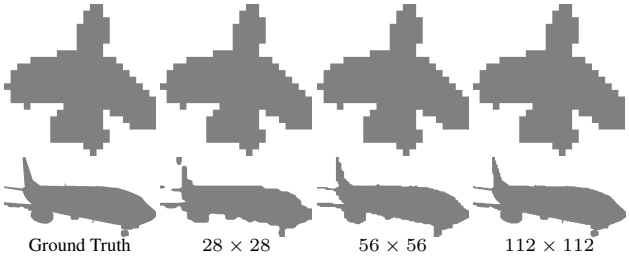

Figure 4: **Synthetic predictions visualization.** First, we downscale cropped ground truth masks to a fixed resolution (from 28×28 to 112×112) mask with continuous values, we then upscale it back using bilinear interpolation, and finally binarize it. Synthetic prediction with low effective resolution are close to the ground truth masks for smaller objects (top row), however the discrepancy grows with object size (bottom row).

Method	AP^{mask}	$AP^{boundary}$	$AP^{boundary}_S$	$AP^{boundary}_M$	$AP^{boundary}_L$
Mask R-CNN	51.5	38.3	45.4	48.7	29.2
PointRend	56.8 (+5.3)	45.9 (+7.6)	49.6 (+4.2)	56.0 (+7.3)	42.2 (**+13.0**)
BMask R-CNN	57.8 (**+6.3**)	46.1 (**+7.8**)	50.8 (**+5.4**)	56.3 (**+7.6**)	40.7 (+11.5)

(a) The models are trained on COCO and evaluated on LVIS*v0.5 `val` set, which has higher annotation quality.

Method	AP^{mask}	$AP^{boundary}$	$AP^{boundary}_S$	$AP^{boundary}_M$	$AP^{boundary}_L$
Mask R-CNN	35.5	16.4	22.3	22.0	9.7
PointRend	42.2 (+6.7)	23.6 (+7.2)	29.9 (+7.6)	29.0 (+7.0)	19.8 (**+10.1**)
BMask R-CNN	43.3 (**+7.8**)	24.0 (**+7.6**)	36.0 (**+13.7**)	29.4 (**+7.4**)	18.7 (+9.0)

(b) The models are trained and evaluated on Cityscapes.

Table 2: Mask R-CNN comparison with the methods designed to improve the mask quality. All models are fed with ground truth boxes. Boundary AP better captures improvements in the mask quality. BMask R-CNN, which outputs 28×28 resolution predictions, outperforms PointRend for smaller objects but trails it for large objects where 224×224 output resolution of PointRend improves boundary quality.

Name	Backbone	LRS	Mask AP						Boundary AP					
			AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN [9]	R50 [10]	1×	35.2	56.3	37.5	17.2	37.7	50.3	21.2	46.4	16.8	17.1	31.6	19.7
	R50 [10]	3×	37.2	58.6	39.9	18.6	39.5	53.3	23.1	49.6	19.0	18.6	33.4	22.2
	R101 [10]	3×	38.6	60.4	41.3	19.5	41.3	55.3	24.5	51.7	20.3	19.4	35.0	23.9
	X101-32×8d [16]	3×	39.5	61.7	42.6	20.7	42.0	56.5	25.4	53.2	21.0	20.6	35.8	24.7
Mask R-CNN with deformable conv. [18]	R50 [10]	1×	37.5	59.4	40.2	18.4	39.7	54.8	22.8	49.6	18.1	18.3	33.4	22.1
	R50 [10]	3×	38.5	60.8	41.1	19.7	40.6	55.7	24.1	51.8	19.4	19.6	34.4	23.4
Mask R-CNN with cascade box head [2]	R50 [10]	1×	36.4	56.9	39.2	17.5	38.7	52.5	22.5	47.9	18.7	17.5	32.6	21.7
	R50 [10]	3×	38.5	59.6	41.5	19.5	41.1	54.5	24.5	51.2	20.8	19.5	34.8	23.6
PointRend [13]	R50 [10]	1×	36.2	56.6	38.6	17.1	38.8	52.5	23.5	48.4	20.2	17.1	33.0	24.1
	R50 [10]	3×	38.3	59.1	41.1	19.1	40.7	55.8	25.4	51.3	22.3	19.1	34.8	26.4
	R101 [10]	3×	40.1	61.1	43.0	20.0	42.9	58.6	27.0	54.1	24.2	19.9	37.0	28.7
	X101-32×8d [16]	3×	41.1	62.8	44.2	21.5	43.8	59.1	28.0	55.6	25.3	21.5	37.8	29.1
BMask R-CNN [5]	R50 [10]	1×	36.6	56.7	39.4	17.3	38.8	53.8	23.5	48.4	20.2	17.2	33.0	24.5
	R50 [10]	3×	38.6	59.2	41.7	19.6	41.1	55.7	25.4	51.4	22.3	19.5	35.2	26.3

Table 3: Boundary AP for recent and classic models on COCO `val`. All models are based on Detectron2. LRS: learning rate schedule, a 1× learning rate schedule refers to 90,000 iterations and a 3× learning rate schedule refers to 270,000 iterations, with batch size 16.

Backbone	Mask AP									Boundary AP								
	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP_r	AP_c	AP_f	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP_r	AP_c	AP_f
R50 [10]	24.4	37.7	26.0	16.7	31.2	41.2	16.0	24.0	28.3	18.8	33.9	18.0	16.7	28.0	19.3	11.9	18.2	22.3
R101 [10]	25.8	39.7	27.3	17.6	33.0	43.7	15.5	26.0	29.6	20.1	35.2	19.8	17.6	29.9	20.8	11.8	20.1	23.5
X101-32×8d [16]	27.0	41.4	28.7	19.0	35.1	43.7	15.4	27.3	31.3	21.4	37.6	21.2	19.0	32.0	21.7	11.2	21.6	25.1

Table 4: Boundary AP of Mask R-CNN baselines on LVISv0.5 `val`. All models are from the Detectron2 model zoo.

Reference Boundary AP evaluation. We provide Boundary AP evaluation for various recent and classic models on COCO (Table 3), LVIS (Table 4), and Cityscapes (Table 5) datasets. We do not train any models ourselves and use the Detectron2 framework [15] or official implementations instead. These results can be used as a reference to simplify the comparison for future methods.

Method	Backbone	Mask AP		Boundary AP	
		AP	AP_{50}	AP	AP_{50}
Mask R-CNN [9]	R50 [10]	33.8	61.5	11.4	37.4
PointRend [13]	R50 [10]	35.9	61.8	16.7	47.2
BMask R-CNN [5]	R50 [10]	36.2	62.6	15.7	46.2
Panoptic-DeepLab [4]	X71 [6]	35.3	57.9	16.5	47.7

Table 5: Boundary AP evaluation on Cityscapes `val` set for models implemented in Detectron2 [15]. Note that we set the dilation width to 0.5% image diagonal for Cityscapes.

2.2. Panoptic Segmentation

The standard evaluation metric for panoptic segmentation is panoptic quality (PQ or Mask PQ) [12], defined as:

$$\text{PQ} = \underbrace{\frac{\sum_{(p,g)\in \text{TP}} \text{IoU}(p,g)}{|\text{TP}|}}_{\text{segmentation quality (SQ)}} \times \underbrace{\frac{|\text{TP}|}{|\text{TP}| + \frac{1}{2}|\text{FP}| + \frac{1}{2}|\text{FN}|}}_{\text{recognition quality (RQ)}}$$

Mask IoU is presented in two places: (1) calculating the average Mask IoU for true positives in Segmentation Quality (SQ) component and (2) matching prediction and ground truth masks to split them into true positives, false positives, and false negatives. Similarly to Boundary AP, we replace Mask IoU with **min**(**Mask IoU**, **Boundary IoU**) in both places and refer the new metric as Boundary PQ.

Datasets. We use two popular datasets with panoptic annotation: COCO panoptic [12] and Cityscapes [7].
COCO panoptic [12] combines annotations from COCO instance segmentation [14] and COCO stuff segmentation [1] into a unified panoptic format with no overlaps. COCO panoptic has 80 things and 53 stuff categories.
Cityscapes [7] has 8 thing and 11 stuff categories.

Dataset	Method	Backbone	Mask PQ			Boundary PQ		
			PQ	SQ	RQ	PQ	SQ	RQ
COCO panoptic [12]	Panoptic FPN [11]	R50 [10]	41.5	79.1	50.5	30.8	70.0	41.7
		R101 [10]	43.0	80.0	52.1	32.5	70.9	43.7
		X101-32×8d [16]	44.4	80.4	53.8	33.9	71.4	45.5
	UPSNet [17]	R50 [10]	42.5	78.2	52.4	31.0	68.7	43.3
	DETR [3]	R50 [10]	43.4	79.3	53.8	32.8	71.0	45.2
Cityscapes [7]	UPSNet [17]	R50 [10]	59.4	79.7	73.1	33.4	63.1	51.9
	Panoptic-DeepLab [4]	R50 [10]	59.8	80.0	73.5	36.3	64.3	55.6
		X71 [6]	63.0	81.7	76.2	41.0	65.5	61.7

Table 6: Reference Boundary PQ evaluation for various models on COCO panoptic `val` and Cityscapes `val`.

Similar to the instance segmentation task, we set dilation width to 2% image diagonal for COCO panoptic and 0.5% image diagonal for Cityscapes.

Analysis with synthetic predictions. Following our experimental setup for instance segmentation, we evaluate Boundary PQ on low-fidelity synthetic predictions generated from ground truth annotations to avoid any potential bias toward a specific model. The synthetic predictions are generated by downscaling ground truth panoptic segmentation maps for each image and then upscaling it back using nearest neighbor interpolation in both cases. This image-level generation process ensures a unified treatment of both things and stuff segments following the idea behind the panoptic segmentation task.

In Table 7, we report Panoptic Quality and its two components: Segmentation Quality (SQ) and Recognition Quality (RQ) for synthetic predictions with various downscaling ratios across different datasets. Similar to our findings for AP, Boundary PQ better tracks boundary quality improvements than Mask PQ for panoptic segmentation. Furthermore, we find that the difference between Boundary PQ and Mask PQ is mainly caused by the difference in SQ. This observation confirms that Boundary IoU better tracks the mask quality of predictions and does not significantly change other aspects like the matching procedure between prediction and ground truth segments.

References Boundary PQ evaluation. We provide Boundary PQ evaluation for various models on COCO panoptic and Cityscapes datasets in Table 6. We do not train any models ourselves and use models trained by their authors. These results can be used as a reference to simplify the comparison for future methods.

Downscaling ratio	Evaluation metric	COCO panoptic [12]			Cityscapes [7]		
		PQ	SQ	RQ	PQ	SQ	RQ
8	Mask PQ	62.6	78.5	78.4	66.3	77.8	83.7
	Boundary PQ	52.8	68.1	77.0	47.1	58.6	80.2
4	Mask PQ	81.0	85.9	93.7	84.3	85.7	98.2
	Boundary PQ	76.6	81.4	93.7	75.0	76.3	98.2
2	Mask PQ	92.5	93.4	99.0	94.2	94.2	99.9
	Boundary PQ	90.8	91.6	99.0	90.7	90.7	99.9

Table 7: Boundary PQ and Mask PQ evaluated on COCO panoptic `val` and Cityscapes `val` sets for synthetic prediction with 8, 4, and 2 downscaling ratios generated from the ground truth. Boundary PQ is more sensitive than Mask PQ in its Segmentation Quality (SQ) component while the Recognition Quality (RQ) component is comparable.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *CVPR*, 2018.

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *CVPR*, 2018.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[4] 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.

[5] Tianheng Cheng, Xinggang Wang, Lichao Huang, and Wenyu Liu. Boundary-preserving Mask R-CNN. In *ECCV*, 2020.

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

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[8] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *ICCV*, 2019.

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

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

[11] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, 2019.

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

[13] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. PointRend: Image segmentation as rendering. In *CVPR*, 2020.

[14] 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.

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

[16] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.

[17] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *CVPR*, 2019.

[18] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, 2019.