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)

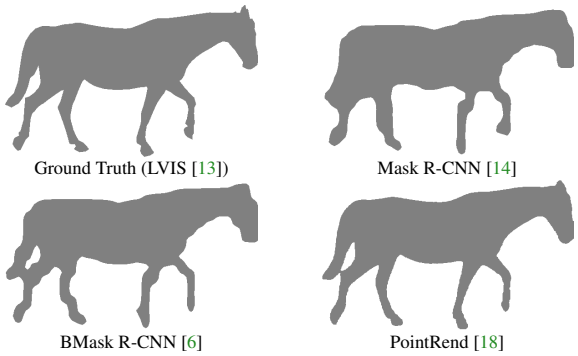
Abstract

We present Boundary IoU (Intersection-over-Union), a new segmentation evaluation measure focused on boundary quality. We perform an extensive analysis across different error types and object sizes and show that Boundary IoU is significantly more sensitive than the standard Mask IoU measure to boundary errors for large objects and does not over-penalize errors on smaller objects. The new quality measure displays several desirable characteristics like symmetry w.r.t. prediction/ground truth pairs and balanced responsiveness across scales, which makes it more suitable for segmentation evaluation than other boundary-focused measures like Trimap IoU and F-measure. Based on Boundary IoU, we update the standard evaluation protocols for instance and panoptic segmentation tasks by proposing the Boundary AP (Average Precision) and Boundary PQ (Panoptic Quality) metrics, respectively. Our experiments show that the new evaluation metrics track boundary quality improvements that are generally overlooked by current Mask IoU-based evaluation metrics. We hope that the adoption of the new boundary-sensitive evaluation metrics will lead to rapid progress in segmentation methods that improve boundary quality.[1]

1. Introduction

The Common Task Framework [23], in which standardized tasks, datasets, and evaluation metrics are used to track research progress, yields impressive results. For example, researchers working on the instance segmentation task, which requires an algorithm to delineate objects with pixel-level binary masks, have improved the standard Average Precision (AP) metric on COCO [24] by an astonishing 86% (relative) from 2015 [9] to 2019 [20].

However, this progress is not equal across all error modes, because different evaluation metrics are sensitive (or insensitive) to different types of errors. If a metric is used for a prolonged time, as in the Common Task Framework, then the corresponding sub-field most rapidly resolves the types of errors to which this metric is sensitive. Research directions that improve other error types typically advance more slowly, as such progress is harder to quantify.



	Mask R-CNN	BMask R-CNN	PointRend
Mask IoU	89%	92% (+3%)	97% (+8%)
Boundary IoU	69%	78% **(+9%)**	91% **(+22%)**

Figure 1: Given the bounding box for a horse, the mask predicted by Mask R-CNN scores a high Mask IoU value (89%) relative to the ground truth despite having low-fidelity, blobby boundaries. The recently proposed BMask R-CNN [6] and PointRend [18] methods predict masks with higher fidelity boundaries, yet these clear visual improvements only marginally improve Mask IoU (+3% and +8%, respectively). In contrast, our proposed **Boundary IoU** measure demonstrates greater sensitivity to boundary errors, and thus provides a clear, quantitative gradient that rewards improvements to boundary segmentation quality.

This phenomenon is at play in instance segmentation, where, among the multitude of papers contributing to the impressive 86% relative improvement in AP (*e.g.*, [36, 4, 1, 16, 21]), only a few address mask boundary quality.

Note that mask boundary quality is an essential aspect of image segmentation, as various downstream applications directly benefit from more precise object segmentations [34, 29, 30]. However, the dominant family of Mask R-CNN-based methods [14] are well-known to predict low-fidelity, blobby masks (see Figure 1). This observation suggests that the current evaluation metrics may have limited sensitivity to mask prediction errors near object boundaries.

To understand why, we start by analyzing Mask Intersection-over-Union (Mask IoU), the underlying mea-

*Work done during an internship at Facebook AI Research.
[1]Project page: https://bowenc0221.github.io/boundary-iou

sure used in AP to compare predicted and ground truth masks. Mask IoU divides the intersection area of two masks by the area of their union. This measure values all pixels equally and, therefore, is less sensitive to boundary quality in *larger objects*: the number of interior pixels grows quadratically in object size and can far exceed the number of boundary pixels, which only grows linearly. In this paper we aim to identify a measure for image segmentation that is sensitive to boundary quality across all scales.

Towards this goal we start by studying standard segmentation measures like Mask IoU and boundary-focused measures such as Trimap IoU [19, 5] and F-measure [26, 8, 28]. We study error-sensitivity characteristics of each measure by generating a variety of error types on top of the high-quality ground truth masks from the LVIS dataset [13]. Our analysis confirms that Mask IoU is less sensitive to errors in larger objects. In addition, the analysis reveals limitations of existing boundary-focused measures, such as asymmetries and instability to small changes in mask quality.

Based on these insights we propose a new **Boundary IoU** measure. Boundary IoU is simple and easy to compute. Instead of considering all pixels, it calculates the intersection-over-union for mask pixels within a certain distance from the corresponding ground truth or prediction boundary contours. Our analysis demonstrates that Boundary IoU measures boundary quality of large objects well, unlike Mask IoU, and it does not over-penalize errors on small objects. An illustrative examples compares Boundary IoU to Mask IoU in Figure 1.

Boundary IoU enables new task-level evaluation metrics. For the task of instance segmentation [24], we propose **Boundary Average Precision** (Boundary AP), and for panoptic segmentation [17], we propose **Boundary Panoptic Quality** (Boundary PQ).

Boundary AP assesses all relevant aspects of instance segmentation, simultaneously taking into account categorization, localization, and segmentation quality, unlike prior boundary-focused metrics for instance segmentation like AF [22] that ignore false positive rates. We test Boundary AP on three common datasets: COCO [24], LVIS [13], and Cityscapes [7]. With real predictions from recent instance segmentation methods that directly aim to improve boundary quality [18, 6], we verify that Boundary AP tracks improvements better than Mask AP. With synthetic predictions, we show that Boundary AP is significantly more sensitive to large-object boundary quality than Mask AP.

For panoptic segmentation, we apply Boundary PQ to the COCO [17] and Cityscapes [7] panoptic datasets. We test the new metric with synthetic predictions and show that it is more sensitive than the previous metric based on Mask IoU. Finally, we evaluate the performance of various recent instance and panoptic segmentation models with the new evaluation metrics to ease comparison for future research.

These new metrics reveal improvements in boundary quality that are generally ignored by Mask IoU-based evaluation metrics. We hope that the adoption of these new boundary-sensitive evaluations can enable faster progress towards segmentation models with better boundary quality.

2. Related Work and Preliminaries

Image segmentation tasks like semantic, instance, or panoptic segmentation are evaluated by comparing segmentation masks predicted by a system to ground truth masks provided by annotators. Modern evaluation metrics for these tasks are based on segmentation quality measures that evaluate consistency between ground truth object shape G and prediction shape P represented by binary masks of a fixed resolution. We define the most common segmentation quality measures and the new Boundary IoU measure in Table 1 using the unified notation presented in Table 2. We split the measures into mask- and boundary-based types and discuss their differences next.

Mask-based segmentation measures take into account all pixels of an object mask. The first PASCAL VOC semantic segmentation track in 2007 [11] used Pixel Accuracy measure to evaluate predictions. For each class it calculates the ratio of correctly labeled ground truth pixels (see Table 1). Pixel accuracy is not symmetric and biased toward prediction masks that are larger than ground truth masks. Subsequently, PASCAL VOC [10] switched its evaluation to the Mask Intersection-over-Union (Mask IoU) measure.

Mask IoU segmentation consistency measure divides the number of pixels in the intersection of the prediction and ground truth masks by the number of pixels in their union (see Table 1). The measure is widely used in the evaluation metrics for most popular semantic, instance, and panoptic segmentation tasks [10, 24, 17] and datasets [7, 3, 35, 24]. Unlike Pixel Accuracy, Mask IoU is symmetric, however, as we will show in this paper, it demonstrates unbalanced responsiveness to the boundary quality across object sizes.

Boundary-based segmentation measures evaluate segmentation quality by estimating contour alignment between predicted and ground truth masks. Unlike mask-based measures, these measures only evaluate the pixels that lie directly on the masks' contours or in their close proximity.

Trimap IoU [19, 5] is a boundary-based segmentation measure that calculates IoU in a narrow band of pixels within a pixel distance d from the contour of the ground truth mask (see Table 1). In contrast to Mask IoU, Trimap IoU reacts similarly to comparable pixel errors across object scales because it calculates IoU only for pixels around the contour. However, unlike Mask IoU, the measure is not symmetric and favors predictions whose masks are larger than the corresponding ground truth masks. Moreover, the

Name	Type	Definition	Symmetric	Preference	Insensitivity
Pixel accuracy	mask-based	$\frac{\lvert G \cap P \rvert}{\lvert G \rvert}$	✗	larger prediction	–
Mask IoU	mask-based	$\frac{\lvert G \cap P \rvert}{\lvert G \cup P \rvert}$	✓	–	boundary errors
Trimap IoU	boundary-based	$\frac{\lvert G_d \cap (G \cap P) \rvert}{\lvert G_d \cap (G \cup P) \rvert} = \frac{\lvert (G_d \cap G) \cap P \rvert}{\lvert (G_d \cap G) \cup (G_d \cap P) \rvert}$	✗	larger prediction	errors far from ground truth boundary
F-measure	boundary-based	$\frac{2 \cdot \tilde{p} \cdot \tilde{r}}{\tilde{p} + \tilde{r}}, \ \tilde{p} = \frac{\lvert P_1 \cap G_d \rvert}{\lvert P_1 \rvert}, \ \tilde{r} = \frac{\lvert G_1 \cap P_d \rvert}{\lvert G_1 \rvert}$	✓	–	errors on small objects
Boundary IoU	boundary-based	$\frac{\lvert (G_d \cap G) \cap (P_d \cap P) \rvert}{\lvert (G_d \cap G) \cup (P_d \cap P) \rvert}$	✓	–	errors far from predicted and ground truth boundaries

Table 1: Existing segmentation measures and the new Boundary IoU defined with the unified notation from Table 2. For each measure we detail several properties. **Symmetric**: whether the swap of ground truth and prediction masks changes measure's value. **Preference**: whether the measure biases better scores to a certain type of prediction. **Insensitivity**: types of errors the measure is less sensitive to.

Notation	Definition
G	ground truth binary mask
P	prediction binary mask
G_1, P_1	set of pixels on the contour line of the binary mask
G_d, P_d	set of pixels in the boundary region of the binary mask
d	pixel width of the boundary region

Table 2: Notation used in this paper. We define a contour as the 1d line comprised of the set of mask pixels that touches the background. The boundary is a 2D region consisting of pixels within pixel distance d from the contour pixels. A boundary region can be constructed by dilating the contour line by d pixels.

measure ignores prediction errors that appear outside the band around the ground truth contour.

F-measure was initially proposed for edge detection [27], but it is also used to evaluate segmentation quality [8, 28]. F-measure $= 2 \cdot p \cdot r / (p + r)$, where p and r denote precision and recall. In the original definition, p and r are calculated by matching prediction and ground truth contour pixels within the pixel distance threshold d via bipartite matching. However, the matching process is computationally expensive for high resolution masks and large datasets and, therefore, [8, 28] proposed an approximation procedure to compute the precision and recall by allowing duplicate matches, which we denote by $\tilde{p}$ and $\tilde{r}$. In this case, $\tilde{p}$ computes the ratio of pixels in the prediction contour that lie within a distance d from the ground truth contours, whereas $\tilde{r}$ computes a similar ratio for the pixels of the ground truth contour, see Table 1. In the rest of the paper we use the approximate formulation of F-measure. The measure is symmetric and tolerates small contour misalignments that can be attributed to ambiguity, however, it ignores significant errors when the object size is comparable to d. For example, this occurs with reasonable choices of d and small objects commonly found in datasets (*e.g.*, COCO [24]).

Trimap and F-measure are often used to evaluate boundary quality for semantic segmentation tasks in an ad-hoc fashion. For example, Trimap IoU is used as an extra evaluation to show boundary quality improvement [5, 25], but it is not reported by most segmentation methods. In the next section we will study both measures in detail and analyze their behavior across different error types and object sizes.

3. Sensitivity Analysis

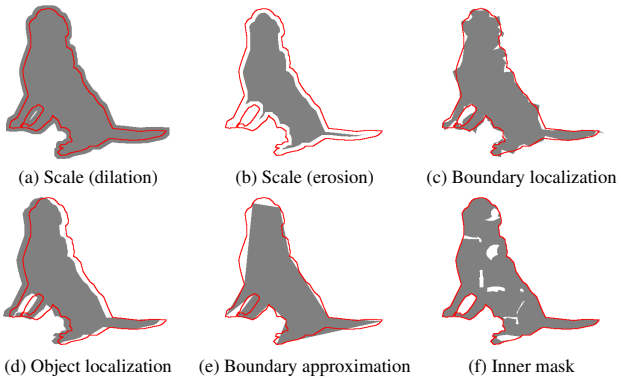
In §4 and §5 we will compare several mask consistency measures by observing how a measure's value changes in response to errors of different magnitudes. We will observe and interpret these curves to draw conclusions about the behavior of these measures, a methodology that we refer to as *sensitivity analysis*.

To enable a systematic comparison, we *simulate* a set of common segmentation errors across different mask sizes by generating pseudo-predictions from ground truth annotations. This approach allows us explicitly control the type and severity of the errors used in the analysis. Moreover, the use of pseudo-predictions avoids any bias toward specific segmentation models which makes the analysis more robust and general. A potential limitation of this approach is that simulated errors may not fully represent errors created by real models. We aim to counteract this limitation by using a diverse set of error types. Figure 2 depicts an example of each error type we consider:

• **Scale error.** Dilation/erosion are applied to the ground truth masks. The error severity is controlled by the kernel radius of the morphological operations.

• **Boundary localization error.** Random Gaussian noise is added to the coordinate of each vertex in polygons that represent ground truth masks. The error severity is controlled by the standard deviation (std) of the Gaussian noise.

• **Object localization error.** Ground truth masks are shifted with random offsets. The error severity is controlled by the pixel length of the shift.



(a) Scale (dilation) (b) Scale (erosion) (c) Boundary localization

(d) Object localization (e) Boundary approximation (f) Inner mask

Figure 2: **Examples of error types** generated from a single ground truth mask. The red contour represents the ground truth contour. *Scale errors*: (a) dilation and (b) erosion of the mask. *Boundary localization error*: (c) adding random Gaussian noise to each vertex in polygons. *Object localization error*: (d) shifting masks. *Boundary approximation error*: (e) simplifying polygons. *Inner mask error*: (f) adding holes to masks.

• **Boundary approximation error.** The `simplify` function from Shapely [12] removes vertices from the polygons that represent ground truth masks while keeping the simplified polygons as close to the original ones as possible. The error severity is controlled by the error tolerance parameter of the `simplify` function.

• **Inner mask error.** Holes of random shape are added to ground truth masks. The error severity is controlled by the number of holes added. While this error type is not common for modern segmentation approaches, we include it to assess the effect of interior mask errors.

Implementation details. For the analysis, we randomly sample instance masks from the LVIS v0.5 [13] validation set. The dataset is selected due to its high-quality annotations. Using these masks, for each segmentation error type we create multiple sets of pseudo-predictions by varying the severity of the error. To analyze a segmentation measure, we report its mean and standard deviation across a set of pseudo-predictions that represent a given error type of a fixed severity. We will also compare segmentation measures across different object sizes by generating a separate set of pseudo-predictions using ground truth objects within a specific mask area range. For all boundary-based measures that use pixel distance parameter, d, we set it to 2% of the image diagonal for fair comparison.

4. Analysis of Existing Segmentation Measures

First, we analyze the standard Mask IoU segmentation consistency measure from both theoretical and empirical perspectives. Then, we study two existing alternatives – Trimap IoU and F-measure boundary-based measures.

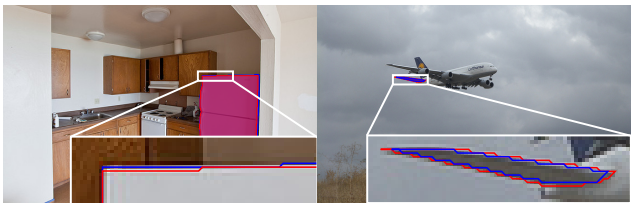


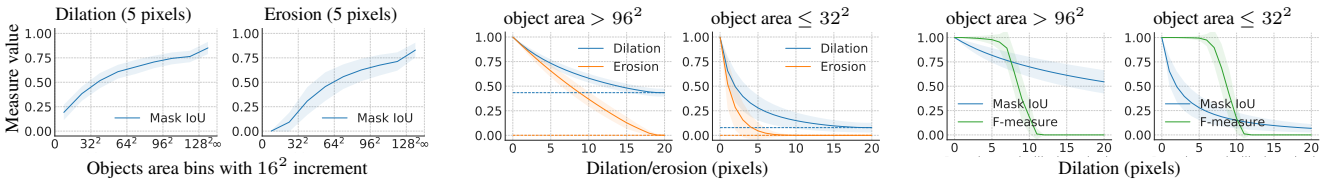
Figure 3: The two objects from LVIS [13] dataset annotated independently twice. While the fridge (left) has more than 100 times larger area than the wing (right), on the crop of the same resolution the discrepancy between annotations is visually similar. This simple example supports the observation that the boundary quality in the ground truth is independent of the object size. Mask IoU counts all pixels equally and gives a higher score to fridge, 0.97, vs wing, 0.81, due to the larger internal region with consistent labeling. The bias toward larger objects render Mask IoU inadequate for boundary quality evaluation across object sizes. In contrast, the new Boundary IoU yields much closer scores (0.87 *vs*. 0.81).

4.1. Mask IoU

Theoretical analysis. Mask IoU is scale-invariant w.r.t. object area. For a fixed Mask IoU value, a larger object will have more incorrect pixels and the change in incorrect pixel count grows in proportional to the change in object area (as Mask IoU is a ratio of areas). However, when scaling up a typical object, the number of interior pixels grows quadratically, whereas the number of contour pixels only grows linearly. These different asymptotic growth rates cause Mask IoU to tolerate a larger number of misclassified pixels *per each unit of contour length* for a larger object.

Empirical analysis. This property corresponds to an assumption that boundary localization error in ground truth annotations (*i.e.*, intrinsic annotation ambiguity) also grows with the object size. However, a classic study on multi-region segmentation [26] shows that the pixel distance between two contours of the same object labeled by different annotators seldomly exceeds 1% of the image diagonal, irrespective of the object size. We confirm this observation by exploring double annotations that are provided for a subset of images in LVIS [13]. In Figure 3 we present a random pair of objects with significant size difference. While one of the objects is 100 times larger, the boundary discrepancy within the cropped part, which has the same resolution, is similar between the two objects. Observed results suggest that boundary ambiguity is fixed and independent of objects area. This is likely a consequence of the annotation tool, which includes the ability to zoom while drawing contours.

Using simulated scale errors (described in §3) we confirm Mask IoU's bias in favor of large objects. The dilation/erosion of the ground truth mask by a fixed number of pixels significantly decreases Mask IoU for small objects while Mask IoU grows as object area increases (see Fig-



(a) Mask IoU is biased toward large objects. The pseudo-predictions for larger objects receive higher score under a fixed error severity.

(b) Trimap IoU is not symmetric. It favors predictions larger than ground truth masks (*e.g.* dilated pseudo-predictions).

(c) F-measure completely ignores small contour misalignments and rapidly drops to zero within a short range of severities.

Figure 4: Sensitivity analysis across object scales for Mask IoU (a), Trimap IoU (b), and F-measure (c) with scaling error type.

ure 4a). Note that Mask IoU's insensitivity to boundary errors on large objects cannot be addressed by simply increasing the lowest Mask IoU threshold in evaluation metrics like AP or PQ. Such a change does not remedy the bias and will lead to relative over-penalization for smaller objects.

4.2. Boundary-Based Measures

Next, we will analyze the boundary-based measures Trimap IoU and F-measure. These measures focus on pixels within a distance d from object contours. The parameter d is usually fixed on the dataset [5] or image level [28] which results in these measures treating boundary localization errors independently of the size of the object. By matching the natural characteristic of the ground truth segmentation data, these boundary-based measures are better suited to evaluate improvements in boundary quality across object sizes.

Trimap IoU computes IoU for a region around the ground truth boundary only (*i.e.*, the region is independent of the prediction), and therefore is not symmetric: swapping the prediction and ground truth masks will give a different score. In Figure 4b we show that this asymmetry favors predictions that are larger than ground truth masks. For larger (dilated) pseudo-predictions Trimap IoU does not drop below some positive value irrespective of the error severity, whereas for smaller (eroded) pseudo-predictions it drops to 0. Moreover, the measure ignores any errors outside of the ground truth boundary region, penalizing inner mask errors less than Mask IoU (see the supplement for details).

F-Measure matches the pixels of the predicted and ground truth contours if they are within the pixels distance threshold d. Hence, it ignores small contour misalignments that can be attributed to ambiguity. While robustness to ambiguity is good in principle, in Figure 4c we observe that F-measure can be nearly discontinuous, rapidly stepping from 1 to 0 when the error severity changes by a small amount. Sharp response curves can lead to task metrics with high variance. In comparison, Mask IoU is more continuous. Further, d may be large relative to small objects, causing F-measure to award significant errors a perfect score.

Discussion. Given the limitations presented above, we conclude that neither Trimap IoU nor F-measure can replace Mask IoU as the main segmentation consistency measure for a broad range of evaluation metrics. At the same time, Mask IoU is biased towards large objects in a way that discourages improvements to boundary segmentations. Next, we propose *Boundary IoU* as a new measure to evaluate segmentation boundary-quality that does not have any of the previously mentioning limitations.

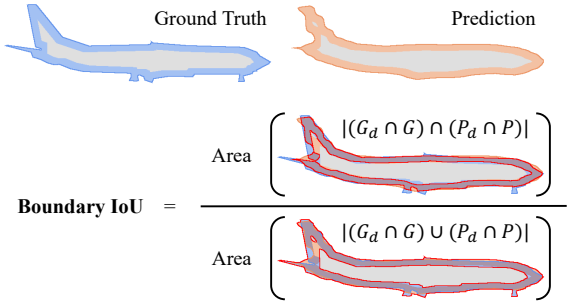


Figure 5: **Boundary IoU computation illustration**. (top) Ground truth and prediction masks. For both, we highlight *mask* pixels that are within distance d from the contours. (bottom) Boundary IoU segmentation consistency measures computes the intersection-over-union between the highlighted regions. In this example, Boundary IoU is 0.73, whereas Mask IoU is 0.91.

5. Boundary IoU

In this section we introduce a new segmentation measure and compare it with existing consistency measures using simulated errors. The new measure should have a weaker bias toward large objects than Mask IoU. Furthermore, we aim for a measure that neither over-penalizes nor ignores errors in small objects similarly to Mask IoU.

Guided by these principals, we propose the *Boundary IoU* segmentation consistency measure. The new measure is simultaneously simple and satisfies the principals charted above. Given two masks G and P, Boundary IoU first computes the set of the original masks' pixels that are within distance d from each contour, and then computes the intersection-over-union of these two sets (see Figure 5). Using the notation from Table 2:

$$\text{Boundary IoU}(G, P) = \frac{|(G_d \cap G) \cap (P_d \cap P)|}{|(G_d \cap G) \cup (P_d \cap P)|}, \quad (1)$$

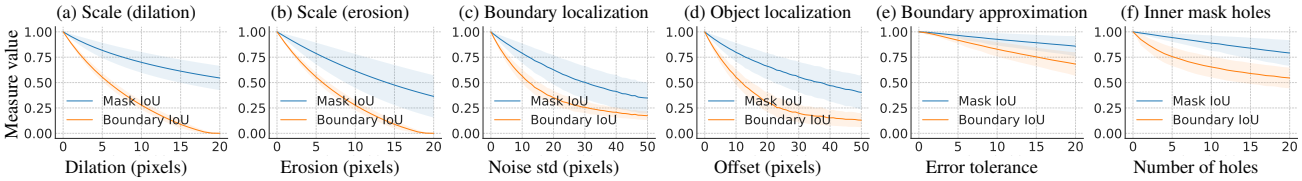


Figure 6: Boundary IoU sensitivity curves *across error severities*. We use pseudo-predictions for objects with area $> 96^2$. For each error type, Boundary IoU makes better use of the 0-1 value range demonstrating an improved ability to differentiate between error severity.

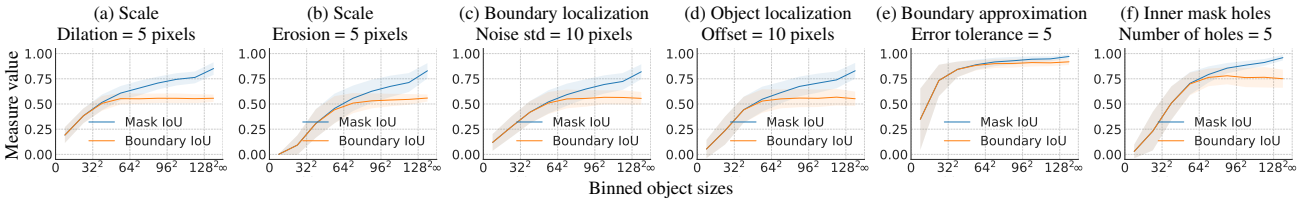


Figure 7: Boundary IoU sensitivity curves *across object sizes*. We use pseudo-predictions for objects of different sizes with *fixed* error severities. Objects are binned by their area with 16^2 increment. For larger objects, Boundary IoU remains flat given the fixed error severity, while Mask IoU demonstrates a clear preferential bias for large objects. For small objects, both metrics have similar curves indicating that neither over-penalizes small objects w.r.t. to the other.

where boundary regions G_d and P_d are the sets of all pixels within d pixels distance from the ground truth and prediction contours respectively. Note, that the new measure evaluates only *mask* pixels that are within pixel distance d from the contours. A simpler version with IoU calculated directly for boundary regions G_d and P_d loses information about sharp contour corners that are smoothed by considering all pixels within distance d from the contours.

The distance to the contour parameter d in Boundary IoU controls the sensitivity of the measure. With d large enough to include all pixels inside the prediction and ground truth masks, Boundary IoU becomes equivalent to Mask IoU. With a smaller parameter d Boundary IoU ignores interior mask pixels, which makes it more sensitive to the boundary quality than Mask IoU for larger objects. For smaller objects, Boundary IoU is very close or even equivalent to Mask IoU depending on the parameter d which prevents it from over-penalizing errors in smaller objects where the number of inner pixels is comparable with the number of pixels close to the contours.

In Figure 6 and Figure 7 we present the results of our analysis for Boundary IoU. The analysis shows that Boundary IoU is less biased than Mask IoU towards large object sizes across all considered error types (Figure 6). Varying object sizes while keeping error severities constant (Figure 7), Boundary IoU behaves identically to Mask IoU for smaller objects avoiding over-penalization for any error type. For larger objects Boundary IoU reduces the bias that Mask IoU exhibits and its value grows more slowly with the object area given fixed error severities.

Comparison with Trimap IoU. We note that the new measure appears quite similar to Trimap IoU (see Table 1). However, unlike Trimap IoU, Boundary IoU considers pixels close to the contours of both prediction and ground truth together. This simple change remedies two main limitations of Trimap IoU. The new measure is symmetric and penalizes the errors that appear away from the ground truth boundary (see Figure 6 (a), (b), and (f)).

Comparison with F-measure. F-measure uses hard matching between the contours of the predicted and ground truth masks. If the distance between contour pixels is within d pixels, then both precision and recall are perfect, but once the distance goes beyond d the matching does not happen at all. In contrast, Boundary IoU evaluates consistency in a soft manner. Intersection over union is 1.0 if two contours are perfectly aligned and as the contours diverge Boundary IoU gradually decreases. In the supplement, we also compare Boundary IoU with a soft generalization of F-measure that averages multiple scores across different parameters d. Our analysis shows that it under-penalizes errors in small objects in comparison with Mask IoU and Boundary IoU.

The pixel distance parameter *d*. If d is large enough Boundary IoU is equivalent to Mask IoU. On the other hand, if d is too small, Boundary IoU severely penalizes even the smallest misalignment ignoring possible ambiguity of the contours. To select d that does not over-penalize possible ambiguity of the contours, we use Boundary IoU to measure the consistency of two expert annotators who independently delineated the same objects. The creators of LVIS [13] have collected such expert annotations for images in COCO [24] and ADE20k [35] datasets. Both datasets have images of similar resolution and we find that median Boundary IoU between the annotations of the two experts exceeds 0.9 for both datasets when d equals 2% of an image diagonal (15 pixels distance on average). For Cityscapes [7] that has

higher resolution images and excellent annotation quality we suggest to use the same distance in pixels which results in d set to 0.5% of an image diagonal for the dataset.

For other datasets, we suggest two considerations for selecting the pixel distance d: (1) the annotation consistency sets the lower bound on d, and (2) d should be selected based on the performance of current methods and decreased as performance improves.

Limitations of Boundary IoU. The new measure does not evaluate mask pixels that are further than d pixels away from the corresponding ground truth or prediction boundary. Therefore, it can award a perfect score to non-identical masks. For example, Boundary IoU is perfect for a disc-shaped mask and a ring-shaped 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 (we show this example in the supplement). For these two masks, all non-matching pixels of the disc-shaped mask are further than d pixels away from its boundary. To penalize such cases, we suggest a simple combination of Mask IoU and Boundary IoU by taking their minimum. In our experiments with both real and simulated predictions, we found that Boundary IoU is smaller or equal to Mask IoU in the absolute majority of cases (99.9%) and the inequality is violated when a prediction with accurate boundaries misses interior part of an object (similar to the toy example above).

6. Applications

The most common evaluation metrics for instance and panoptic segmentation tasks are Average Precision (AP or Mask AP) [24] and Panoptic Quality (PQ or Mask PQ) [17] respectively. Both metrics use Mask IoU and inherit its bias toward large objects and, subsequently, the insensitivity to the boundary quality observed before in [18, 31].

We update the evaluation metrics for these tasks by replacing Mask IoU with **min(Mask IoU, Boundary IoU)** as suggested in the previous section. We name the new evaluation metrics Boundary AP and Boundary PQ. The change is simple to implement and we demonstrate that the new metrics are more sensitive to the boundary quality while able to track other types of improvements in predictions as well.

We hope that adoption of the new evaluation metrics will allow rapid progress of boundary quality in segmentation methods. We present our results for instance segmentation in the main text and refer to the supplement for our analysis of Boundary PQ.

Boundary AP for instance segmentation. The goal of the instance segmentation task is to delineate each object with a pixel-level mask. An evaluation metric for the task is simultaneously assessing multiple aspects such as categorization, localization, and segmentation quality. To compare different evaluation metrics, we conduct our experiments with both synthetic predictions and real instance segmentation models. Synthetic predictions allow us to assess the segmentation quality aspect in isolation, whereas real predictions provide insights into the ability of Boundary AP to track all aspects of the instance segmentation task.

Evaluation metric	AP	AP_S	AP_M	AP_L
Mask AP	96.5	98.9	95.7	95.0
Boundary AP	85.9	98.9	93.0	73.0

Table 3: Boundary AP and Mask AP on COCO `val` set for synthetic 28×28 predictions generated from the ground truth. Unlike Mask AP_L, Boundary AP_L successfully captures the lack of fidelity in the synthetic prediction for large objects (area $> 96^2$).

We compare Mask AP and Boundary AP on the COCO instance segmentation dataset [24] in the main text. In addition, our findings are supported by similar experiments on Cityscapes [7] and LVIS [13] presented in the supplement. Detailed description of all datasets can be found in the supplement, along with Boundary AP evaluation for various recent and classic models on all three datasets. These results can be used as a reference to simplify the comparison for future methods.

6.1. Evaluation on Synthetic Predictions

Using synthetic predictions we evaluate the segmentation quality aspect of instance segmentation in isolation without a bias that any particular model can have. We simulate predictions by capping the effective resolution of each mask. First, we downscale cropped ground truth masks to a 28×28 resolution[2] mask with continuous values, we then upscale it back using bilinear interpolation, and finally binarize it. Such synthetic masks are close to the ground truth masks for smaller objects, however the discrepancy grows with object size. In Table 3 we report overall AP and AP_S, AP_M, and AP_L for object size splits defined in COCO [24]. Mask AP follows the behavior of Mask IoU, showing little sensitivity to the error growth between AP_S and AP_L. In contrast, Boundary AP successfully captures the difference with significantly lower score for larger objects. In the supplement, we provide an example of the synthetic predictions and more results using different effective resolutions.

6.2. Evaluation on Real Predictions

We use outputs of existing segmentation models to further study Boundary AP. Unless specified, to isolate the segmentation quality from categorization and localization errors for purposes of analysis, we supply ground truth boxes to these methods and assign a random confidence score to each box. We use Detectron2 [32] with a ResNet-50 [15] backbone unless otherwise specified.

[2]This is a popular prediction resolution used in practice [14].

Method	Evaluation metric	AP	AP_S	AP_M	AP_L
Mask R-CNN	Mask AP	52.5	44.9	55.0	66.0
Mask R-CNN	Boundary AP	36.1	44.8	46.5	25.9

(a) AP of Mask R-CNN **with ground truth boxes**. Mask R-CNN makes blobby predictions with large defects around boundaries for large objects (see Figure 1). Boundary AP successfully captures these errors with much lower AP_L than Mask AP.

Method	Backbone	Ground truth boxes AP^{mask}	Ground truth boxes $AP^{boundary}$	Predicted boxes AP^{mask}	Predicted boxes $AP^{boundary}$
Mask R-CNN	R50	53.6	37.7	37.2	23.1
Mask R-CNN	R101	53.8 **(+0.2)**	38.2 **(+0.5)**	38.6 (+1.4)	24.5 (+1.4)
Mask R-CNN	X101	53.4 (-0.2)	38.1 (+0.4)	39.5 **(+2.3)**	25.4 **(+2.3)**

(b) Larger backbones do not improve segmentation quality significantly (*i.e.*, Boundary AP stays roughly the same when ground truth boxes are used). With real box predictions, Boundary AP tracks the categorization and localization improvements similarly to Mask AP.

Method	AP^{mask}	$AP^{boundary}$	$AP^{boundary}_S$	$AP^{boundary}_M$	$AP^{boundary}_L$
Mask R-CNN	52.5	36.1	44.8	46.5	25.9
PointRend-28	57.0 (+4.5)	41.6 (+5.5)	48.2 (+3.4)	52.3 (+5.8)	33.3 (+7.4)
PointRend-224	57.2 **(+4.7)**	42.1 **(+6.0)**	48.3 **(+3.5)**	52.5 **(+6.0)**	34.4 **(+8.5)**

(c) PointRend (with either 28×28 or 224×224 output resolution) is designed to give higher-quality output masks than Mask R-CNN (which has 28×28 output resolution). Boundary AP captures these improvements well, especially for the higher-resolution output variant of PointRend and for AP_L.

Method	AP^{mask}	$AP^{boundary}$	$AP^{boundary}_S$	$AP^{boundary}_M$	$AP^{boundary}_L$
Mask R-CNN	52.5	36.1	44.8	46.5	25.9
PointRend	57.2 (+4.7)	42.1 (+6.0)	48.3 (+3.5)	52.5 (+6.0)	34.4 **(+8.5)**
BMask R-CNN	57.4 **(+4.9)**	42.3 **(+6.2)**	48.7 **(+4.1)**	52.7 **(+6.2)**	33.9 (+8.0)

(d) Boundary-preserving Mask R-CNN (BMask R-CNN) which uses 28×28 output *vs*. PointRend which uses 224×224 output. The Boundary AP metric reveals that BMask R-CNN outperforms PointRend for small objects but trails it for large objects where the high output resolution of PointRend improves boundary quality.

Table 4: Comparison of Mask AP and Boundary AP on COCO `val` set for different instance segmentation models **fed with ground truth boxes unless specified otherwise**. Using ground truth boxes disentangles segmentation errors from localization and classification errors.

Mask AP *vs*. Boundary AP. Table 4a shows both Mask AP and Boundary AP for the standard Mask R-CNN model [14]. Mask R-CNN is well-known to predict blobby masks with significant visual defects for larger objects (see Figure 1). Nevertheless, Mask AP_L is larger than Mask AP_S. In contrast, we observe that Boundary AP_L is smaller than Boundary AP_S for Mask R-CNN suggesting that the new measure is more sensitive to the boundary quality of the large objects. Note that in this experiment the use of ground truth boxes removes any categorization and localization errors that are usually larger for small objects.

Segmentation *vs*. categorization and localization. A general evaluation metric for instance segmentation should track the improvements in all aspects of the task including segmentation, categorization, and localizations. In Table 4b we first evaluate Mask R-CNN with several backbones (ResNet-50, ResNet-101, and ResNeXt-101-32×8d [33]), again supplying ground truth boxes. Note that both Mask AP and Boundary AP do not change significantly with different backbones, suggesting that more powerful backbones do not directly influence the segmentation quality. Next, we evaluate Boundary AP requiring each model to predict its own boxes as is standard. We observe that Boundary AP is able to track improvements from better localization and categorization similarly to Mask AP.

Mask quality improvements. We explore Boundary AP's ability to capture the improvements in segmentation quality by the methods designed for this purpose in Tables 4c and 4d. To compare the segmentation quality aspect across models we again supply ground truth boxes to each model.

PointRend [18] was developed to improve pixel-level prediction quality of models like Mask R-CNN and can produce predictions of varying resolution. PointRend significantly improves mask quality, while this can be measured via mask AP, it is more pronounced in Boundary AP, especially for large objects and for a higher resolution PointRend variant. See Table 4c for details.

Boundary-preserving Mask R-CNN [6] (BMask R-CNN) improves boundary quality by adding a direct boundary supervision loss and increasing the resolution of feature maps used in its mask head. In Table 4d Boundary AP shows that BMask R-CNN with its 28×28 output resolution outperforms PointRend for small objects, whereas for larger objects the 224×224 resolution output of PointRend is preferable, which matches a subjective visual quality assessment (see an example in Figure 1). We hope that the improved sensitivity of the new Boundary AP metric will lead to a rapid progress in the methods that improve boundary quality for instance segmentation.

7. Conclusion

Unlike the standard Mask IoU, Boundary IoU segmentation quality measure provides a clear, quantitative gradient that rewards improvements to boundary segmentation quality. We hope that the new measure will challenge our community to develop new methods with high-fidelity mask predictions. In addition, Boundary IoU allows a more granular analysis of segmentation-related errors for the complex multifaceted tasks like instance and panoptic segmentation. Incorporation of the measure in the performance analysis tools like TIDE [2] can provide better insights into specific error types of instance segmentation models.

References

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

[2] Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. TIDE: A general toolbox for identifying object detection errors. In *ECCV*, 2020.

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

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

[5] 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.

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

[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] Gabriela Csurka, Diane Larlus, Florent Perronnin, and France Meylan. What is a good evaluation measure for semantic segmentation?. In *BMVC*, 2013.

[9] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016.

[10] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes challenge: A retrospective. *IJCV*, 2015.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.

[12] Sean Gillies et al. Shapely: manipulation and analysis of geometric objects, 2007.

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

[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] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring R-CNN. In *CVPR*, 2019.

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

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

[19] Pushmeet Kohli, Philip HS Torr, et al. Robust higher order potentials for enforcing label consistency. *IJCV*, 2009.

[20] Zeming Li, Yuchen Ma, Yukang Chen, Xiangyu Zhang, and Jian Sun. Joint coco and mapillary workshop at iccv 2019: Coco instance segmentation challenge track. *arXiv:2010.02475*, 2020.

[21] Zeming Li, Yueqing Zhuang, Xiangyu Zhang, Gang Yu, and Jian Sun. COCO instance segmentation challenges 2018: winner. http://presentations.cocodataset.org/ECCV18/COCO18-Detect-Megvii.pdf, 2018.

[22] Justin Liang, Namdar Homayounfar, Wei-Chiu Ma, Yuwen Xiong, Rui Hu, and Raquel Urtasun. Polytransform: Deep polygon transformer for instance segmentation. In *CVPR*, 2020.

[23] Mark Liberman. Reproducible research and the common task method. https://www.simonsfoundation.org/event/reproducible-research-and-the-common-task-method, 2015.

[24] 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.

[25] Dmitrii Marin, Zijian He, Peter Vajda, Priyam Chatterjee, Sam Tsai, Fei Yang, and Yuri Boykov. Efficient segmentation: Learning downsampling near semantic boundaries. In *ICCV*, 2019.

[26] David Royal Martin. *An empirical approach to grouping and segmentation*. University of California Berkeley, 2003.

[27] David R Martin, Charless C Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *PAMI*, 2004.

[28] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016.

[29] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. In *BMVC*, 2020.

[30] Akash Sharma, Wei Dong, and Michael Kaess. Compositional scalable object SLAM. *arXiv:2011.02658*, 2020.

[31] Zian Wang, David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Object instance annotation with deep extreme level set evolution. In *CVPR*, 2019.

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

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

[34] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *ECCV*, 2020.

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

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