SortedAP: Rethinking evaluation metrics for instance segmentation

Long Chen[1] Yuli Wu[1] Johannes Stegmaier[1]
Dorit Merhof[2]
[1] Institute of Imaging & Computer Vision, RWTH Aachen University, Germany
[2] Faculty of Informatics and Data Science, University of Regensburg, Germany
{Long.Chen, Yuli.Wu, Stegmaier.Johannes}@lfb.rwth-aachen.de,
Dorit.Merhof@informatik.uni-regensburg.de

Abstract

Designing metrics for evaluating instance segmentation revolves around comprehensively considering object detection and segmentation accuracy. However, other important properties, such as sensitivity, continuity, and equality, are overlooked in the current study. In this paper, we reveal that most existing metrics have a limited resolution of segmentation quality. They are only conditionally sensitive to the change of masks or false predictions. For certain metrics, the score can change drastically in a narrow range which could provide a misleading indication of the quality gap between results. Therefore, we propose a new metric called sortedAP, which strictly decreases with both object- and pixel-level imperfections and has an uninterrupted penalization scale over the entire domain. We provide the evaluation toolkit and experiment code at https://www.github.com/looooongChen/sortedAP.

1. Introduction

Recently, considerable work has been conducted in instance segmentation due to its wide scope of application [9, 19, 4, 3], such as autonomous driving [5], medical diagnosis [13] and agricultural phenotyping [18, 2]. In the field of bioimage computing, segmenting instances of animals [16], cells [6], and subcellular structures [1, 8] is also common and infrastructural processing for further analysis and study. Instance segmentation not only localizes the object of interest but also delineates the exact boundary, which can be seen as performing object detection and semantic segmentation concurrently.

Correspondingly, a qualified evaluation metric should consider three fundamental types of imperfections: missed ground truth objects (false negative), falsely predicted objects (false positive), and segmentation inaccuracy. Existing metrics all incorporate the three error types above, but are not discussed with respect to properties, including sensitivity, continuity, and equality.

Sensitivity. An ideal metric should be sensitive to all occurrences of imperfections of all types. Any additional errors are supposed to lead monotonically to a worse score, not ignored or obscured by the occurrence of other errors. A metric that monotonically decreases with any errors will enable a more accurate comparison.

Continuity. The penalization scale of a metric should be relatively consistent locally across the score domain. Intuitively, gradually and evenly changing segmentations should correspond to a smoothly changing metric score as well. Abrupt changes are not desired.

Equality. Without any assumed importance of different objects, all objects should have an equal influence on the metric score. A common case of inequality is that the score is biased towards larger objects. Although larger objects may be prioritized in some applications, as a general metric, the metric should treat all objects equally. Analysis with respect to object size can be easily performed by evaluating different size groups using a metric of equal property.

Although all metrics discussed in this paper implement a penalization of false positive, false negative, and segmentation inaccuracy, the majority of metrics, even very widely used ones, such as the mean Average Precision (mAP) [1], are only conditionally sensitive to errors. This violates the sensitivity property, as some differences in segmentation results are not reflected in the score. For match-based approaches, such as Average Precision (AP) [1] and Panoptic Quality (PQ) [11], the score will change abruptly at the match threshold. There is actually a paradox in choosing thresholds, which is discussed in Section 3.

To address the gap, we propose a new metric called the sorted Average Precision (sortedAP). Unlike mAP [1], which queries the AP score at a sequence of fixed intersection over union (IoU) thresholds, sortedAP detects every exact IoU value at which the AP score drops. This is achieved through our proposed *Unique Matching* approach and sort-

ing all possible matches according to the IoU values (Section 4). The Unique Matching method explicitly preserves the one-to-one relationship between two sets of instances. This also allows the use of IoU thresholds smaller than 0.5, or under object overlap, in all match-based metrics.

2. Related work: A review

This section provides an overview of proposed evaluation metrics in the literature. We use the notion $\mathcal{G} = \{g_1, g_2, \ldots, g_M\}$ and $\mathcal{P} = \{p_1, p_2, \ldots, p_N\}$ to represent the set of ground truth and predicted objects in the following context. The capitalized symbols $\mathcal{G}$ and $\mathcal{P}$ can represent a set, or the number of elements in the set, for notation simplicity.

2.1. Overlap-based metrics

The Dice coefficient (Dice) and the Intersection over Union (IoU) are the most commonly used metrics to measure the similarity between two binary masks. The IoU, also known as Jaccard Index (JI), is defined as the ratio of the intersection area to the union area between two masks:

$$IoU(p, g) = \frac{|p \cap g|}{|p \cup g|}. \tag{1}$$

Instead of the union, Dice use accumulated area:

$$Dice(p, g) = \frac{2 \cdot |p \cap g|}{|p| + |g|}. \tag{2}$$

Although they have slightly different definitions, both metrics utilize the same fact that the intersection area is maximized when two masks are identical. Furthermore, the two metrics are directly related in values:

$$Dice(p, g) = \frac{2 \cdot IoU(p, g)}{1 + IoU(p, g)}. \tag{3}$$

Aggregated Jaccard Index (AJI). The AJI [13] extends the Jaccard Index to instance segmentation by accumulating the object-level intersection and union area, which is computed between each ground truth object and the prediction yielding the maximum IoU. The area of predicted objects without any matched ground truth objects is also aggregated to the union area as the penalization to false positives.

Symmetric Best Dice (SBD). SBD [18] is based on an asymmetric score Best Dice (BD). For each object in one set, BD finds the maximal Dice with any object in the other set (the reference set) for averaging.

$$BD(\mathcal{P}, \mathcal{G}) = \frac{1}{N} \sum_{i=1}^{N} \max_{j=1:M} Dice(p_i, g_j), \tag{4}$$

The BD does not fully penalize all errors, since unmatched objects in the reference set are excluded and have no impact on the score. Therefore, the SBD computes BD using both sets under comparison as the reference and takes the worse score as the final score:

$$SBD(\mathcal{P}, \mathcal{G}) = min\{BD(\mathcal{P}, \mathcal{G}), BD(\mathcal{G}, \mathcal{P})\}. \tag{5}$$

2.2. Match-based metrics

Another category of metrics is based on object-level detection errors at one or multiple segmentation quality thresholds. A matching criterion t, typically an IoU value, is defined as a prerequisite. Each ground truth object searches for a successful match in the predicted objects, or vice versa. Based on the match results, all objects can be grouped into one of the three categories: true positives (TP_t), false positives (FP_t), and false negatives (FN_t).

Fundamentally, the match between predicted objects and ground truth objects should satisfy a one-to-one relationship. This ensures that the number of true positives is equal to the number of ground truth objects that have a successful match. We will discuss how to explicitly maintain this relationship in Section 4.1.

Average precision (AP). The term AP can refer to different evaluation metrics in the literature. For ease of discussion, we refer to them as the P-R AP [7] and the point AP [1]. Despite being based on different perspectives, both metrics are defined in terms of precision and recall:

$$Pre_t = \frac{TP_t}{TP_t + FP_t}, \; Rec_t = \frac{TP_t}{TP_t + FN_t}. \tag{6}$$

The P-R AP was first proposed for the evaluation of object detection tasks [7, 14]. As a summary of the Precision-Recall curve (P-R curve), it evaluates a model from a more comprehensive view by considering the precision performance over the entire recall domain. Although very widely used, the P-R AP suffers from certain deficiencies, as pointed out by recent works. Firstly, the definition requires a confidence score for each prediction, while not all approaches naturally score the outputs. For example, most bottom-up approaches do not directly deliver object-level confidence scores as most detection-based pipelines do. In terms of discrimination capability, P-R AP does not really distinguish between different shapes of P-R curves [17]. The neglect of low-confidence duplicates (hedged prediction) is another important deficiency of P-R AP [10].

In comparison, the point AP is oriented towards the end result and corresponds to a point on the P-R curve that achieves a certain precision-recall trade-off. In this case, all predictions are treated equally regardless of scoring. The point AP is formulated as follows:

$$AP_t = \frac{TP_t}{TP_t + FP_t + FN_t}. \tag{7}$$

The point AP relates to the P-R curve according to the following equation:

$$AP_t = \frac{1}{Pre_t + Rec_t + 1}. \tag{8}$$

While the P-R AP favors precision improvements at any recall level, the point AP only focuses on the single point of best precision-recall trade-off. From the user's perspective, higher precision in the extreme recall range is of limited practical significance. Therefore, point AP obligates the processing pipeline to screen predictions, including determining the optimal cutoff confidence. In the following context, we refer to the point AP when using the term AP.

Mean Average Precision (mAP). The AP score is based on the matching results under a certain IoU threshold t. Segmentation imperfections better than the matching criterion will not be further penalized. Similarly, objects worse than the threshold are viewed as equally bad.

To compensate for the neglect of segmentation imperfections, the mean Average Precision (mAP) [1] averages a series of AP scores over progressively higher IoU thresholds:

$$mAP = \frac{1}{N} \sum_{t \in T} \frac{TP_t}{TP_t + FP_t + FN_t}, \tag{9}$$

where $T = \{t_1, t_2, \ldots, t_N\}$. A typical choice for the threshold range is from 0.5 to 0.95, with a step size of 0.05.

It is worth mentioning that when referring to mAP, it generally means the averaging of multiple AP scores, rather than scores under different matching thresholds specifically. For example, the PASCAL dataset [7] computes P-R AP scores of different semantic classes for averaging. The COCO challenge [14] considers both the semantic categories and varying matching thresholds. In this work, we only discuss averaging across matching thresholds, as it is directly relevant to metric design.

Panoptic Quality (PQ). The PQ is defined as the multiplication of the Recognition Quality (RQ)

$$RQ = \frac{2 \cdot TP_{t=0.5}}{2 \cdot TP_{t=0.5} + FP_{t=0.5} + FN_{t=0.5}} \tag{10}$$

and the Segmentation Quality (SQ)

$$SQ = \frac{\sum_{(p,g) \in TP_{T=0.5}} IoU(p, g)}{|TP_{t=0.5}|}, \tag{11}$$

where (p, g) indicates a matched prediction and ground truth pair. The RQ measures the detection accuracy as the AP and they are related as

$$RQ = \frac{2 \cdot AP_{t=0.5}}{1 + AP_{t=0.5}}. \tag{12}$$

The SQ term is basically the mean IoU of all true positive pairs, explicitly modeling the segmentation quality of objects above the match threshold.

Metrics	AJI	SBD	PQ	mAP	sortedAP
Case-1	**.5125**	**.4925**	**.4229**	**.3778**	.4261
Case-2	.4587	.4325	.3771	**.3778**	.3839
Case-3	.6252	**.4925**	.4933	.4722	.5283
Case-4	.5159	.4975	**.4975**	.4000	.4288
Case-5	**.5125**	.3940	.3700	.3148	.3572

Table 1. Scores of different metrics for the examples shown in Figure 1. In each column, the pair of cases marked in bold demonstrate the deficiency of a metric.

3. An analysis of deficiencies

3.1. Sensitivity to errors

While existing metrics account for all three types of errors, few of them are sensitive to all occurrences of errors.

Exempted error. SBD takes the worse BD score between using the ground truth and the prediction as the reference. This only considers false positives or false negatives, except the segmentation inaccuracy, respectively. As illustrated in Figure 1a and Figure 1c, predictions with and without an additional false positive have the same SBD score. Although the false prediction decreases BD$(\mathcal{P}, \mathcal{G})$, the impact on SBD is exempted by the lower BD$(\mathcal{G}, \mathcal{P})$.

Resolution of segmentation difference. As stated previously, the mAP score reflects the segmentation quality by computing AP scores at varying IoU thresholds, with a certain step size. Despite having a good practical utility with an appropriate step size, mAP is only definitely sensitive to IoU changes greater than the step size. A smaller difference in IoU may or may not result in score changes, depending on whether the change crosses a predefined IoU threshold or not. From Figure 1a to Figure 1b, all IoUs decrease by 0.08. However, the mAP score remains unchanged in the case of step size 0.1 (Figure 2).

3.2. Match thresholds and score continuity

Match-based metrics use hard thresholds to determine true and false positives. As a result, objects can abruptly transition from true positives to false positives, even if they are only slightly different in IoU. PQ and mAP introduce a continuous or quasi-continuous measure of the segmentation, but only in the domain above the minimum IoU threshold. A discontinuous change always occurs at the lower IoU threshold. An example is shown in Figure 1, where increasing the IoU of only one prediction from 0.49 to 0.51 leads to a PQ change of 17.64%, from 0.4229 to 0.4975 (Table 1).

Threshold dilemma. A discontinuous score is not completely unacceptable. The IoU threshold can be set low enough so that two useful results (away from the low IoU range) will not be assigned drastically different scores. However, a single AP or PQ score reported with a low match threshold becomes less informative. PQ makes the

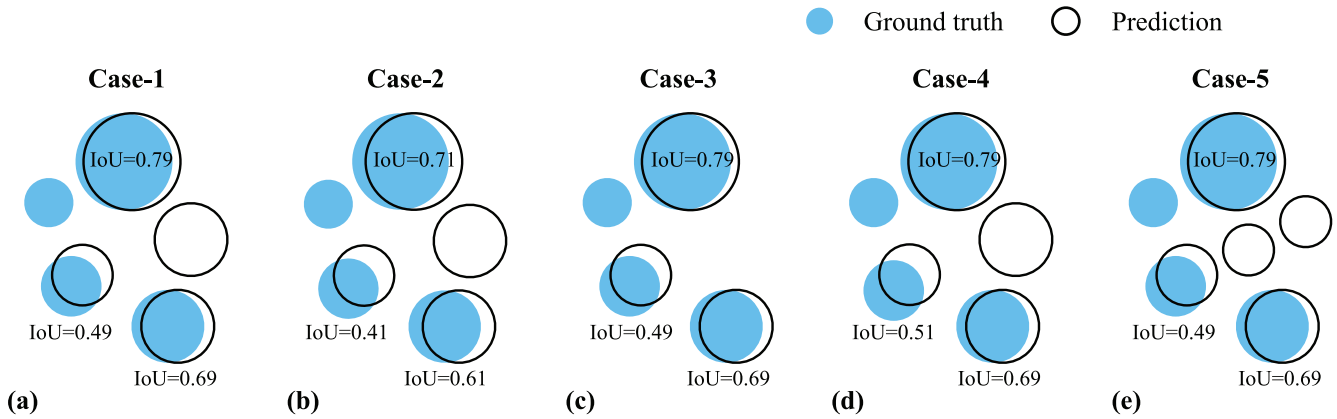

Figure 1. Examples to illustrate the deficiencies of evaluation metrics. (a) Case-1 is the base example. (b) All IoUs get worse in Case-2, but the mAP score remains unchanged. (c) Case-3 contains one less false positive, but SBD score is the same as Case-1. (d) In Case-4, only one object segmentation improves by 0.02 in IoU, but the PQ score increases by 17.64%. (e) Two false positives are present in Case-5, while only one exists in Case-1. AJI score penalizes them equally due to the smaller size of objects in Case-5.

compromise at the IoU of 0.5. The mAP only alleviates the amplitude of abrupt changes by dividing them into multiple levels (Figure 3c and Figure 3f).

3.3. Equality of object-level errors

Without specific assumptions, objects should be treated equally. A missed small object is supposed to place the same impact on the score as a larger object. Object segmentation accuracy should also be measured relative to their size, rather than the absolute area. Match-based approaches satisfy this property by constructing the metric using the object counts and object-level IoU. SBD takes the average of object Dice, therefore also area-independent. In contrast, AJI does not have a notion of objects. For instance, the scenario of having two false positives in Figure 1e yields the same AJI score as the scenario of having one larger false positive in Figure 1a. And accumulating absolute area will also bias the score towards the quality of larger objects.

4. Sorted Average Precision (sortedAP)

4.1. Unique Matching

For match-based metrics, each ground truth object can match at most one prediction, and vice versa. This rule ensures that the number of true positives is consistent with the number of ground truth objects that have a successful match. In the greedy match used by mAP and PQ, the one-to-one relationship is implicitly maintained by using matching IoUs larger than 0.5. This is because, under the non-overlapping assumption, no two objects can match with the same object while both having IoUs larger than 0.5 [11].

We propose using the Hungarian algorithm [12] to determine true positive matches. This involves the following steps: constructing the cost matrix, padding the cost

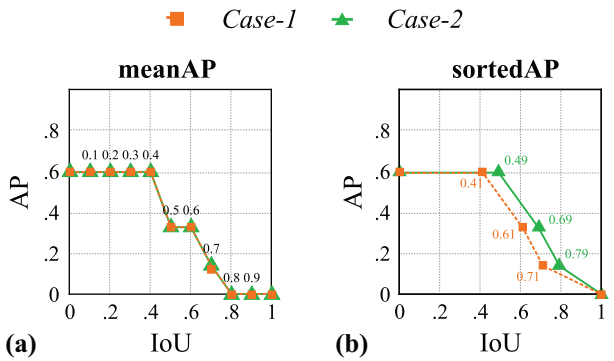

Figure 2. Computation of meanAP and sortedAP on the Case-1 and Case-2 in Figure 1. The mAP estimates the AP curve by querying AP values at fixed IoUs, while sortedAP identifies the exact IoU value where the AP curve drops.

matrix to square, solving the maximal assignment problem using the Hungarian algorithm, and removing matches of zero cost. The implementation details are depicted in Algorithm 1.

The Hungarian matching algorithm not only maintains the one-to-one match relationship but also maximizes the accumulated IoUs of true positive matches. The Unique Matching as a plug-in extension can be applied to both AP and PQ, making them applicable with low match thresholds and object overlap.

4.2. AP scores over the entire IoU domain

To avoid the drastic score change (Section 3.2), we propose to summarize the AP scores over the entire IoU threshold domain as a metric, instead of a single AP score or scores covering only part of the domain. By using our proposed Unique Matching approach, the mAP can be straight-

Algorithm 1 Unique Matching

Require: ground truth $\mathcal{G} = \{g_1, g_2, \ldots, g_M\}$, prediction $\mathcal{P} = \{p_1, p_2, \ldots, p_N\}$, matching threshold t
Ensure: the match matrix $\mathbf{TP} \in \{true, false\}^{N \times M}$
  Initialize the cost matrix $\mathbf{Cost} \in R^{N \times M}$
  for each prediction $p_i \in \mathcal{P}$ **do**
    for each ground truth object $g_j \in \mathcal{G}$ **do**
      if $IoU(g_i, p_j) > t$ **then**
        $\mathbf{Cost}(i, j) = 1 - IoU(p_i, g_j)$
      end if
    end for
  end for
  if $N > M$ **then**
    pad $N - M$ dummy zero columns to $\mathbf{Cost}^{N \times M}$
  else
    pad $M - N$ dummy zero rows to $\mathbf{Cost}^{N \times M}$
  end if
  $\mathbf{TP}^{N \times M} \leftarrow$ run standard Hungarian algorithm, remove dummy rows or columns
  for i from 1 to N **do**
    for j from 1 to M **do**
      if $\mathbf{Cost}(i, j) == 0$ **then**
        $\mathbf{TP}(i, j) = false$
      end if
    end for
  end for

forwardly extended to the entire IoU domain, such as using a threshold collection of $\{0.1, 0.2, ..., 0.9\}$. However, querying AP scores at fixed IoU values can ignore small segmentation changes, noted as the limited resolution in Section 3.1.

We propose sorted Average Precision (sortedAP) as a new metric that is sensitive to all segmentation changes. The concept of sortedAP involves identifying all IoU values at which the AP score drops, instead of querying AP scores at fixed IoUs as the mAP. The AP score can only change at the IoUs of each object where the object transitions from true positive to false positive. Raising the matching threshold from 0 to 1 will turn all matches into non-matches one by one in the ascending order of IoU. In consequence, one non-match will diminish a true positive and introduce a false negative. Considering the sum of true and false positives is constant, we rewrite the AP score as:

$$AP_t = \frac{TP_t}{TP_t + FP_t + FN_t} = \frac{TP_t}{P + FN_t}. \quad (13)$$

We let TP_0 and FN_0 be true positives and false negatives of the maximal possible match between two sets. This can be obtained by the Unique Matching (Section 4.1) with a tiny but non-zero fuzzy threshold. All possible AP scores can then be computed by:

$$AP_{t_k} = \frac{TP_0 - k}{P + FN_0 + k}, \; k = 1, 2, ..., TP_0, \quad (14)$$

where t_k is the k-th lowest IoU of all matches. As shown in Figure 2b, any segmentation differences will be reflected by the positions of turning points. The sortedAP is defined as the area under the AP curve and can be computed by Algorithm 2. In the computation of sortedAP, the Unique Matching runs only once, while it has to be performed multiply times for different IoUs in mAP.

Algorithm 2 Sorted Average Precision

Require: ground truth $\mathcal{G} = \{g_1, g_2, \ldots, g_M\}$, prediction $\mathcal{P} = \{p_1, p_2, \ldots, p_N\}$
Ensure: the sortedAP score $s \in R$
  Match: run Unique Matching with a fuzz threshold $1e^{-6}$
  Count: true positives TP_0 and false negatives FN_0
  Sort: arrange IoUs of all matches in increasing order $[IoU_1, IoU_2, \ldots, IoU_{TP_0}]$
  Initialize: $AP_{prev} \leftarrow \frac{TP_0}{P + FN_0}$, $t_{prev} \leftarrow IoU_1$
  Initialize: $s \leftarrow t_{prev} \cdot AP_{prev}$
  for k from 1 to TP_0 **do**
    $AP_k \leftarrow \frac{TP_0 - k}{P + FN_0 + k}$, $t_k \leftarrow IoU_k$
    $s \leftarrow s + \frac{1}{2} \cdot (t_k - t_{prev}) \cdot (AP_k + AP_{prev})$
    $AP_{prev} \leftarrow AP_k$, $t_{prev} \leftarrow t_k$
  end for

5. Experiments and results

We also simulate imperfect results on the basis of ground truth segmentation from real datasets, in order to observe the behavior of different metrics. We choose the CVPPP dataset [18] and the CervicalCell dataset [15], containing clustered instances. We perform experiments per image because the effects, such as abrupt changes, will be covered when averaged over a large population. We design three experiments based on the fact that introducing errors gradually and evenly will result in a smooth decrease in the evaluation score.

Incremental falses. This experiment starts with two identical sets of objects and alternately introduces new objects into each set. At each step, we randomly duplicate an object and place it in a position where it does not overlap with any existing objects. This ensures that the newly introduced object is always a false positive or false negative. In our experiment, we add two objects to one set, then switch to the other set and repeat the process. The experiment only concerns detection errors, as objects are either perfectly matched or not matched at all.

Object erosion. At each step, morphological erosion is performed to a random object with a 3×3 structuring ele-

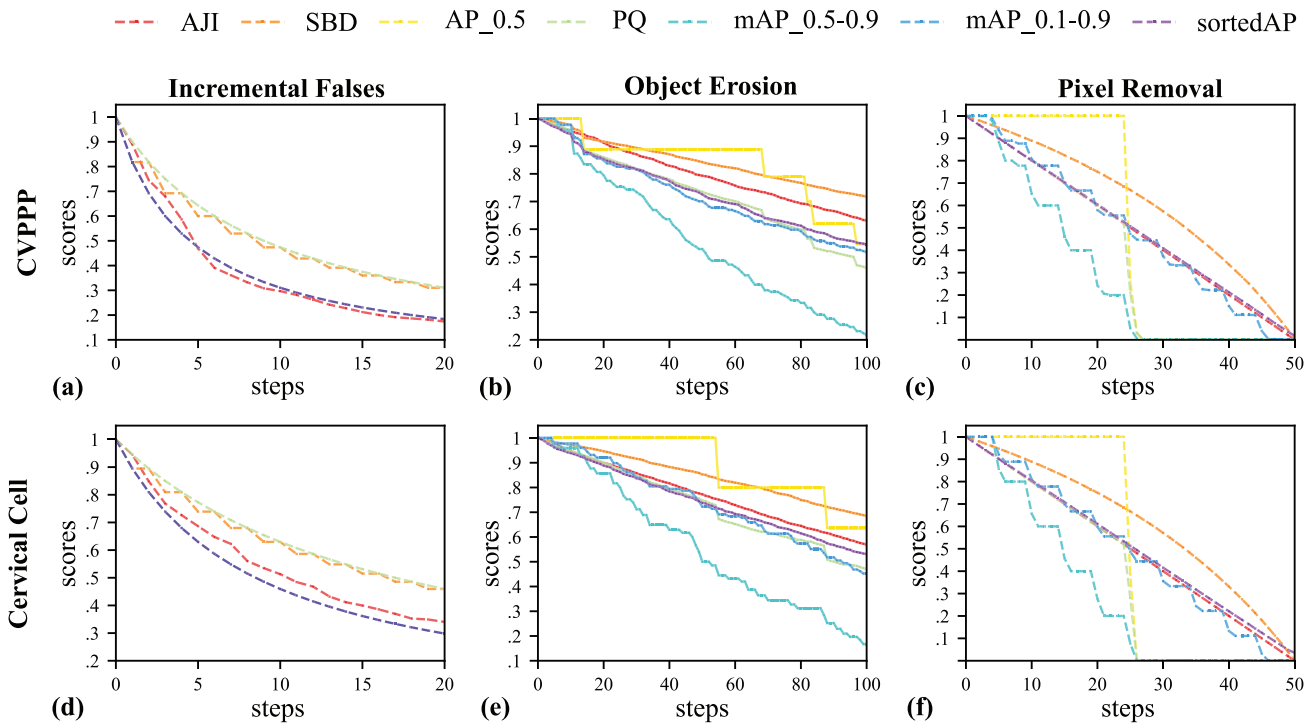

Figure 3. Comparison of different metric scores on simulated imperfect segmentation results. Three experiments (Incremental Falses, Object Erosion, and Pixel Removal) create increasingly degraded results from the ground truth of real datasets (CVPPP and CervicalCell). Since errors are gradually and evenly introduced, the evaluation score is supposed to smoothly decrease in response. In Figure 3a and Figure 3d, the curve of AP, mAP, and sortedAP are identical, shown in mixed dark blue.

ment. Consequently, the segmentation quality will steadily deteriorate. But we do not completely remove any objects. Metric scores are reported between the continuously eroded masks and the original set.

Pixel removal. Similar to the object erosion experiment, we construct a sequence of increasingly degraded results by deteriorating the segmentation quality of objects. However, instead of handling one object per step, we randomly remove a fixed portion of pixels from all objects at each step. This process simulates a situation where the segmentation of most objects is at a similar quality level. The deficiencies are more pronounced in this experiment.

The experiments conducted on the CVPPP and CervicalCell datasets yielded similar results. In the incremental false experiment (Figure 3a and Figure 3d), objects are either a perfect match with an IoU of 1 or not matched at all. Thus, segmentation inaccuracy does not play any role. The AP, mAP, and sortedAP all degrade to the same score in this case. All match-based metrics decrease smoothly as expected. In contrast, the AJI fluctuates depending on the size of introduced objects. The SBD score does not decrease in a strictly monotonic manner but instead exhibits periodic plateaus. This is an instance of the error exemption (Section 3.1). In the alternating introduction of false matches into two sets, errors introduced earlier can obscure subsequent ones.

In the object erosion and pixel removal experiment, the segmentation quality gets worse step by step. The AP and mAP also show plateaus but for a different reason from SBD in the incremental false experiment. This is due to AP's insensitivity to segmentation differences above or below the match threshold. Using multiple thresholds by mAP only improves sensitivity up to the scale of the threshold interval. PQ explicitly considers segmentation quality in the IoU range above the threshold. However, it faces a common issue of abrupt change at the match threshold as AP and mAP. Combining the two factors above, mAP exhibits a step-wise change, which is more noticeable in the pixel removal experiment (Figure 3c and Figure 3f). PQ scores will not be completely flat, but can drastically drop in a narrow IoU range. In comparison, our proposed sortedAP maintains sensitivity and continuity in all cases where other metrics fail.

6. Conclusion

In this paper, we have analyzed existing evaluation metrics for instance segmentation from the perspective of sensitivity, continuity, and equality. Although some metrics are widely used in practice, we have found that no metric

strictly satisfies all the properties under discussion. To address this gap, we propose the sortedAP, which is sensitive to any small segmentation changes, continuous over the entire IoU domain, and treats objects equally. The proposed Unique Matching approach can also be applied to AP, mAP, and PQ, allowing its use under object overlap and match IoU thresholds smaller than 0.5.

References

[1] Juan C Caicedo, Allen Goodman, Kyle W Karhohs, Beth A Cimini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, et al. Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nature methods*, 16(12):1247–1253, 2019. 1, 2, 3

[2] Long Chen, Matthias Daub, Hans-Georg Luigs, Marcus Jansen, Martin Strauch, and Dorit Merhof. High-throughput phenotyping of nematode cysts. *Frontiers in Plant Science*, page 3124, 2022. 1

[3] Long Chen, Yuli Wu, and Dorit Merhof. Instance segmentation of dense and overlapping objects via layering. *Proceedings of the British Machine Vision Conference*, 2022. 1

[4] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. Tensormask: A foundation for dense object segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2061–2069, 2019. 1

[5] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 7–9, 2017. 1

[6] Christoffer Edlund, Timothy R Jackson, Nabeel Khalid, Nicola Bevan, Timothy Dale, Andreas Dengel, Sheraz Ahmed, Johan Trygg, and Rickard Sjögren. Livecell—a large-scale dataset for label-free live cell segmentation. *Nature methods*, 18(9):1038–1045, 2021. 1

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 2, 3

[8] Estelle Glory and Robert F Murphy. Automated subcellular location determination and high-throughput microscopy. *Developmental cell*, 12(1):7–16, 2007. 1

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[10] Rohit Jena, Lukas Zhornyak, Nehal Doiphode, Pratik Chaudhari, Vivek Buch, James Gee, and Jianbo Shi. Beyond map: Towards better evaluation of instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11309–11318, 2023. 2

[11] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 1, 4

[12] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 4

[13] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging*, 36(7):1550–1560, 2017. 1, 2

[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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 3

[15] Zhi Lu, Gustavo Carneiro, Andrew P Bradley, Daniela Ushizima, Masoud S Nosrati, Andrea GC Bianchi, Claudia M Carneiro, and Ghassan Hamarneh. Evaluation of three algorithms for the segmentation of overlapping cervical cells. *IEEE journal of biomedical and health informatics*, 21(2):441–450, 2016. 5

[16] Magdalena Mazur-Milecka, Tomasz Kocejko, and Jacek Ruminski. Deep instance segmentation of laboratory animals in thermal images. *Applied Sciences*, 10(17):5979, 2020. 1

[17] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. Localization recall precision (lrp): A new performance metric for object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 504–519, 2018. 2

[18] Hanno Scharr, Massimo Minervini, Andrew P French, Christian Klukas, David M Kramer, Xiaoming Liu, Imanol Luengo, Jean-Michel Pape, Gerrit Polder, Danijela Vukadinovic, et al. Leaf segmentation in plant phenotyping: a collation study. *Machine vision and applications*, 27:585–606, 2016. 1, 2, 5

[19] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33:17721–17732, 2020. 1