arXiv:2008.03533v4 [cs.CV] 22 Jul 2022

How Trustworthy are Performance Evaluations for Basic Vision Tasks?

Tran Thien Dat Nguyen*, Hamid Rezatofighi*, Ba-Ngu Vo, Ba-Tuong Vo, Silvio Savarese, and Ian Reid

Abstract—This paper examines performance evaluation criteria for basic vision tasks involving sets of objects namely, object detection, instance-level segmentation and multi-object tracking. The rankings of algorithms by an existing criterion can fluctuate with different choices of parameters, *e.g.* Intersection over Union (IoU) threshold, making their evaluations unreliable. More importantly, there is no means to verify whether we can trust the evaluations of a criterion. This work suggests a notion of trustworthiness for performance criteria, which requires (i) robustness to parameters for reliability, (ii) contextual meaningfulness in sanity tests, and (iii) consistency with mathematical requirements such as the metric properties. We observe that these requirements were overlooked by many widely-used criteria, and explore alternative criteria using metrics for sets of shapes. We also assess all these criteria based on the suggested requirements for trustworthiness.

Index Terms—Performance evaluation, metric, object detection, instance-level segmentation, multi-object tracking.

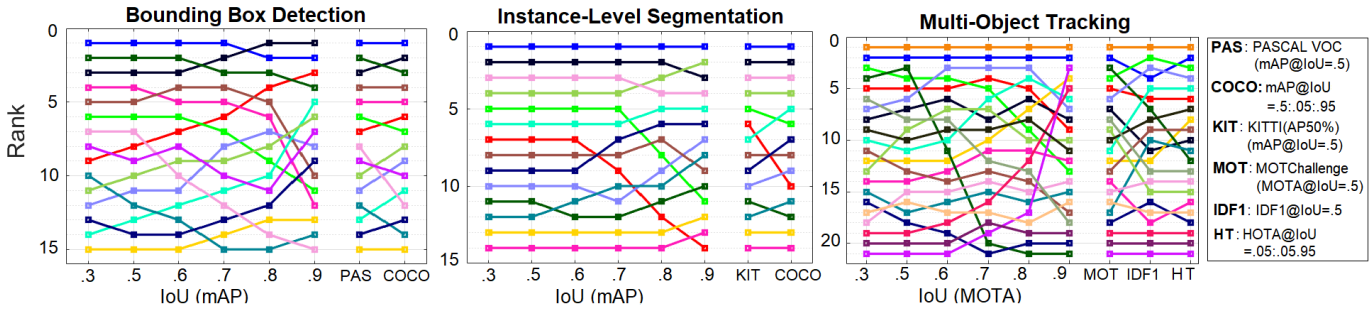

Fig. 1: Rankings of some established algorithms (details given in Fig. 14) on public datasets and challenge benchmarks, according to various (performance) criteria. The Bounding Box Detection and Instance-Level Segmentation tasks are evaluated on the COCO validation dataset, while the Multi-Object Tracking task is evaluated on the MOT17 training dataset. For a given task, each ranked algorithm is represented by a unique color. The plots show the ranking varies across IoU thresholds, with some algorithms switching from high to low ranks, and vice-versa. The algorithms are also ranked differently on different benchmarks. Such ranking variability begs the question of how trustworthy are the evaluations by these criteria.

1 Introduction

In addition to technological developments, performance evaluation is indispensable to the advancement of machine vision. It is difficult to envisage how improvements or advances can be demonstrated without performance evaluation. In this work we restrict ourselves to *basic vision tasks* involving sets of objects, namely object detection, instance-level segmentation, and multi-object tracking, where several benchmarks have been proposed to evaluate their performance, see for example [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11].

Given the importance of performance evaluation, its consistency and rigor have not received proportionate attention in computer vision. The standard practice is to rank the solutions according to certain criteria based on their outputs or *predictions/estimates* on prescribed datasets [1], [2], [4]. In general, these criteria aim to capture the similarities/dissimilarities between the *predictions* and prescribed *references*, with higher similarities (lower dissimilarities) indicating better performance. In practice, performance criteria are chosen, largely, via intuition (*e.g.* see [2], [9], [12]), while formal consideration on fairness or consistency is overlooked.

While the widely-used performance criteria for basic vision tasks are important to the progress of the field, there are a number of drawbacks.

- Firstly, the rankings by these criteria may fluctuate with the choice of parameters (*e.g.* IoU-thresholds as shown in Fig. 1). Hence, their evaluations are dubious because tuning of parameters could shift low-ranking predictions to high-ranking ones, and vice-versa. Note that the widely-used 0.5 IoU-threshold is rather arbitrary, and there are no formal justifications for its preference over other choices [1], [2], [13].
- Secondly, while these criteria are formulated based on intuition and intent, there is no principled framework to assess how meaningful their evaluations actually are, or how well they capture the intent of the evaluation exercise.

* *authors have equal contribution.*
T.T.D. Nguyen, B.-N. Vo and B.-T. Vo are with School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Australia (emails: t.nguyen1@curtin.edu.au, {ba-ngu.vo, ba-tuong.vo}@curtin.edu.au.). H. Rezatofighi is with Faculty of Information Technology, Monash University, Australia (email: hamid.rezatofighi@monash.edu). S. Savarese is with Stanford University, USA (email: ssilvio@stanford.edu). I. Reid is with School of Computer Science, University of Adelaide, Australia (email: ian.reid@adelaide.edu.au).

- Thirdly, in basic vision tasks, exact or ground truths are not available as references, and it is assumed that high similarities with approximate truths (acquired *e.g.* via annotations) imply high similarities with ground truths. However, this is not the case as demonstrated in Section 3.3 (Figs. 5, 6, 7). Consequently, there is no assurance that high-ranking predictions actually perform better than low-ranking ones, which undermines the whole purpose of performance evaluation.

In view of such drawbacks, the ensuing scientific questions are: what would a trustworthy performance criterion entail, and how to formulate trustworthy performance evaluation strategies?

This paper suggests a formalism for the trustworthiness of performance criteria, and provides an independent assessment of some widely-used criteria in basic computer vision tasks together with criteria borrowed from point pattern theory. In particular, this formalism is stipulated as a set of guidelines, whereby a trustworthy performance criteria is required to be:

(i) robust to variations in parameters for reliability;
(ii) meaningful in *sanity tests* - systematically constructed test scenarios with pre-determined rankings to capture the intent of the evaluation;
(iii) mathematically consistent - suitable analytical properties *e.g.* metric properties.

Noting that the above requirements were overlooked in widely-used criteria, such as F1, log-Average Miss Rate (log-AMR), mean Average Precision, Multi-Object Tracking Accuracy (MOTA), IDF1, and Higher Order Tracking Accuracy (HOTA), we explore some alternative performance criteria for object detection, instance-level segmentation, and multi-object tracking. These alternative criteria are *(mathematical) metrics* for sets of shapes, which integrate point pattern metrics with shape metrics. We also assess the trustworthiness of these metrics (and the above criteria) via the suggested requirements.

2 Related Work

Several performance evaluation methods have been proposed for the basic vision tasks of object detection, instance-level segmentation, and multi-object tracking.

Intersection over Union (IoU) and Generalized-IoU (GIoU) is the most commonly used family of similarity measures between two arbitrary shapes. IoU captures the similarity of the objects under comparison by a normalized measure based on the overlap in areas (or volumes) of the regions they occupy. This construction makes IoU scale-invariant, and hence the defacto base-similarity measure of many performance criteria. However, IoU is insensitive to the shape and proximity of non-overlapping shapes. To this end, a generalization that covers non-overlapping shapes, namely Generalized IoU (GIoU), was proposed in [14].

Performance evaluations for object detection and instance-level segmentation consider the similarity (or dissimilarity) between the reference and predicted sets of bounding boxes or masks. Popular performance criteria are based on the notion of *true positives*, determined by matching predictions with references such that the IoU (or GIoU) value between them is larger than a specified threshold, usually 0.5 [1], [6], [7]. Note that, the subset of true positives is dependent on the choice of thresholds. The (subset of) *false positives* is then defined to be the prediction set excluding all true positives. Similarly, the (subset of) *false negatives* (or misses) is the truth set excluding all true positives.

F1-score [11] is one of the simplest similarity measure for object detections, where the predictions are sets of bounding box coordinates with no confidence scores nor category labels, *e.g.* salient object detection [11]. F-measure captures the similarity with the harmonic mean of precision (the ratio of true positives to predictions) and recall (the ratio of true positives to truths). Specifically, let FP be the number of false positives, FN the number of false negatives and TP the number of true positives. Then the precision (P), recall (R) and F1 are are defined respectively as

$$P = \frac{TP}{TP+FP}\,,\, R = \frac{TP}{TP+FN}\,,\, F1 = 2 \times \frac{P \times R}{P+R}$$

Average Precision (AP) and mean AP (mAP) [1], [2] are perhaps the most popular performance criteria for single-category and multi-category label object detection/instance-segmentation, respectively. When predictions include confidence scores, true positives are determined by a non-optimal greedy assignment strategy that matches (with references) those with higher confidence scores first [1], [2]. Precision and recall can be expressed as a curve generated from different confidence threshold values. Let p denote the precision in order of confidence scores, and r denote the recall. Then, the AP score is defined as the area under the $p(r)$ curve, *i.e.*

$$AP = \int_0^1 p(r)dr.$$

In practice, this area is approximated by summing over a finite set of recall points [1], [2]. Given N selected recall points $r_1, ..., r_N$ such that $r_n < r_{n+1}, \forall n < N$, the approximate AP score is:

$$\widetilde{AP} = \sum_{n=1}^{N-1} (r_{n+1} - r_n)\widetilde{p}(r_{n+1}),$$

where $\widetilde{p}(r)$ is the approximation of $p(r)$ such that $\widetilde{p}(r) = \max_{\widetilde{r} \geq r} p(\widetilde{r})$.

For multi-category label predictions, the mean AP (mAP) over all categories is used. Conversely, the MS COCO Benchmark challenge [2] averages mAP across multiple IoU thresholds to reward detector with higher localization accuracy

Log-average miss rate (log-AMR) [13] is another popular performance criterion for object detection. Given the reference-prediction matches as per AP, the miss rate (MR) is plotted against the false positives per image (FPPI) rate. Similar to AP, log-AMR approximates the area under the MR-FPPI curve from a finite number of samples. For a miss rate m and FPPI rate f (sorted in the order of the prediction score), the log-AMR is given by

$$AMR = \exp\left(\frac{1}{N}\sum_{n=1}^{N} \ln\left(m(f_n)\right)\right),$$

where $f_1, ..., f_N$ are the sampled FPPI rates.

Performance evaluations for multi-object tracking consider the similarity/disimilarity between sets of reference and predicted tracks. Performance criteria usually rely on IoU or Euclidean distance to match reference tracks with predicted tracks, at each time step [9], or on the entire duration [8]. Other performance criteria such as trajectories-based measures [15], configuration distance and purity measure [10], or global mismatch error [16] were also developed based on similar constructions. A criterion based on high order matching is also recently proposed in [17].

Multi-Object Tracking Accuracy (MOTA) [9] is based on pairing, at each frame, reference and predicted objects within a separation

threshold. From this pairing, the mismatch error that captures label inconsistency is the total number of times that track identities are switched. The MOTA score is defined as one minus the normalized (by the total number of reference tracks) sum of mismatch error, and the total number (over all frames) of false positives and false negatives. Specifically, given FP_t, FN_t, $IDSW_t$ and GT_t, which are respectively the number of false positives, false negatives, ID switches and ground truth track instances at time t, the MOTA score is given by [9]:

$$MOTA = 1 - \frac{\sum_t FP_t + FN_t + IDSW_t}{\sum_t GT_t}.$$

IDF1 [8] is based on pairing reference tracks to predicted tracks so as to minimize the sum of, false positives and false negatives from each pair, for a given distance/IoU threshold. Dummy trajectories are used to account for the cardinality mismatch between the reference and predicted sets. From the optimal pairing, the IDPrecision, IDRecall, and subsequently IDF1 scores are given by the total number of false positives and false negatives of the pairs. The IDF1 score is defined as:

$$IDF1 = \frac{2IDTP}{2IDTP + IDFP + IDFN},$$

where $IDTP$, $IDFP$ and $IDFN$ are respectively the numbers of true positive ID, false positive ID and false negative ID.

Higher Order Tracking Accuracy (HOTA) [17] is designed to evaluate the long-term high-order association between predictions and references. In particular, HOTA measures the degree of alignment between trajectories and matching detections given the matches. Relying on thresholds to declare matches, the score is first evaluated over a set of localization thresholds α,

$$HOTA^{(\alpha)} = \sqrt{\frac{\sum_{c\in\{TP\}} \mathcal{A}(c)}{|TP| + |FN| + |FP|}},$$

where:

$$\mathcal{A}(c) = \frac{|TPA(c)|}{|TPA(c)| + |FNA(c)| + |FPA(c)|};$$

TP, FN, and FP are, respectively, the sets of true positives, false negatives and false positives for all predicted and ground truth instances; $TPA(c)$, $FNA(c)$, and $FPA(c)$ are, respectively, the sets of true positive associations, false negative associations and false positive associations for a given c, see [17] for details.

The final score is then obtained via marginalizing out the thresholds. In this work, we use the term "HOTA" to refer to the thresholding version of the measure while the marginalized score will be treated independently for consistent comparison with other performance criteria.

3 Guidelines for Performance Criteria

A performance criterion quantifies (by a numerical value) the similarity/dissimilarity of the output of an algorithm to a nominal reference. For basic vision tasks, namely detection, instance-level segmentation and multi-object tracking, our interest lies not only in the dissimilarity between two shapes, but dissimilarity between two (finite) sets of shapes. This dissimilarity measure can be constructed in many ways, from hand-crafted criteria based on intuition to using actual human assessments, each with its own merits and drawbacks. Regardless of its conception, the fundamental question is: how can we trust that a performance criterion does what we expect it to do?

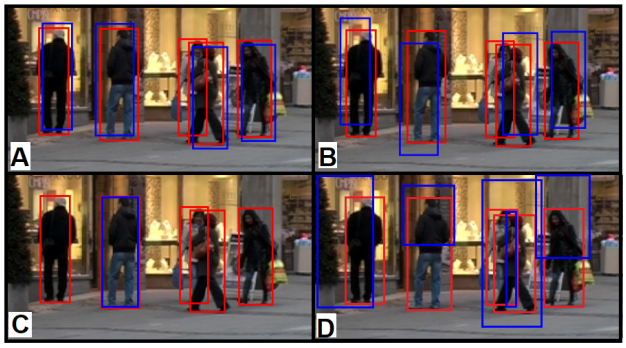

Fig. 2: People detection, with red/blue boxes representing truths/predictions. A and B detect 4 out of 5 people, with 0.8 and 0.55 IoU per person, respectively. C detects one person perfectly out of 5 people. D detects 5 out of 5 people but with an IoU of 0.3 per person. For criteria based on IoU thresholding, *e.g.* F1-score: (i) A is indistinguishable from B, if the commonly used IoU threshold of 0.5 is applied; (ii) C can rank above A and B if a high IoU threshold (above 0.8) is chosen; (iii) D is the worst detector at IoU threshold above 0.3 but becomes the best detector if an IoU threshold below 0.3 is selected.

This section attempts to answer the above question by suggesting guidelines for certifying trustworthiness of criteria based on the notions of *reliability*, *meaningfulness*, and *mathematical consistency*. Specifically, a trustworthy criterion must be reliable, meaningful and mathematically consistent. In the following, we discuss the meaning and rationale of these concepts.

3.1 Reliability

The rankings produced by a performance criterion should be robust to variations of the parameters, *e.g.* the IoU thresholds in Fig. 1. Intuitively, a criterion whose rankings are independent of the parameters is more robust than one whose rankings wildly fluctuate with variation of the parameters. More specifically, for a reliable criterion we expect that a small change in parameter values will not result in a drastic change in rankings. For example, in Fig. 2 for an IoU threshold below 0.8, detector C has the worst performance among A, B, and C. However, when the threshold is above 0.8 (no matter how small above 0.8), C becomes the best detector. Similarly, if a threshold above 0.3 is chosen, D is the worst detector. However, D becomes the best detector when the threshold is below 0.3 (no matter how small below 0.3). Such sensitivity may allow dubious promotion of certain solutions via parameter tuning. Averaging the evaluation score over a set of thresholds (*e.g.* mAP implementation in COCO multi-object detection challenge [2]) may lead to even larger ranking discrepancies for criteria with higher parameter sensitivity, although averaging the score over the a wider range of thresholds seems to improve the ranking performance (as indicated by our experiment). However, the problem with this strategy is its sensitivity to how the averaging is implemented, *i.e.* the parameters of the averaging implementation.

3.2 Meaningfulness

Reliability alone does not guarantee that a criterion is *meaningful*, *i.e.* captures the intent of the performance evaluation exercise. Consider *e.g.* the people detection task in Fig. 3, where: detector A correctly detected all 3 people with a small error for each person;

detector B correctly detected the only person but incurs a large error, and detector C has the same output as B with an additional spurious positive. Unequivocally, the detection performance of A is better than B, which, in turn, is better than C. Any performance criteria that proclaim otherwise are not meaningful.

Given that there are no analytical means in the computer vision literature for ensuring meaningfulness of performance criteria, the best option is to consider *experimental validation*–a common practice in the empirical sciences. This approach tests the criteria on a series of scenarios (real or simulated) to verify corroboration with the intent of the performance evaluation. The better the criteria fare, and the more extensive the test scenarios, the more trust we have in their meaningfulness when applied to real data.

A popular experimental validation strategy is to use humans to evaluate whether the performance criteria are meaningful [12], [18]. However, this practice inherently suffers from a number of drawbacks. Firstly, human evaluation is not scalable, and can only be applied to evaluate a small number of scenarios. Hence, extensive validation on complex scenarios involving multiple error sources, large number of objects, and large datasets is not feasible. Secondly, human evaluation is subjective and invariably leads to inconsistencies due to differences in expertise, experience and capability. For example, in object detection one prediction set may contain more false positives/negatives while another set has more severe localization error. In this case, human judgment can be subjective and assessments by different humans can be inconsistent with one another. Finally, humans are not capable of differentiating small differences in performance, and thus unable to assess the granularity of the criteria.

3.2.1 Sanity Testing

Our suggestion for assessing meaningfulness is to systematically construct a series of sanity tests, consisting of scenarios with pre-determined prediction rankings, based on the intent of the performance evaluation exercise (*e.g.* the edge-cases in Fig. 3), and verify whether the criterion's rankings corroborate the pre-determined rankings. A criterion that does not corroborate the pre-determined rankings cannot provide meaningful evaluation. On the other hand, the better the corroboration with the pre-determined rankings, the more confidence/trust we have in its ability to provide meaningful performance evaluation in practice. This strategy allows extensive validation involving multiple error sources, large number of objects, and large dataset.

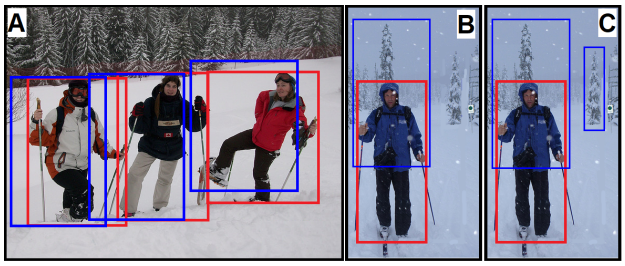

Fig. 3: Scene A: a correct prediction that there are 3 people in the scene, with an accuracy of 0.75 IoU per person (red/blue boxes represent truths/predictions). Scene B: A correct prediction that there is only person in the scene, with 0.3 IoU accuracy. Scene C is formed by adding a spurious detection to B. Any meaningful criterion should rank A above B, and B above C.

Suppose that the sources of errors for the application can be identified, *e.g.* false negatives/positives, location/shape errors, *etc.*.

- *First*, we generate/use a number of reference sets based on typical data from the application.
- *Second*, we generate a number of prediction sets with pre-determined performance ranking by perturbing the reference sets with simulated errors. Predictions generated from small perturbations are ranked higher than those generated from large perturbations. A prediction with lower rank can be generated from a given prediction by perturbing it with additional sources of error, see *e.g.* scenarios B and C in Fig. 3. This strategy enables the generation of complex scenarios with a combination of error sources and large number of objects, where the pre-determined rankings might not be obvious to the human eye, thereby enabling extensive validation not achievable with human evaluation.
- *Third*, we rank the generated predictions according to the criterion under investigation, and determine how meaningful it is by measuring the ranking discrepancy or error (with respect to the pre-determined rankings). For a given a collection of predictions, we measure the ranking error of a criterion by the *Kendall-tau* distance between its own ranking and the pre-determined ranking. This distance (also called bubble-sort distance), is a well-established (mathematical) metric for measuring dissimilarity between two rankings by counting the number of pairwise disagreements between two ranking lists [19] and has been widely used in the literature (see [20], [21], [22], [23], [24] for examples). The smaller the ranking error, the better the criterion corroborates the intent of the performance evaluation[1].

Fig. 4 shows a single trial of the proposed sanity test. The performance of predictions (a) and (b) are almost impossible for humans to distinguish via visual inspection. In contrast, from the parameters characterizing the perturbations (the dislocation magnitude that the experimenter prescribes), it is clear that prediction (a) is better than (b). Similarly, without any context, it is not clear how we would rank the performance of predictions (c) and (d) due to the complexity of the scene. However, based on the prescribed magnitude of dislocation, number of misses, falses, it is clear that (c) is better than (d). If a performance criterion corroborates well with a series of predetermined rankings, we would have more trust in its ability to capture the intent of the evaluation in ambiguous scenarios such as (e), where the ordering of the perturbation parameters provide no information to rank the predictions.

We stress that no performance criteria in the literature are guaranteed to provide meaningful evaluations in general (whether real or simulated). Moreover, there are no analytical implements nor frameworks to assess how meaningful criteria are. Our proposed methodology offers a sensible and pragmatic way to address the meaningfulness of criteria in the context of performance evaluation.

3.3 Mathematical Consistency

Relying purely on intuitive indicators is not adequate for rigorous scientific performance evaluation. This is especially true in basic vision problems, where *ground truths* are not available (except for simulated data) and only *approximate truths* can be used.

1. Other distances such as Manhattan distance and Spearman correlation (in distance form) also show almost identical behaviors to the Kendall-tau distance in our experiments

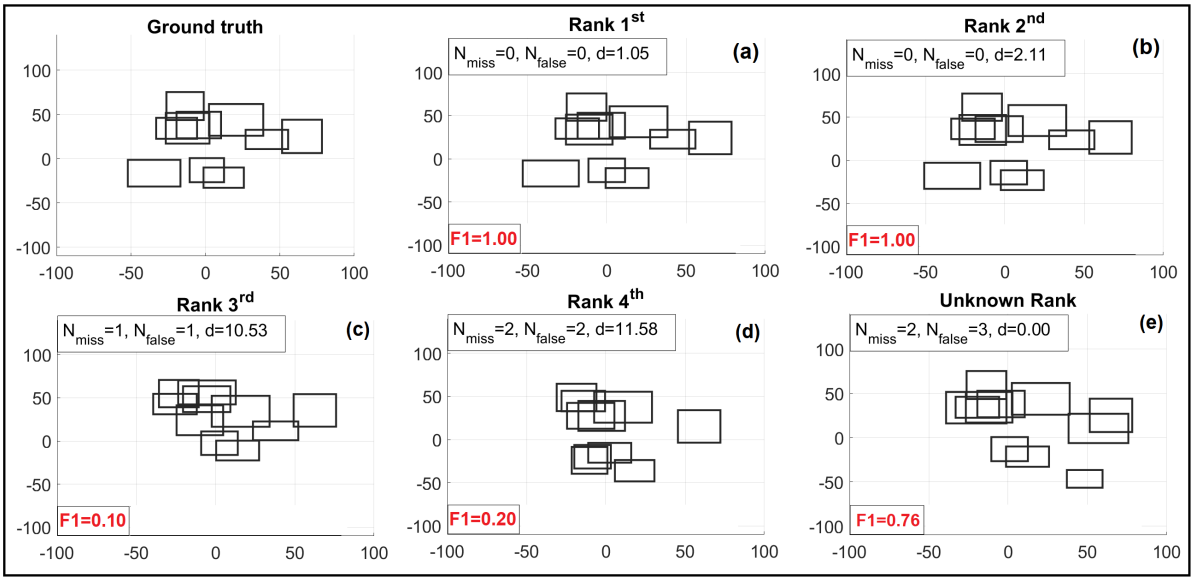

Fig. 4: Visual demonstration on the concept of parameters characterizing the perturbation in the sanity test where N_{miss}, N_{false} are the number of missed and false objects, d is the dislocation of centroid (Euclidean distance); F1 is the F1 score at IoU=0.5. Prediction (**a**) is very competitive compared to prediction (**b**) and F1 criterion cannot distinguish their performances. It is uncertain to tell if (**c**) is better than (**d**) by visual inspection but it is clear via parameters characterizing the perturbation; however, F1 criterion produces incorrect ranking order for this pair. It is uncertain to rank (**e**) among other predictions via either visualization or parameters characterizing the perturbation hence we need to solely rely on performance criteria to rank the predictions.

Keeping in mind that approximate truths are acquired through some measurement processes, *e.g.* manual annotation (which is rather subjective) and differ from the ground truth, a performance criterion only captures the similarity/dissimilarity between the predictions and approximate truths. It is implicitly assumed that the similarity/dissimilarity measure is *mathematically consistent* in the following sense: suppose that the approximate truth is "close" (*i.e.* highly similar) to the ground truth, then being "close" to the approximate truth means being "close" to the ground truth. However, this assumption does not necessarily hold even for similarity/dissimilarity between two shapes, let alone two sets of shapes, as illustrated in Fig. 5. According to the F1 criteria, even though the prediction is "closest" (indicated by the best F1 score) to the approximate truth, which in turn is "closest" to the ground truth, it bears no similarity with the ground truth whatsoever (zero F1 score). Thus, without mathematical consistency, the best possible predictions according to a criterion could be the furthest (most dissimilar) from the truth.

To further illustrate the role of mathematical consistency in prediction errors for basic vision tasks, we simulated ground truths, and approximate truths/predictions by perturbing ground truths with small/large random dislocations, and consider the F1 and mAP dissimilarity measures, *i.e.* $1-$F1 and $1-$mAP (the mAP score is calculated by assuming there is only one class, and the confidence score is 0.9 for all predictions). The red curve in Fig. 6 indicates zero dissimilarity between ground truth and approximate truth, while the blue curve shows that the normalized (prediction) error measured from approximate truth is not close to 1 (the normalized prediction error measured from ground truth). This demonstrates large discrepancies between the (prediction) errors measured from ground truth and that measured from approximate truth, even though there is no dissimilarity between these truths.

To illustrate the effect of mathematical consistency on performance

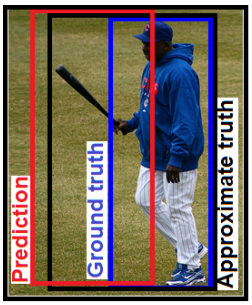

Fig. 5: For an IoU threshold of 0.5, the Prediction is "closest" to the Approximate truth ($F1=1$), which is "closest" to Ground truth ($F1=1$). Thus, the Prediction should be "close" to Ground truth, but it is as "far" as possible from Ground truth ($F1=0$)!

rankings, we generate the ground truth and prediction sets for the multi-class multi-object detection and multi-object tracking tests (by introducing perturbations to the ground truth). The true ranking order of the predictions are known (via the severity of the perturbations). Sets of approximate truth are also generated from the ground truth sets by perturbing the bounding boxes with small random dislocations (the minimum allowable IoU index between ground truth and approximate truth is 90%). Fig. 7 plots the normalized Kendall-tau distance between the true ranking vectors and the evaluated ranking vectors using ground truth references and approximate truth references, for a number of traditional criteria. Observe that at high IoU and GIoU thresholds, the (Kendall-tau) ranking error is substantially higher with approximate truth reference compared to ground truth reference. Thus, in practice where only the approximate truths are available, mathematically inconsistent criteria may not provide fair evaluations because being close to the approximate truth does not mean much.

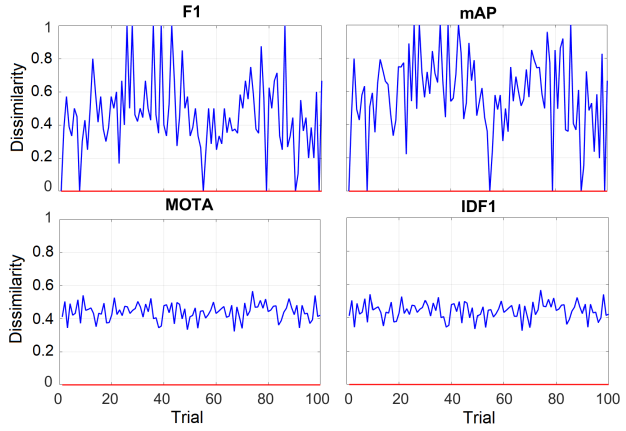

Fig. 6: **(Red)** Dissimilarity between ground truth and approximate truth. **(Blue)** Dissimilarity between approximate truth and prediction. Both are normalized against the dissimilarity between ground truth and prediction. Note that the normalized the dissimilarity between ground truth and prediction is 1. Hence for a consistent criterion the blue lines should be close to 1 (assuming the red line is close to 0).

One way to ensure mathematical consistency is to consider *(mathematical) metrics*–dissimilarity measures with certain mathematical properties. Specifically, a function $d : \mathcal{S} \times \mathcal{S} \to [0, \infty)$ is called a metric (or distance function) on the space $\mathcal{S}$, if for all $x, y, z \in \mathcal{S}$ it satisfies:

1) (Identity) $d(x, y) = 0$ if and only if $x = y$;
2) (Symmetry) $d(x, y) = d(y, x)$;
3) (Triangle inequality) $d(x, z) \leq d(x, y) + d(y, z)$.

The *triangle inequality* warrants mathematical consistency, *i.e.* if the prediction z is "close" to the approximate truth y, and assuming that the approximate truth y is "close" to the ground truth x, then the triangle inequality asserts that the prediction z is also "close" to the ground truth x. Violating the triangle inequality results in the inconsistencies of the performance criteria depicted in Fig. 6. It is also important to note that without the *Identity* property, imperfect predictions can have the same rank as the perfect prediction. Violation of this property can result in the inability to distinguish relatively clear performance differences, as illustrated in our earlier discussion on Fig. 2.

Remark: All criteria discussed in Section 2 are not mathematically consistent because they rely on thresholding the base-similarity/dissimilarity to determine the number of true positives (that solely define the criteria). In fact, (the dissimilarity forms of) these criteria violate the *Triangle Inequality* and *Identity* property which is shown in the following 1-D counter example. Let $\{x\}$ and $\{y\}$ denote the reference set and prediction set (in the case of multi-object tracking x and y would represent tracks with unit-length). Given a threshold $\theta > 0$, (keeping in mind that these sets are singletons) the number of true positives is given by the indicator function $\mathbf{1}(|x - y| \leq \theta)$ (which equals 1 if $|x - y| \leq \theta$, and 0 otherwise). Despite differences amongst the criteria in Section 2, we can abstract that any dissimilarity measure $d(\{x\}, \{y\})$ of a criterion is a function of only $\mathbf{1}(|x - y| \leq \theta)$, since number of false positives and false negatives also depend on this value. More concisely, $d(\{x\}, \{y\}) = D(\mathbf{1}(|x - y| \leq \theta))$, where D is a function such that: $D(1) = 0$ (because $d(\{x\}, \{x\}) = 0$ and $d(\{x\}, \{x\}) = D(1)$); and $D(0) > 0$ (because if $D(0) = 0$, then $d(\{x\}, \{y\}) = 0$, for all x, y, making this a trivial criterion). Now, the dissimilarity measure d violates the *Triangle Inequality* because $d(\{x - 0.6\theta\}, \{x + 0.6\theta\}) = D(0) > 0$, but $d(\{x - 0.6\theta\}, \{x\}) + d(\{x\}, \{x + 0.6\theta\}) = D(1) + D(1) = 0$. It also violates the *Identity* property because $\{x\} \neq \{x + 0.6\theta\}$ but $d(\{x\}, \{x + 0.6\theta\}) = D(1) = 0$.

4 Metric Performance Criteria

Fundamentally, performance evaluations for all three basic vision tasks in this work can be cast in terms of measuring the dissimilarity between two sets of shapes (see Fig. 8). To ensure mathematical consistency, we seek dissimilarity measures that avoid the notion of true positives–the source of unreliability and mathematical inconsistency. In this section, we explore (mathematical) metrics or distances between two sets of shapes. This is accomplished by using suitable metrics for shapes (Section 4.1) as the base-distance to construct a number of metrics for sets of shapes from various point pattern metrics (Section 4.2).

4.1 Metrics for Shapes

For any two arbitrary shapes x, y, the Intersection over Union (IoU) similarity index is given by $IoU(x, y) = \lambda(x \cap y)/\lambda(x \cup y) \in [0, 1]$, where $\lambda(\cdot)$ denotes hyper-volume. For convex shapes, the Generalized IoU index is given by $GIoU(x, y) = IoU(x, y) - \lambda(C(x \cup y) \setminus (x \cup y))/\lambda(C(x \cup y))$, where $C(x \cup y)$ is the convex hull of $x \cup y$ [14]. Note that unlike $IoU(x, y)$, $GIoU(x, y) \in [-1, 1]$. For arbitrary shapes, the definition of GIoU is given in the supplementary section of [14]. As the defacto base-similarity measure for many performance criteria, IoU/GIoU is a natural base-distances between shapes, required to construct distances between sets of shapes. The metric forms of IoU and GIoU, respectively are $\underline{d}_{IoU}(x, y) = 1 - IoU(x, y)$ and $\underline{d}_{GIoU}(x, y) = \frac{1-GIoU(x,y)}{2}$ [14], which are indeed metrics bounded by 1.

IoU/GIoU extension for shapes with confidence score: The IoU/GIoU distance can also be extended to accommodate basic vision solutions that attach to each shape a confidence score. Note that such scores can be normalized to the interval $(0, 1]$ since reference shapes have maximum confidence scores of one. To determine the IoU/GIoU distance between shapes with confidence scores, we take the Cartesian products of the shapes with their corresponding confidence scores to form augmented shapes in a higher dimensional space, and then compute the IoU/GIoU distance between these augmented shapes.

4.2 Metrics for Sets of Shapes

Our interest is the distance between two point patterns (or finite subsets) of a metric space $(\mathbb{W}, \underline{d})$, where $\underline{d} : \mathbb{W} \times \mathbb{W} \to [0, 1]$ denotes the *base-distance* between the elements of $\mathbb{W}$. Specifically, $\mathbb{W}$ is the space of arbitrary/convex shapes and the base-distance $\underline{d}$ is the IoU/GIoU distance.

One option is to consider classical set distances such as *Chamfer* [25], *Hausdorff* [26] and *Earth Mover Distance (EMD)* [27] (or Wasserstein distance [28] of order one).

The Hausdorff distance between two non-empty point patterns X and Y of $\mathbb{W}$ is defined by [26], [29]

$$d_{\mathrm{H}}(X, Y) = \max \left\{ \max_{x \in X} \min_{y \in Y} \underline{d}(x, y), \max_{y \in Y} \min_{x \in X} \underline{d}(x, y) \right\}. \quad (1)$$

This metric was traditionally used as a measure of dissimilarity between binary images. It gives a good indication of the dissimilarity

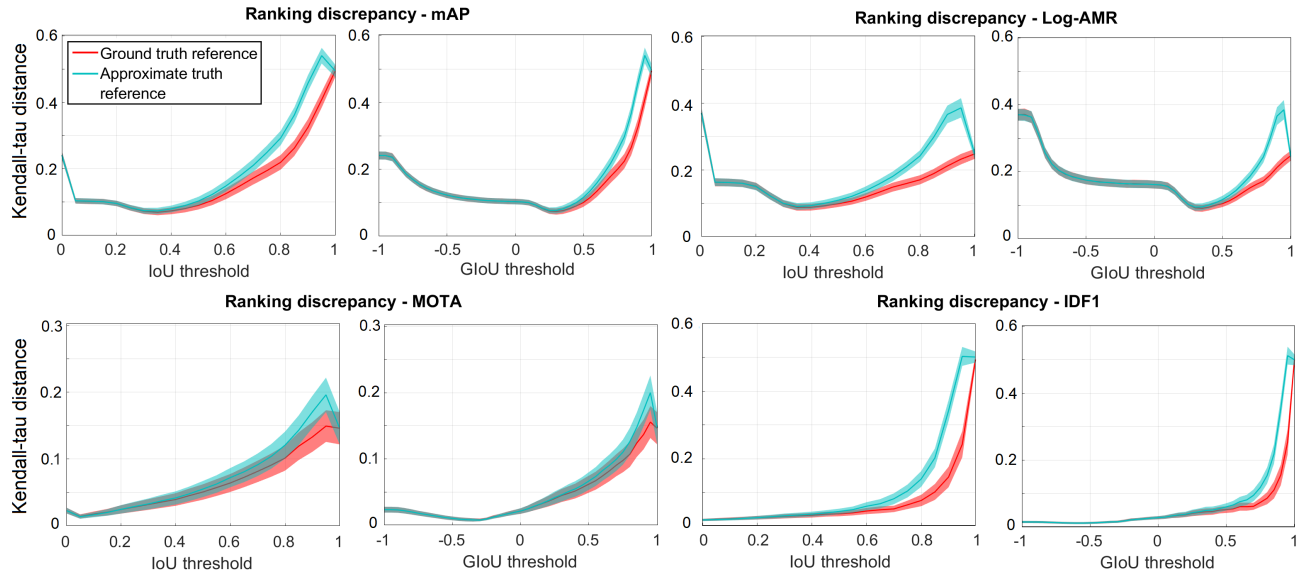

Fig. 7: Monte Carlo means of normalized Kendall-tau ranking errors (from the true ranking) for various traditional criteria at different thresholds with ground truth and approximate truth reference sets, in **detection test (top row)** and **tracking test (bottom row)**. Shaded area around each curve indicates 0.2-sigma bound.

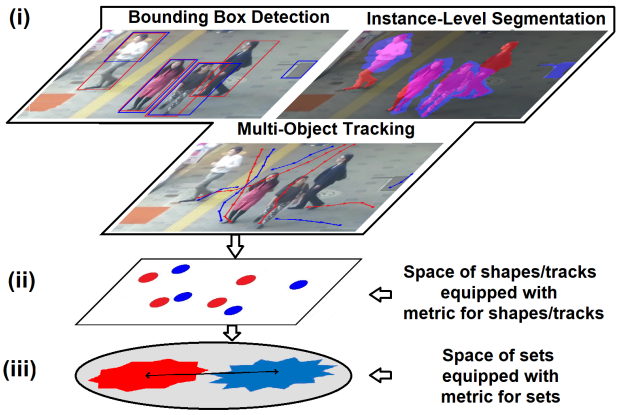

Fig. 8: (i) Truths (red) and predictions (blue) for bounding box detection, instance-level segmentation, and multi-object tracking tasks; (ii) Conceptualization (for all three tasks) of truths and predictions as two sets of points in an abstract space; (iii) Dissimilarity of these sets is measured by the (set) distance between them.

in the visual impressions that a human would typically perceive between two binary images.

In general, the Wasserstein distance (also known as Mallows distance) of order $p \geq 1$ between two non-empty point patterns $X = \{x_1, ..., x_m\}$ and $Y = \{y_1, ..., y_n\}$ is defined by [28], [29]

$$d_{\mathtt{W}}^{(p)}(X,Y) = \min_{C} \left(\sum_{i=1}^{m} \sum_{j=1}^{n} c_{i,j} \underline{d}(x_i, y_j)^p \right)^{\frac{1}{p}}, \qquad (2)$$

where $C = (c_{i,j})$ is an $m \times n$ transportation matrix, *i.e.*, the entries $c_{i,j}$ are non-negative, each row sum to $1/m$, and each column sum to $1/n$. The order p in the Wasserstein distance plays the same role as the order of the ℓ_p-distance for vectors, which is usually assumed to be 1 or 2 in most applications.

For an IoU/GIoU base-distance, which is a ratio of hyper-volumes, the Wasserstein distance of order 1 has a more natural interpretation than its higher order counterparts. This special case is commonly known as the EMD. If we consider the sets X and Y as collections of earth piles and suppose that the cost of moving a mass of earth over a distance is given by the mass times the distance. Then EMD can be considered as the minimum cost needed to build one collection of earth piles from the other.

Note that, in general, the Hausdorff and Wasserstein metrics are not defined when either of the set is empty. This is problematic for performance evaluation because it is not uncommon for the prediction set or reference set to be empty. However, when $\underline{d}$ is bounded by 1 (as per the IoU/GIoU distance), this problem can be resolved (while observing the metric properties) by defining $d_{\mathtt{H}}(X,Y) = d_{\mathtt{W}}^{(p)}(X,Y) = 1$ if one of the set is empty, and $d_{\mathtt{H}}(\emptyset,\emptyset) = d_{\mathtt{W}}^{(p)}(\emptyset,\emptyset) = 0$.

The Hausdorff and Wasserstein metrics are constructed for arbitrary sets and probability distributions. Thus, whether they capture the intent of performance evaluation in basic vision tasks, remain to be verified. The intent behind the performance criteria discussed in Section 2 is to capture the dislocation and cardinality error. What these criteria have in common is the pairing of predicted and reference points so as to minimize the sum of base-distances between the pairs, either by greedy assignment or optimal assignment. Despite differences amongst various criteria, the dislocation is determined from the matched pairs (those with base-distances below a threshold), and the cardinality error from unmatched elements, which are then combined to produce a normalized or averaged score.

An alternative to classical set distances is to find a metric that captures the above intent. Instead of thresholding the base-distance between the pairs to determine true positives, which violates the metric properties, we can capture the same intent simply by adding the minimum sum of base-distances (representing dislocation) with the number of unpaired elements (representing cardinality error), and normalize by the total number of pairs and unpaired

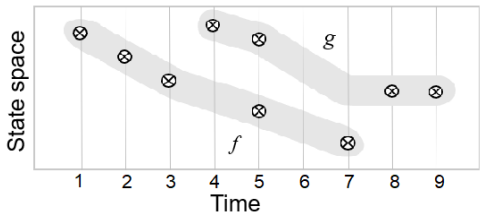

Fig. 9: Two fragmented tracks f and g in a 1-D state space. Note that at k = 6 both tracks are undefined (or non-existent).

elements. Simply put, this is the best-case per-object dislocation and cardinality error, *i.e.* for $X = \{x_1, ..., x_m\}$ and $Y = \{y_1, ..., y_n\}$,

$$d_{\mathsf{0}}(X,Y) = \frac{1}{n}\left(\min_{\pi\in\Pi_n}\sum_{i=1}^{m}\underline{d}\left(x_i, y_{\pi(i)}\right) + (n-m)\right), \quad (3)$$

if $n \geq m > 0$, where Π_n is the set of all permutations of $\{1, 2, ..., n\}$, additionally: $d_{\mathsf{0}}(X,Y) = d_{\mathsf{0}}(Y,X)$, if $m > n > 0$; $d_{\mathsf{0}}(X,Y) = 1$, if one of the set is empty; and $d_{\mathsf{0}}(\emptyset,\emptyset) = 0$. This normalized error is indeed the *Optimal Sub-Pattern Assignment (OSPA)* metric [30], which can be computed efficiently in polynomial time via optimal assignment algorithms.

Note that, although the current formulation of the metric is suitable for generic evaluation tasks where no preference is given to cardinality or localization, the original OSPA metric (see appendix Section 3.2) allows such emphasis via a cut-off parameter. Recently, an attempt to distinguish the false positives and false negatives components of cardinality error in OSPA, called the *Deficiency Aware Sub-pattern Assignment (DASA)* metric, has been introduced in [31], [32], [33].

4.3 Metrics for Sets of Tracks

For performance evaluation of multi-object tracking, the metrics for sets of shapes discussed earlier are not directly applicable because a track cannot be treated as a shape or a set of shapes due to the temporal ordering of its constituents. A *track* in a metric space $(\mathbb{W}, \underline{d})$ and discrete-time window $\mathbb{T}$, is defined as a mapping $f : \mathbb{T} \mapsto \mathbb{W}$ [34]. Its *domain* $\mathcal{D}_f \subseteq \mathbb{T}$, is the set of time instants when the object/track has a state in $\mathbb{W}$. This definition accommodates the so-called fragmented tracks, *i.e.* tracks with domains that are not intervals, see Fig. 9 for visualization in a 1-D state-space.

A meaningful distance between two sets of tracks requires a meaningful base-distance between two tracks. The most suitable for multi-object tracking is the time-averaged OSPA distance over instants when at least one of the tracks exists [34], *i.e.* for two tracks f and g

$$\widetilde{\underline{d}}(f,g) = \sum_{t\in\mathcal{D}_f\cup\mathcal{D}_g}\frac{d_{\mathsf{0}}\left(\{f(t)\},\{g(t)\}\right)}{|\mathcal{D}_f\cup\mathcal{D}_g|}, \quad (4)$$

if $\mathcal{D}_f \cup \mathcal{D}_g \neq \emptyset$, where $|\cdot|$ denotes cardinality, and $\widetilde{\underline{d}}(f,g) = 0$, if $\mathcal{D}_f \cup \mathcal{D}_g = \emptyset$. For example, the distance between the tracks in Fig. 9 is the average OSPA distance between them over all instances in $\{1, ..., 9\}$ except for $k = 6$, the instance when both tracks are undefined. The distance $\widetilde{\underline{d}}$ is indeed a metric [34] bounded by 1.

Using the Hausdorff, EMD, and OSPA metrics, respectively, with base-distance $\underline{d}$, yield the Hausdorff($\widetilde{\underline{d}}$), EMD($\widetilde{\underline{d}}$), and OSPA($\widetilde{\underline{d}}$) distances between two sets of tracks. The latter is called OSPA$^{(2)}$ (since $\widetilde{\underline{d}}$ is constructed from OSPA) and can be interpreted as the time-averaged per-track error. OSPA$^{(2)}$ takes into account errors in localization, cardinality, track fragmentation and identity switching [34]. A dropped track that later regained with the same identity incurs a smaller penalty than if it were regained with a different identity.

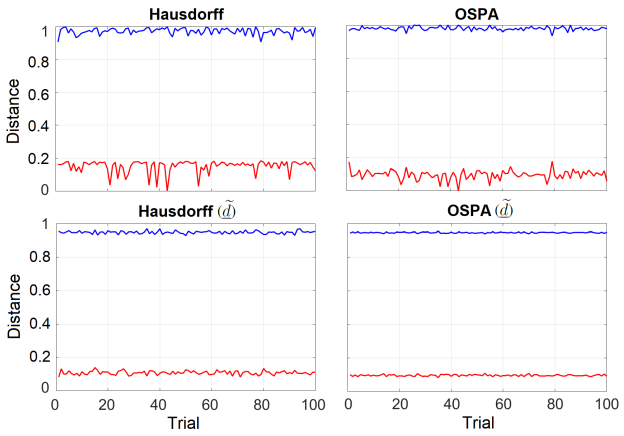

Fig. 10: **(Red)** Dissimilarity between ground truth and approximate truth. **(Blue)** Dissimilarity between approximate truth and prediction. Both are normalized against the dissimilarity between ground truth and prediction. The results are based on the same dataset as per Fig. 6. In contrast to Fig. 6, the blue lines are close to 1, which demonstrates the consistency of metric criteria .

Remark: The Hausdorff, EMD, and OSPA metrics (with both base-distances $\underline{d}$ and $\widetilde{\underline{d}}$) above are mathematically consistent (by default) and reliable (no parameters). How meaningful they are will be examined in Section 5, while further discussions can be found in Section 3 of the appendix.

In contrast to the inconsistencies of various criteria shown in Fig. 6, the (prediction) errors measured from ground truth and approximate truth are similar (for the OSPA and Hausdorff metrics) given the difference between ground truth and approximate truth is small (the same observation holds for the EMD). Moreover, compared to the criteria in Fig. 7, Tab. 1 shows that for metric criteria, the differences in (Kendall-tau) ranking errors between ground truth reference and approximate truth reference are negligible.

Note that mathematical consistency and/or reliability are not sufficient to warrant meaningful performance evaluation. Consider the simple sanity check for people detection in Fig. 3. Detector A achieves an IoU error of 0.25 for each of the 3 objects in the scene, while detector B incurs an IoU error of 0.7 even with only one object. Reiterating our previous discussion, unequivocally, detector A performs better than B. A naive metric such as the unnormalized OSPA distance (no dividing by the number of objects) is mathematically consistent (because the normalizing factor does not affect the metric axioms) and reliable (because there are no parameters). However, according to this metric B (0.7 total IoU error) has smaller prediction error than A (0.75 total IoU error), *i.e.* B performs better A, which is nonsensical. In contrast, a mathematically inconsistent criterion like F1 is more meaningful, confirming (for a 0.5 threshold) that A performs better than B, and even if the threshold is varied, would never declare B to be the better.

When the number of detected object is correct, it is obvious that a criterion should not assign a larger error to a scenario with an accurate prediction than a (different) scenario with an inaccurate

TABLE 1: Monte Carlo means (and standard deviations) of normalized Kendall-tau ranking errors for various metric criteria with ground truth and approximate truth reference sets (using the same dataset as per Fig. 7). The Kendall-tau errors between rankings based on ground truth and approximate truth are similar.

Multi-Class Multi-Object Detection: Normalized Kendall-tau ranking error (in units of 10^{-2})						
	Hausdorff		**EMD**		**OSPA**	
	IoU	**GIoU**	**IoU**	**GIoU**	**IoU**	**GIoU**
Ground truth reference	7.75 (4.62)	9.36 (4.79)	4.16 (3.64)	5.45 (4.33)	3.06 (3.35)	4.20 (3.80)
Approximate truth reference	7.94 (4.58)	9.48 (4.80)	4.48 (3.66)	5.61 (4.25)	3.37 (3.34)	4.41 (3.92)
Multi-Object Tracking: Normalized Kendall-tau ranking error (in units of 10^{-2})						
	Hausdorff($\widetilde{d}$)		**EMD($\widetilde{d}$)**		**OSPA($\widetilde{d}$)**	
	IoU	**GIoU**	**IoU**	**GIoU**	**IoU**	**GIoU**
Ground truth reference	11.3 (9.85)	11.0 (5.41)	3.63 (2.46)	5.90 (3.28)	0.536 (0.617)	0.538 (0.609)
Approximate truth reference	11.3 (9.95)	11.1 (5.37)	3.68 (2.46)	5.94 (3.29)	0.611 (0.663)	0.582 (0.636)

prediction. Hence, it is necessary to sanity-test a criterion across different scenarios, along the line of the example in Fig. 3.

To this extent, we present a sanity test that assesses the criterion's meaningfulness across different scenarios numbered from 1 to 10. In scenario k, the number of true objects is 2^k. The objects are 10 pixels by 10 pixels squares, evenly spaced so that the nearest object is more than 20 pixels away. The prediction set is the true set with each object shifted to the left by $2^{-0.5k}$ pixels. Since the predicted cardinality is correct, unequivocally, scenario 1 must have larger prediction error than scenario 2 and so on as the localization error decreases from scenario 1 to scenario 10.

Fig. 11 plots the F1_{IoU} prediction error ($1 - \text{F1}_{\text{IoU}}$), OSPA_{IoU}, un-normalized OSPA_{IoU}, EMD_{IoU}, and $\text{Hausdorff}_{\text{IoU}}$ distances for each scenario. Note that the EMD_{IoU}, $\text{Hausdorff}_{\text{IoU}}$ and OSPA_{IoU} distances exhibit identical behavior that corroborate with physical intuition as they decrease with better performance. The F1_{IoU} distance can only take the value of either 0 or 1, and is not granular enough to distinguish the prediction errors in scenarios 1, 2 and 3 to 10. Nonetheless, it still shows the general trend of improving performance. In contrast, the un-normalized OSPA_{IoU} metric[2] produces non-sensical prediction error that increases drastically with unequivocally better performance.

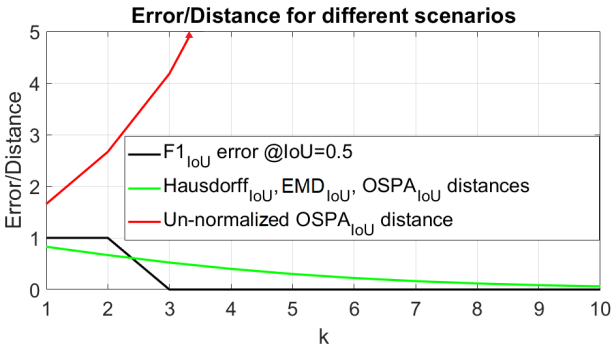

Fig. 11: The distances between true and predictions sets in scenario k of the sanity test with only dislocation of the centroid.

2. This distance takes same the form as Eq. 3 but without the normalizing constant $1/n$.

5 Assessing Criteria via Sanity Tests

In Section 3, we suggested guidelines to certify the trustworthiness of a performance criterion via its reliability, meaningfulness and mathematical consistency. In this experiment, we use sanity tests (Section 3.2) to examine the meaningfulness of different performance criteria for bounding box multi-object detection and multi-object tracking. Tests on instance-level segmentation are omitted as bounding boxes can be interpreted as masks, with both having similar properties in terms of similarity measure. The sanity test for each task is performed via 100 randomly sampled ground truth and 100 predictions sets of pre-determined ranking for each ground truth (totalling 10000 Monte Carlo trials for each task). The construction of the tests (each trial) are briefly described in the following, details can be found in Sections 4.1 and 4.2 of the appendix.

5.1 Sanity Test for Multi-Object Detection

We first sample a set of bounding boxes for the reference set, and then perturb this set to form 20 prediction sets with pre-determined ranks. The lower the prediction set is ranked: the higher the disturbance in locations and sizes, the higher the number of missed objects, false positives. Additionally, for multi-class detection test, the lower the prediction set is ranked: the higher number of predicted objects with incorrect classes and the lower the detection confidence scores for objects with correct class. In the multi-class detection test, the evaluation score/rate/distance is averaged across all classes.

5.2 Sanity Test for Multi-Object Tracking

First, we simulate the initial states of the tracks by generating a random number of random bounding boxes at random instances in the 100 time-step window. We then simulate the track lengths randomly from the interval $\{50, ..., 100\}$ and, accordingly, propagate the initial states in time via the constant velocity model to simulate a reference set (of tracks). We generate 20 predictions sets (of tracks) with pre-determined ranks by perturbing the reference set. The simulated numbers of missed objects at each time step and false tracks increase from the best prediction set to the worst. Simulated false tracks randomly appear in the scene during their active periods while their sizes vary without any dynamics. Identities swapping events are simulated so that the lower rank prediction sets have, at the same level of mutual IoU, more tracks identity swapping.

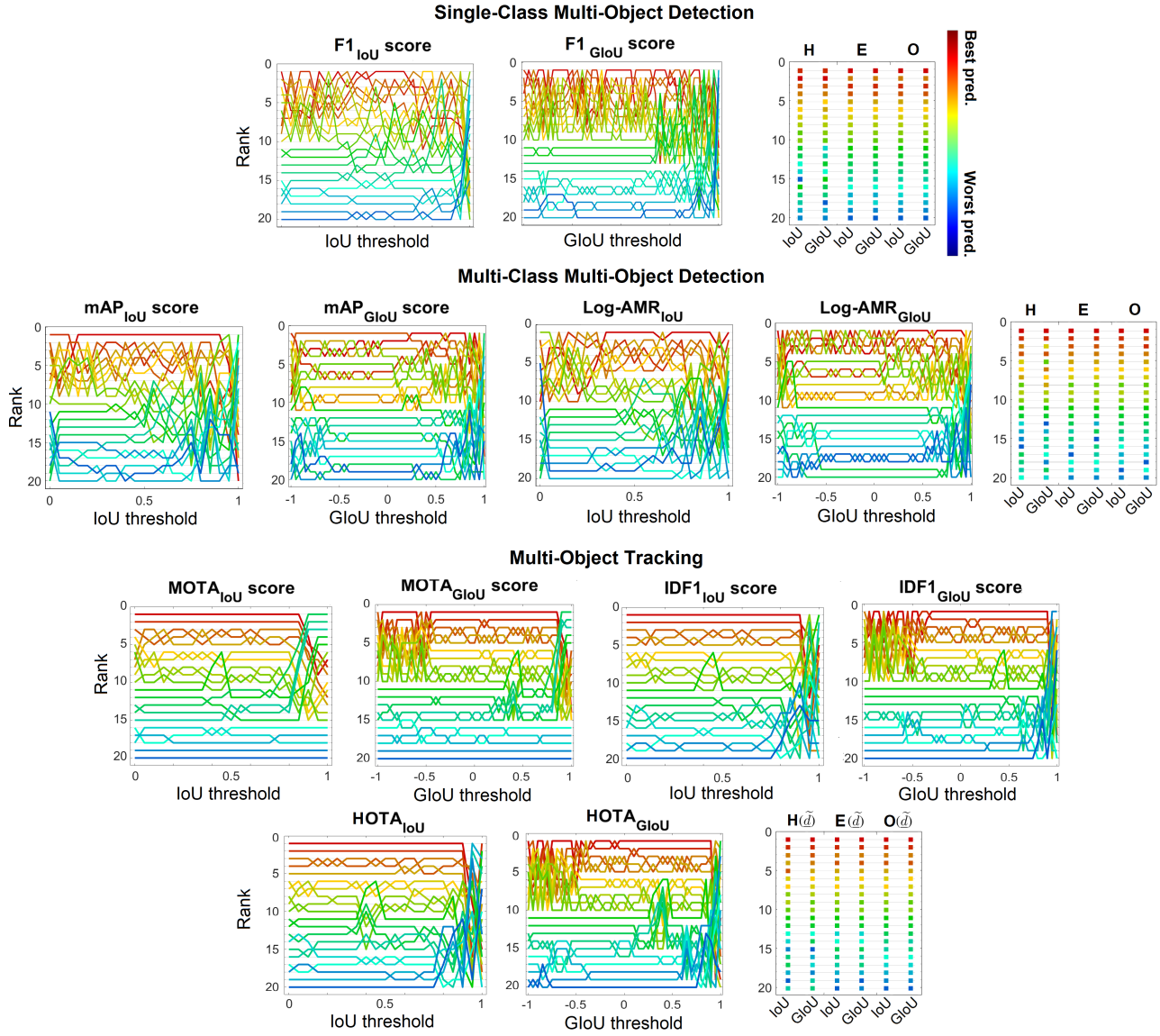

Fig. 12: Ranks of prediction sets (for a sample reference set) according to various traditional criteria over a range of IoU/GIoU thresholds, and according to Hausdorff (**H**), EMD (**E**), OSPA (**O**) metrics. The pre-determined ranks are color-coded from worst (blue) to best (red).

5.3 Results and Discussions

For completeness, we use both IoU and GIoU metrics for performance criteria in our experiment. Fig. 12 shows traditional performance criteria producing ranking orders switching severely across different IoU/GIoU thresholds. In general, more meaningful criteria should incur smaller ranking errors. Hence, Fig. 13 further confirms that the ranking accuracy (meaningfulness) of these criteria also vary considerably across the range of IoU/GIoU thresholds, albeit generally better at low thresholds. Tab. 2 shows that ranking performance at mid-scale threshold is usually not optimal, while the optimal threshold varies depending on the characteristics of the data. It also shows that partially marginalizing the parameters may produce less meaningful rankings compared to the optimal threshold in the detection test (see mAP score with IoU). While marginalizing over the entire range of threshold seems to improve the ranking performance, especially, for single-class multi-object detection and multi-object tracking tests, there is nothing to guarantee this in general. Given its insensitivity to the cardinality error, Hausdorff metric tends to have worse ranking performance than other criteria. In contrast, EMD and OSPA metrics show improved ranking performance compared to traditional criteria, with OSPA being the better metric because it also captures the intuition of traditional criteria (but without thresholding). Further, ranking results using the OSPA metric at different cut-off values are given in the appendix Section 4.

6 Real Benchmark Datasets Ranking

This section presents some observations on the traditional benchmarks and suggested metrics, in the context of how they rank various real detectors and trackers on public datasets.

COCO 2017 validation set: For bounding box detection, we use different detection models including Faster-RCNN [35], Single Shot Detector (SSD) [36] and Regional based Fully Convolutional Networks (RFCN) [37] with different backbones (Inception Network [38], [39], Residual Network (ResNet) [40], Inception ResNet [41] with atrous pooling strategy [42], Neural Architecture Search

TABLE 2: Monte Carlo means (and standard deviations) of normalized Kendall-tau ranking errors of various criteria at certain thresholds. The subscripts of IoU/GIoU indicate the threshold values; "optimal" threshold is the one with best ranking accuracy; "M-partial" indicates that the evaluation is done via averaging the score/rate over the range 0.5 to 0.95 in steps of 0.05. "M-full" indicates that the evaluation is done via averaging the score/rate over the entire range of the base-measure (excluded two extreme thresholds).

	$\text{IoU}_{0.5}$	$\text{IoU}_{\text{optimal}}$	$\text{IoU}_{\text{M-partial}}$	$\text{IoU}_{\text{M-full}}$	GIoU_{0}	$\text{GIoU}_{\text{optimal}}$	$\text{GIoU}_{\text{M-partial}}$	$\text{GIoU}_{\text{M-full}}$
Single-Class Multi-Object Detection: Normalized Kendall-tau ranking error (in units of 10^{-2})								
F1	10.0 (8.82)	7.33 (5.17)	6.68 (9.39)	2.15 (1.51)	7.89 (3.05)	7.69 (4.90)	7.49 (9.36)	**2.17 (1.36)**
Hausdorff	17.8 (9.87)				22.4 (11.1)			
EMD	3.88 (1.96)				5.16 (3.03)			
OSPA	**1.97 (1.48)**				2.22 (1.43)			
Multi-Object Multi-Class Detection: Normalized Kendall-tau ranking error (in units of 10^{-2})								
mAP	10.0 (8.90)	7.08 (5.56)	7.52 (8.71)	3.62 (2.65)	9.41 (3.81)	7.39 (5.51)	8.27 (8.71)	4.86 (3.00)
Log-AMR	9.91 (5.97)	8.42 (5.29)	6.75 (3.42)	4.31 (2.30)	16.5 (6.57)	8.80 (5.46)	7.33 (3.69)	4.95 (2.83)
Hausdorff	5.43 (2.71)				6.39 (2.88)			
EMD	2.80 (1.83)				3.50 (2.26)			
OSPA	**1.86 (1.64)**				**2.41 (1.90)**			
Multi-Object Tracking: Normalized Kendall-tau ranking error (in units of 10^{-2})								
MOTA	5.18 (5.51)	1.42 (1.60)	7.64 (7.74)	3.26 (3.90)	2.20 (2.28)	1.00 (1.39)	8.46 (8.22)	0.872 (0.806)
IDF1	3.47 (3.51)	1.84 (1.63)	3.04 (3.00)	1.38 (1.54)	2.82 (2.47)	1.24 (1.68)	3.93 (3.62)	0.676 (0.920)
HOTA	4.11 (4.17)	2.95 (2.57)	3.56 (3.70)	1.45 (1.63)	4.34 (3.21)	4.24 (3.19)	4.19 (4.18)	1.02 (1.26)
Hausdorff$(\underline{\tilde{d}})$	12.0 (9.64)				10.6 (5.27)			
EMD$(\underline{\tilde{d}})$	3.53 (2.38)				5.80 (3.29)			
OSPA$(\underline{\tilde{d}})$	**0.518 (0.580)**				**0.539 (0.577)**			

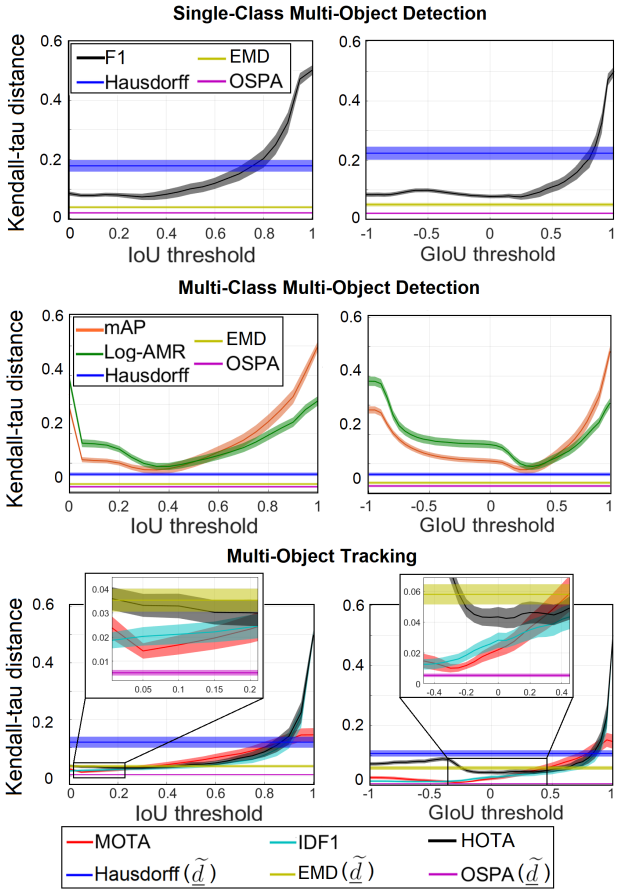

Fig. 13: Monte Carlo means of normalized Kendall-tau ranking errors for various criteria at different thresholds, in **detection tests** and **tracking test**. Shaded area around each curve indicates 0.2-sigma bound.

(NAS) [43], Mobilenets [44], Mobilenets v2 [45], Feature Pyramid Network (FPN) [46] and Pooling Pyramid Network (PPN) [47]) to detect objects. For instance-level segmentation, we use the Mask-RCNN [48] model with different network structures (FPN, ResNet, Inception ResNet) and ResNext model [49] (with FPN) to produce predictions.

MOTChallenge (MOT17) dataset: This experiment ranks predictions from 21 trackers [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70] on the MOT17 [5] leaderboard, according to various criteria. The tracking results are obtained by applying the trackers to track human in 7 training sequences and each with 3 detection methods.

Results and discussion: The rankings of established algorithms via traditional criteria are shown in Fig.1. For a given task, each ranked algorithm is represented by a unique color. Rankings for log-AMR, IDF1, and HOTA are given in the appendix (Section 5). Fig. 14 shows the rankings of these algorithms via the suggested metrics. Observe from Fig. 14, that the same metric with IoU and GIoU base-distances (for bounding boxes detection and multi-object tracking) produce similar ranking order. In addition, rankings amongst different metrics also tend to be similar to each other, especially in the segmentation task. Analogous to Fig. 13, Fig. 15 shows the differences between the rankings of traditional and metric criteria, in terms of the normalized Kendall-tau distance from the OSPA rankings (given they have the lowest ranking discrepancy as shown in Fig. 13). The behaviors of performance criteria shown in Fig. 15 corroborate their behaviors in the sanity tests (Fig. 13 and Tab. 2).

In the detection and segmentation tasks, the difference between EMD and OSPA rankings is smaller than that between Hausdorff and OSPA rankings. The ranking distances (from OSPA) are large at low and high extreme thresholds for mAP and log-AMR. The difference between COCO benchmark (averaging mAP over IoU between 0.5 and 0.95) and OSPA rankings is smaller than that between PASCAL VOC/KITTI(AP50%) benchmark (mAP with IoU of 0.5) and OSPA rankings. The mAP ranking distance at its optimal threshold (respecting to OSPA rankings) is smaller than the distance between COCO and OSPA rankings. These behaviors agree with the sanity test results. The IoU thresholds at which mAP and log-AMR rankings are the closest to of OSPA occur at around 0.8 for both detection and segmentation tasks; in the sanity test, this threshold is around 0.4. This can be explained by the variation

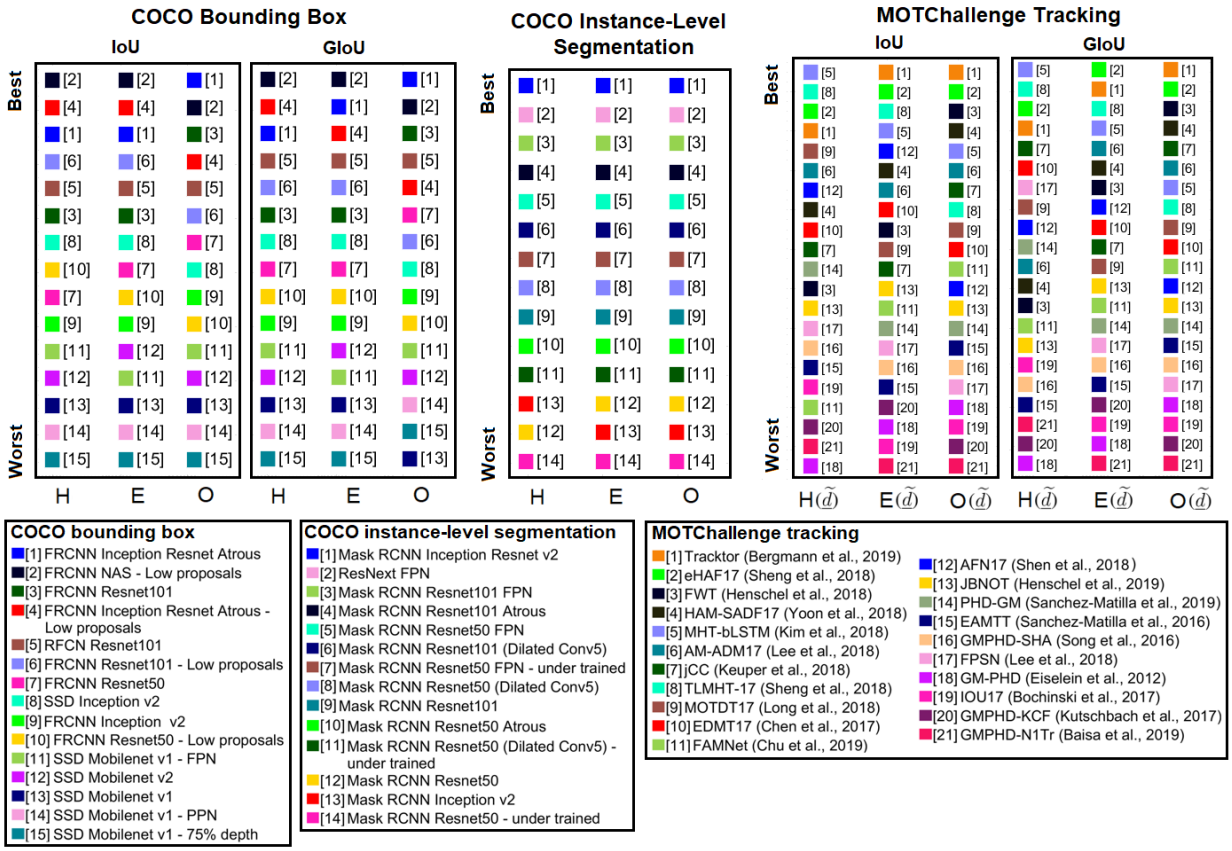

Fig. 14: Ranks of real algorithms via **H**: Hausdorff, **E**: EMD and **O**: OSPA metrics on public datasets in COCO bounding box detection, COCO instance-level segmentation and MOTChallenge tracking experiments.

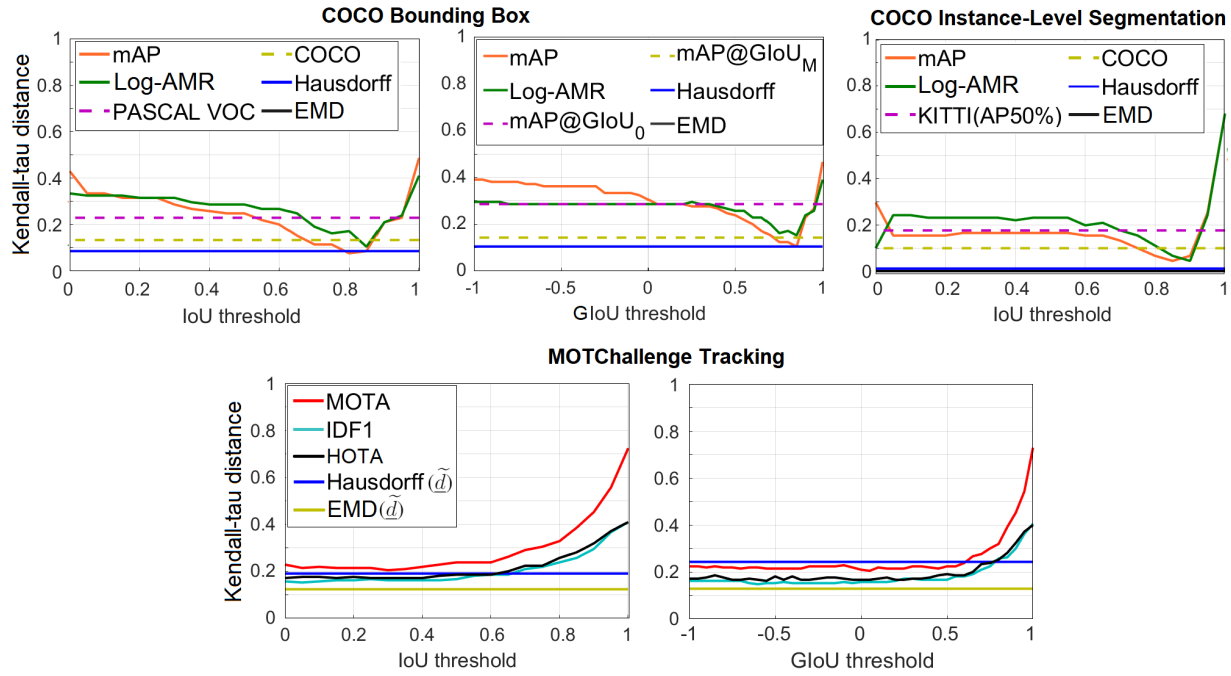

Fig. 15: Normalized Kendall-tau distances between rankings of OSPA and of other performance criteria. **PASCAL VOC** and **COCO** benchmarks use the mAP calculations of [1] and [2] respectively. **KITTI(AP50%)** is mAP calculated at 0.5 IoU overlap.

in prediction sets quality. In the MOTChallenge experiment, the differences between IDF1/HOTA and OSPA rankings are similar and lower than that between MOTA and OSPA rankings at all thresholds. MOTA, IDF1 and HOTA rankings diverge from those of OSPA at the high extreme threshold while being closer at low thresholds. With the IoU base-distance, Hausdorff and EMD rankings are close to those of OSPA. With the GIoU base-distance, the difference between Hausdorff and OSPA rankings is higher than those of between OSPA and IDF1/HOTA/EMD rankings. These trends are similar to the sanity test results in Fig. 13. For completeness, rankings of different algorithms on real benchmark datasets evaluated with the OSPA metric at different cut-off values are also provided in the appendix Section 5.

7 Conclusions

We have suggested the notion of trustworthiness for performance evaluation criteria in basic vision problems by requiring them to be mathematically consistent, meaningful and reliable. We also suggested some metrics for sets of shapes as mathematically consistent and reliable alternatives over the (neither mathematically consistent nor reliable) traditional criteria, and assessed their meaningfulness. Our experiments indicated that metrics which capture the intuition behind traditional criteria are more meaningful than other metrics and the traditional criteria. This also means that the most meaningful metric is indeed the most trustworthy because it is also mathematically consistent and reliable (by default). While our study is by no means comprehensive, we hope it paves the way towards a richer and versatile set of performance evaluation tools for computer vision.

8 Acknowledgments

This work was supported by the Western Australia DSC Collaborative Research Funding Scheme (2020) and the Australian Research Council under Discovery Project DP170104584.

APPENDIX

1 ON TRADITIONAL PERFORMANCE CRITERIA

In this section, we show that criteria based on the notion of true positives violate the triangle inequality and identity property. For a similarity measure s, we define its corresponding dissimilarity measure between a reference set $\{x\}$ and a prediction set $\{y\}$ as $d_s(\{x\},\{y\}) = 1 - s(\{x\},\{y\})$. For traditional set similarity measures, this form of dissimilarity measure has the same property as the abstract counterpart defined in the 1-D counter example at the end of Section 3.3 of the main text. If x and y are bounding boxes, the distance $|x-y|$ can be defined as IoU or GIoU distance (denoted $d_{IoU}(x,y)$ or $d_{GIoU}(x,y)$).

F1-score: For the example in Fig. 16, we can assume that there exists an IoU (or GIoU) distance threshold θ such that (i) the bounding box x can be considered as a true positive for the bounding box y (*i.e.* $d_{IoU}(x,y) < \theta$), (ii) the bounding box y can be considered as a true positive for the bounding box z (*i.e.* $d_{IoU}(y,z) < \theta$), (iii) but the bounding box x is a false positive for the bounding box z (*i.e.* $d_{IoU}(x,z) > \theta$). Therefore, in both pairs of scenarios (x,y) and (y,z), the precision, recall and consequently F1 score values are equal to one, *i.e.* $d_{F1}(\{x\},\{y\}) = d_{F1}(\{y\},\{z\}) = 0$. However, in the pair scenario (x,z), precision, recall and consequently $F1$ scores are equal to zero, *i.e.* $d_{F1}(\{x\},\{z\}) = 1$. Therefore, F1 score, as dissimilarity measure, does not fulfill the following metric properties:

- (Identity) $d_{F1}(\{x\},\{y\}) = d_{F1}(\{y\},\{z\}) = 0$, but $x \neq y \neq z$;
- (Triangle inequality) $\underbrace{d_{F1}(\{x\},\{z\})}_{1} > \underbrace{d_{F1}(\{x\},\{y\})}_{0} + \underbrace{d_{F1}(\{y\},\{z\})}_{0}$.

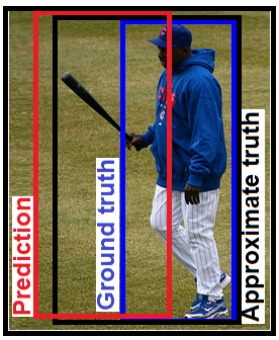

Fig. 16: Ground truth, approximate truth and prediction bounding boxes for demonstration of the inconsistency of the traditional criteria.

By altering the reference and prediction sets, it can be shown that the value of precision and recall are switched. However, $F1$ is symmetrical between the precision and recall and therefore it has the symmetry property.

Average Precision (AP): For the example in Fig. 16, with one prediction and reference in each scenario, AP is turned into the calculation of the precision only [3]. Following the same argument given for F1, precision, p, is equal 1 for the pair scenarios (x,y) and (y,z), but $p = 0$ for the pair scenario (x,z). Consequently, d_{AP} does not fulfill identity and triangle inequality as

- (Identity) $d_{AP}(\{x\},\{y\}) = d_{AP}(\{y\},\{z\}) = 0$, but $x \neq y \neq z$;
- (Triangle inequality) $\underbrace{d_{AP}(\{x\},\{z\})}_{1} > \underbrace{d_{AP}(\{x\},\{y\})}_{0} + \underbrace{d_{AP}(\{y\},\{z\})}_{0}$.

The approximated $\widetilde{AP}$ dissimilarity measure also trivially violates the above metric properties in the same example. Moreover, AP as area under precision-recall curve in exact form is symmetrical, but this property cannot be guaranteed in the approximation, *i.e.*,

- $d_{\widetilde{AP}}(\{x\},\{y\}) \neq d_{\widetilde{AP}}(\{y\},\{x\})\ \forall x,y \in \mathbb{X}$, where $\mathbb{X}$ is the space of all possible predictions.

Note that, as mAP is the average of AP over all classes, it is also not a (mathematical) metric.

Log-Average Miss Rate (log-AMR): We define the dissimilarity measure form of log-AMR as itself, *i.e.*, $d_{\text{Log-AMR}} = AMR$. From the formulation, this dissimilarity measure has the same property as the abstract dissimilarity measure defined in the 1-D counter example in the main text (for the pair of two singleton sets). In Fig. 16, for the pair scenarios (x,y) and (y,z), both the miss rate and false positive per image rate (FPPI) rate are zero hence $d_{\text{Log-AMR}}(\{x\},\{y\}) = d_{\text{Log-AMR}}(\{y\},\{z\}) = 0$. For the pair scenario (x,z), the miss rate and FPPI rate are both 1 hence $d_{\text{Log-AMR}}(\{x\},\{z\}) = 1$. Therefore, the triangle inequality and identity property do not hold.

- (Identity) $d_{\text{Log-AMR}}(\{x\},\{y\}) = d_{\text{Log-AMR}}(\{y\},\{z\}) = 0$, but $x \neq y \neq z$;
- (Triangle inequality) $\underbrace{d_{\text{Log-AMR}}(\{x\},\{z\})}_{1} > \underbrace{d_{\text{Log-AMR}}(\{x\},\{y\})}_{0} + \underbrace{d_{\text{Log-AMR}}(\{y\},\{z\})}_{0}$.

In addition, as the averaging step to calculate log-AMR is carried out over a finite samples of FPPI rate, the symmetrical property cannot be guaranteed, *i.e.*,

- $d_{\text{Log-AMR}}(\{x\},\{y\}) \neq d_{\text{Log-AMR}}(\{y\},\{x\})\ \forall x,y \in \mathbb{X}$, where $\mathbb{X}$ is the space of all possible predictions.

Further, AP and log-AMR rely on the greedy assignment to match the true to the predicted objects. This approach is indeed sub-optimal as the score and the geometrical similarity of the objects are treated independently, where the geometrical matches are conditioned on the order of the confidence score. To this extent, in Section 4.1, via our proposed IoU/GIoU extension to confidence score, we introduce a new approach to compute AP and log-AMR optimally which is shown to produce more meaningful predictions ranks in the experiment in Section 6.

MOTA: Consider unit-length tracks, following the same argument as above, the bounding box (as a single frame track) x can be considered as a true positive for the track y ($FP_t = FN_t = IDSW_t = 0$ and $d_{MOTA}(\{x\},\{y\}) = 0$), and the track y can be considered as a true positive for the track z ($FP_t = FN_t = IDSW_t = 0$ and $d_{MOTA}(\{y\},\{z\}) = 0$) (where FP_t, FN_t, and $IDSW_t$ are respectively the numbers of false positive, false negative and ID switches at time t). However, the track x is considered as false positive for the track z; therefore, there is one false positive and false negative ($FP_t = FN_t = 1$ and $d_{MOTA}(\{y\},\{z\}) = 2$). Consequently, $MOTA$ does not fulfill metric properties, *i.e.*,

- (Identity) $d_{MOTA}(\{x\},\{y\}) = d_{MOTA}(\{y\},\{z\}) = 0$,

3. there is a single prediction with an arbitrary score. Therefore, there exists no range for the confidence score.

but $x \neq y \neq z$;

- (Triangle inequality) $\underbrace{d_{MOTA}(\{x\},\{z\})}_{2} > \underbrace{d_{MOTA}(\{x\},\{y\})}_{0} + \underbrace{d_{MOTA}(\{y\},\{z\})}_{0}$.

Due to its sequential process to indicate ID switches over time, it can be also shown that $MOTA$ does not fulfill the symmetry property, *i.e.*,

- $d_{MOTA}(\{x\},\{y\}) \neq d_{MOTA}(\{y\},\{x\})\ \forall x, y \in \mathbb{T}$ where $\mathbb{T}$ is the space of all possible predicted tracks.

IDF1: Similar to the MOTA example, IDF1 dissimilarity measure between pairs of single-frame tracks (x, y) and (y, z) are $d_{IDF1}(\{x\},\{y\}) = d_{IDF1}(\{y\},\{z\}) = 0$ as the numbers of false negative ID and false positive ID are 0 and the number of true positive ID is 1. For the pair of single-frame track (x, z) the IDF1 dissimilarity measure is $d_{IDF1}(\{x\},\{z\}) = 1$ as the number of true positive ID is 1 and there are no false positive ID and false negative ID. Hence the IDF1 in dissimilarity measure form violates the following metric properties:

- (Identity) $d_{IDF1}(\{x\},\{y\}) = d_{IDF1}(\{y\},\{z\}) = 0$, but $x \neq y \neq z$;
- (Triangle inequality) $\underbrace{d_{IDF1}(\{x\},\{z\})}_{1} > \underbrace{d_{IDF1}(\{x\},\{y\})}_{0} + \underbrace{d_{IDF1}(\{y\},\{z\})}_{0}$.

HOTA: For the HOTA score defined in the main text, given that x is matched with y, hence $\sum_{c\in\{TP\}} \mathcal{A}(c) = 1$ (as $TP = \{c_{xy}\}$, where c_{xy} denotes a true positive match between x and y) hence $d_{HOTA}(\{x\},\{y\}) = 0$. Similarly, $d_{HOTA}(\{y\},\{z\}) = 0$ as y is also matched with z. However, as x is not matched with z then $d_{HOTA}(\{x\},\{z\}) = 1$ (as $TP = \emptyset$). Hence the HOTA in dissimilarity measure form violates the following metric properties:

- (Identity) $d_{HOTA}(\{x\},\{y\}) = d_{HOTA}(\{y\},\{z\}) = 0$, but $x \neq y \neq z$;
- (Triangle inequality) $\underbrace{d_{HOTA}(\{x\},\{z\})}_{1} > \underbrace{d_{HOTA}(\{x\},\{y\})}_{0} + \underbrace{d_{HOTA}(\{y\},\{z\})}_{0}$.

Greedy assignment is used for for mAP and log-AMR calculations. However, as a sub-optimal algorithm, the greedy assignment is not intuitive in some scenarios, *i.e.* see the below Fig. 17.

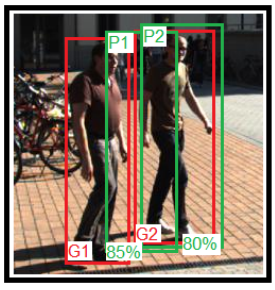

Fig. 17: As prediction P1 has higher confidence score than P2, it is considered first and as it has more overlapping with G2, it is matched to G2. P2 is not matched to any ground truth as G2 has already been taken by P1. Although it would be more intuitive if P2 is matched to G2 and P1 is matched to G1.

2 Ranking Reliability Indicators

In this section, we provide details on three intuitive indicators that can be used to measure the robustness of a performance criterion respecting to the variation of parameters. While there are many alternatives to measure the *ranking consistency*, we are particularly interested in the *purity* of the ranking order, its *distortion level* and *sensitivity* to the change of parameter.

Specifically, to measure the purity of the ranks across m independent parameters, we calculate the average number of ranking switches per predictions set. For an m-D vector $\varrho^{(i)}$ of the ranks of prediction set i^{th} across m parameters, the number of ranking switches is given by $R_S^{(i)} = \left|\{\varrho^{(i)}[j] : j \in \{1, ..., m\}\}\right| - 1$. The *average ranking switches per set* is given by $\overline{R_S} = \sum_{i=1}^{K} R_S^{(i)}/K$, where K is the number of predictions sets in consideration.

On the other hand, the degree of distortion of the ranks is reflected in the standard deviation of the elements of $\varrho^{(i)}$. For the i^{th} set, the ranking distortion is defined as $R_{std}^{(i)} = \text{std}(\varrho^{(i)})$ and *average ranking distortion per set* as $\overline{R_{std}} = \sum_{i=1}^{K} {R_{std}}^{(i)}/K$, where $\text{std}(\cdot)$ is the function to calculate the standard deviation of elements of the vector in its argument.

To indicate the ranking consistency given the sequential nature of the thresholds, we can measure the sensitivity of the ranking order against the change of parameters via taking its first order derivative with respect to the thresholds. In particular, let $\varsigma_{t_1}^{(1)}, ..., \varsigma_{t_m}^{(K)}$ be the ranking vectors (tuple of the ranks) of methods 1 to K across m thresholds from t_1 (sequentially) to t_m, the *average ranking sensitivity* across the set of these m thresholds is defined as $\overline{R_{Sen}} = \sum_{i=1}^{K}\sum_{j=1}^{m-1} \left|\left(\varsigma_{t_j}^{(i)} - \varsigma_{t_{j+1}}^{(i)}\right) / \left((t_{j+1} - t_j) \times (m-1) \times K\right)\right|$. If the thresholds are evenly spaced the factor $(t_{j+1} - t_j)$ can be omitted.

3 Further Discussions on Metrics

In the main text, we propose an extension of IoU/GIoU to accommodate the confidence score implicitly in the calculation (Section 4.1) and the use of (mathematical) metrics as alternatives for the traditional performance criteria (Section 4.2). In this section, we present detailed implementation of the proposed IoU/GIoU extension and further discussions on the optimal sub-pattern assignment (OSPA) metric.

3.1 Metric for Shapes and Confidence Score

Traditional IoU and GIoU measures only reflect the similarity between shapes geometrically but not the confidence scores of the predictions. In the main text, we propose a new method to calculate IoU/GIoU by extending the shapes to an extra dimension to accommodate the confidence score (via taking Cartesian product between the shape and corresponding score). Specifically, for a set of bounding boxes $\mathbb{B} \subset \mathbb{R}^N$ ($N = 4$ for 2-D bounding boxes) and the set of confidence score $\mathbb{S} = (0, 1]$, the set of (confidence score) augmented bounding boxes is $\mathbb{B} \times \mathbb{S}$ (where '$\times$' denotes the Cartesian product operation between sets). Visually, for 2-D bounding box scenario, the augmented bounding box is a rectangular box in 3-D. Computing IoU/GIoU distance between augmented bounding boxes can be performed similarly as for standard IoU/GIoU with steps given in Alg. 1. This extension of IoU/GIoU to the confidence score inherits all mathematical properties discussed in [14].

As discussed previously, current implementations of AP and log-AMR rely on the greedy assignment to determine the truth-to-prediction matches which do not guarantee the optimality of the matches. Basing on the IoU/GIoU extension, we propose an alternative strategy to compute AP (mAP) and log-AMR (can be extended to other criteria relying on greedy assignment). Particularly, we first calculate the pair-wise similarity scores between true and predicted objects via the IoU/GIoU extension. We then propose the use of optimal assignment algorithm to determine the matches. Given the optimal matches and a threshold value, we can determine the numbers of true positives, false positives, false negatives and then sort them in the order from the highest to the lowest confidence score. Subsequently, the standard computation for AP or log-AMR is carried out. As this approach takes into account both the confidence score and the geometrical similarity together, the assignment is indeed optimal. In Section 6, we show that it produces more meaningful ranking order compared to the greedy assignment approach.

Algorithm 1: IoU/GIoU extension to confidence score

Input: two arbitrary *N*-D convex shapes, x, y and their corresponding confidence score, $0 < s_x \leq 1$ and $0 < s_y \leq 1$.
Output: Standard IoU/GIoU distsance, $d_{IoU}(x, y)$, $d_{GIoU}(x, y)$; extended IoU/GIoU distance, $d_{\widetilde{IoU}}(x, s_x, y, s_y)$, $d_{\widetilde{GIoU}}(x, s_x, y, s_y)$

For x and y, find the smallest enclosing convex object C, then
$IoU = \frac{|x \cap y|}{|x \cup y|}$,
$d_{IoU} = 1 - IoU$,
$GIoU = IoU - \frac{|C \setminus (x \cup y)|}{|C|}$,
$d_{GIoU} = \frac{1 - GIoU}{2}$.
Construct $V_x = (x, s_x)$ and $V_y = (y, s_y)$, the *(N+1)*-D shapes which are augmented bounding boxes in $\mathbb{B} \times \mathbb{S}$.
For V_x and V_y, find the smallest enclosing convex object V_C, then
$\widetilde{IoU} = \frac{|V_x \cap V_y|}{|V_x \cup V_y|}$,
$d_{\widetilde{IoU}} = 1 - \widetilde{IoU}$,
$\widetilde{GIoU} = \widetilde{IoU} - \frac{|V_C \setminus (V_x \cup V_y)|}{|V_C|}$,
$d_{\widetilde{GIoU}} = \frac{1 - \widetilde{GIoU}}{2}$.

3.2 Optimal Sub-Pattern Assignment Metric

Consider a metric space $(\mathbb{W}, \underline{d})$, where $\underline{d} : \mathbb{W} \times \mathbb{W} \to [0, \infty)$ is the *base-distance* between the elements of $\mathbb{W}$. In its general form, the OSPA distance of order $p \geq 1$, and cut-off $c > 0$, between two point patterns $X = \{x_1, ..., x_m\}$ and $Y = \{y_1, ..., y_n\}$ is defined by [30]

$$d_0^{(p,c)}(X, Y) = \left(\frac{1}{n} \left(\min_{\pi \in \Pi_n} \sum_{i=1}^{m} \underline{d}^{(c)} \left(x_i, y_{\pi(i)} \right)^p + c^p (n - m) \right) \right)^{\frac{1}{p}}, \quad (5)$$

if $n \geq m > 0$, and $d_0^{(p,c)}(X, Y) = d_0^{(p,c)}(Y, X)$ if $m > n > 0$, where Π_n is the set of permutations of $\{1, 2, ..., n\}$, $\underline{d}^{(c)}(x, y) = \min(c, \underline{d}(x, y))$. Further $d_0^{(p,c)}(X, Y) = c$ if one of the set is empty, and $d_0^{(p,c)}(\emptyset, \emptyset) = 0$. The order p plays the same role as per the Wasserstein distance discussed in the main text, and is taken to be 1 in this work. The cut-off parameter c provides a weighting between cardinality and location errors. A large c emphasizes cardinality error while a small c emphasizes location error. However, a small c also decreases the sensitivity to the separation between the points due to the saturation of $\underline{d}^{(c)}$ at c. The

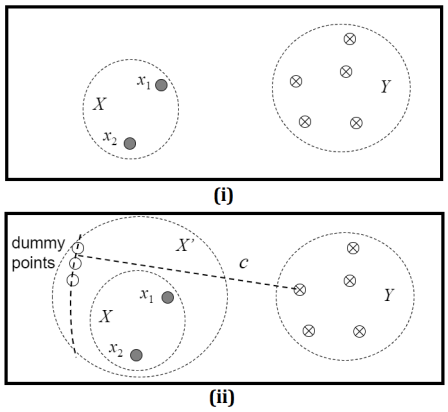

Fig. 18: OSPA distance between X and Y as the average distance between the best pairing of the points of X' and Y.

general OSPA distance above yields the following base-distance between two tracks tracks f and g:

$$\underline{d}^{(c)}(f, g) = \sum_{t \in \mathcal{D}_f \cup \mathcal{D}_g} \frac{d_0^{(c)}(\{f(t)\}, \{g(t)\})}{|\mathcal{D}_f \cup \mathcal{D}_g|},$$

if $\mathcal{D}_f \cup \mathcal{D}_g \neq \emptyset$, and $\underline{d}^{(c)}(f, g) = 0$ if $\mathcal{D}_f \cup \mathcal{D}_g = \emptyset$, where $d_0^{(c)}$ denotes the OSPA distance (the order parameter p is redundant because only sets of at most one element are considered) [34]. Note that, apart from the tracking error over the entire scenario, the OSPA$^{(2)}$ distance (OSPA distance with the above base-distance) between two sets of tracks can be plotted against time. Two algorithms with similar OSPA$^{(2)}$ errors over the entire scenario, may exhibit different OSPA$^{(2)}$ error curves over time. The monitoring of the tracking performance over time is important for the analysis/diagnosis of tracking algorithms. We refer the interested reader to [34] for more details.

The OSPA distance treats a cardinality error as if the set with smaller cardinality contained an additional (dummy) point separated from the remaining set by a base-distance of at least c. For an IoU/GIoU base-distance, such dummy point does not exist when the cut-off $c > 1$, because the largest possible separation between any two points in $\mathbb{W}$ is 1. Hence, there is no physical meaning in penalizing a cardinality error with an IoU/GIoU base-distance of $c > 1$. On the other hand, for evaluation tasks where the users do not give any preference to either localization or cardinality error, to ensure sensitivity to all IoU/GIoU base-distance separations, we require $c \geq 1$. Consequently, for an IoU/GIoU base-distance, the best cut-off choice for the OSPA distance is $c = 1$, as per Eq. 3 of Section 4 of the main text.

For evaluation tasks where it is important to emphasize on either localization or cardinality error, a cut-off $c < 1$ can be used. The smaller the value of c, the less sensitive to localization error since any pairs with base-distance greater than c is counted as a cardinality mismatch (distance saturated at c). Indeed, this cut-off parameter can be interpreted in a similar light to the IoU

threshold in traditional criteria. However, unlike traditional criteria, localization error can also be measured for matched pairs with base-distance lower than the cut-off. Traditional criteria can only count matched pairs as true positives without penalizing the actual localization error. For completeness, in Sections 4.3 and 5 (of this appendix), we also show the error and the corresponding rankings produced by OSPA metric at different cut-off values.

4 Further Details on Sanity Tests

In the main text, we briefly discuss how we set up the sanity tests. In this section, we detail the constructions of the sanity tests and provide further insights on the results.

4.1 Sanity Test for Multi-Object Detection

We first uniformly sample a reference set of N_D bounding boxes (capped at maximum 40 boxes) with centroid range $[-200, 200] \times [-200, 200]$ and size range $[20, 40]$. We then generate 20 sets of predictions (produced by 20 hypothetical detectors) by perturbing the reference set. In this test, the perturbations are dislocation of centroid, scaling of size, mis-detections, state-dependent falses, and random falses. In the multi-class test, each predicted bounding box has an additional confidence score between 0 and 1, and each true box is assigned a random enumerated class between 1 and 5 (true boxes have a score of one). The additional perturbations for the multi-class test include the mis-classifications and the reduction of confidence score (from 1) for the correctly predicted object (class).
To simulate dislocation, we assign each reference box with an enumerated label. For the box with enumerated label n in the k^{th} prediction set, we set its centroid dislocation magnitude to $d^{(k)}(n) = a^{(k)}n$, where $a^{(k)}$ is a unique constant. The centroid dislocation vector is set to

$$\begin{bmatrix} \Delta_x^{(n,k)} \\ \Delta_y^{(n,k)} \end{bmatrix} = \begin{bmatrix} ud^{(k)}(n) \\ \sqrt{\left(d^{(k)}(n)\right)^2 - \left(\Delta_x^{(n,k)}\right)^2} \end{bmatrix}, \quad (6)$$

where u is a random number between 0 and 1. Next we sample a random 2-D vector u_2 whose elements lie between 0 and 1. If $u_2[1] < 0.5$ then $\Delta_x^{(n,k)} = -\Delta_x^{(n,k)}$ and if $u_2[2] < 0.5$ then $\Delta_y^{(n,k)} = -\Delta_y^{(n,k)}$. In the multi-class detection experiment, its confidence score is scaled by $1 - r^{(k)}(n)$, where $r^{(k)}(n) = b^{(k)}n$, and $b^{(k)}$ is a constant associated with the prediction set k. For 20 sets of predictions, we use $a^{(k)} = D[k]/N_D$,where D is a 20-D vector whose elements are evenly spaced (in ascending order) numbers from 10 to 20. Similarly, $b^{(k)} = S[k]/N_D$ where S is another 20-D vectors whose elements are evenly spaced (in ascending order) numbers from 0.2 to 0.8. In this test, each box has a small random disturbance on their size.
Perturbation involving falses and mis-detections are introduced from the 11^{th} prediction set. For each experiment, we sample the 10-D vectors, P_D, P_C uniformly within the range $[0.5, 0.95]$, F_S uniformly within the range $[0.05, 0.5]$, and F_R from Poisson distributions with respective rates $1, 2, ..., 10$. The elements of P_D, P_C are then sorted in descending order and elements of F_S, F_R are sorted in ascending order. To simulate state-dependent falses in detector k, we first set the number of falses to $N_{F_R}^{(k)} = \text{round}(N_D F_S[k])$ (where round$(\cdot)$ rounds its argument to the nearest non-negative whole number). If an object is chosen to have state-dependent false, we generate a false object with the same dislocation magnitude and confidence score as the corresponding predicted object. Amongst the remaining objects (not having state-dependent falses), we simulate mis-detection by discarding the $N_M^{(k)} = \max((N_D - N_{F_R}^{(k)})(1 - P_D[k]), 0)$ objects with the largest enumerated labels, *i.e.*, objects with the highest distortion magnitudes and lowest confidence scores. The $F_R[k]$ false positive boxes are sampled using the same procedure as that per the reference boxes. For the multi-class test, we choose the $N_C^{(k)} = \max((N_D - N_{F_R}^{(k)} - N_M^{(k)})(1 - P_C[k]), 0)$ objects with largest enumerated labels to be mis-classified objects.

4.2 Sanity Test for Multi-Object Tracking

For tracking sanity tests, we set the tracking window to 100 time steps and the number of tracks, N_T, in the reference set is randomly sampled between 5 and 30. The states of the tracks are sampled from the space of bounding boxes and each track is assigned an enumerated label (1 to N_T). The length of the reference tracks are sampled between 50 and 100 time steps. The initial time of the track is then sampled between 1 and the latest possible initial time step conditioned on its length. The initial centroids of the tracks are sampled from the region $[-200, 200] \times [-200, 200]$.
For the initial size, we set a linear correlation between the height and the sampled initial y-coordinate of the tracks such that the height is limited within the range $[20, 40]$ and the higher the y-coordinate the lower the height. After the height is generated, the width is then generated by multiplying the height with a random number drawn from the interval $[0.5, 1.5]$. To generate the initial velocity of the tracks we sample the course angles and speeds uniformly from the intervals $[0, 360]$ and $[1, 5]$.
After initialization, the centroids of the tracks follow a constant velocity model. To simulate the effect of in-out camera in real tracking scenarios, we vary the heights of the tracks linearly with their y-velocity, and cap minimum height at 20. The width is kept unchanged through time.
In this test, we generate 20 sets of predictions (from 20 hypothetical trackers). The error types considered here are the dislocation of centroids, size errors, missed tracks, tracks identities confusion (swapping) and false tracks (both state-dependent and random). Following the multi-object detection test, for a track with label n, predicted by the k^{th} tracker, the centroid dislocation magnitude (at each time step) is set to $\tau^{(k)}(n) = \alpha^{(k)}n$, where $\alpha^{(k)} = T[k]/N_T$, T is a 20-D vector, whose elements are evenly spaced numbers between 20 and 40 (in ascending order). The centroid dislocation vector (at every instances that the track exists) is set to

$$\begin{bmatrix} \Delta_x^{(n,k)} \\ \Delta_y^{(n,k)} \end{bmatrix} = \begin{bmatrix} u\tau^{(k)}(n) \\ \sqrt{\left(\tau^{(k)}(n)\right)^2 - \left(\Delta_x^{(n,k)}\right)^2} \end{bmatrix}, \quad (7)$$

where u is a random number between 0 and 1. Next we sample a 2-D vector u_2 whose elements lie between 0 and 1. If $u_2[1] < 0.5$ then $\Delta_x^{(n,k)} = -\Delta_x^{(n,k)}$, and if $u_2[2] < 0.5$ then $\Delta_y^{(n,k)} = -\Delta_y^{(n,k)}$. We also add small uniform noise to the sizes of objects.
Similar to the detection sanity test, for the first 10 predictions sets we only perturb individual tracks. From the 11^{th} set, perturbations involving false tracks and missed tracks are introduced. For each experiment, we sample 10-D vectors P_{fr}, P_{sft}, and P_{id} uniformly within the range $[0.05, 1]$, and P_{rft} from Poisson distributions with respective rates $1, 2, ..., 10$. The elements of P_{fr}, P_{sft} and P_{rft} are then sorted in ascending order and P_{id} in descending order. To simulate state-dependent falses in tracker k, we first set number of tracks with state-dependent falses to $N_{sft} = N_T N_{sft}[k]$. If a track is chosen (randomly) to have state-dependent falses, we generate an

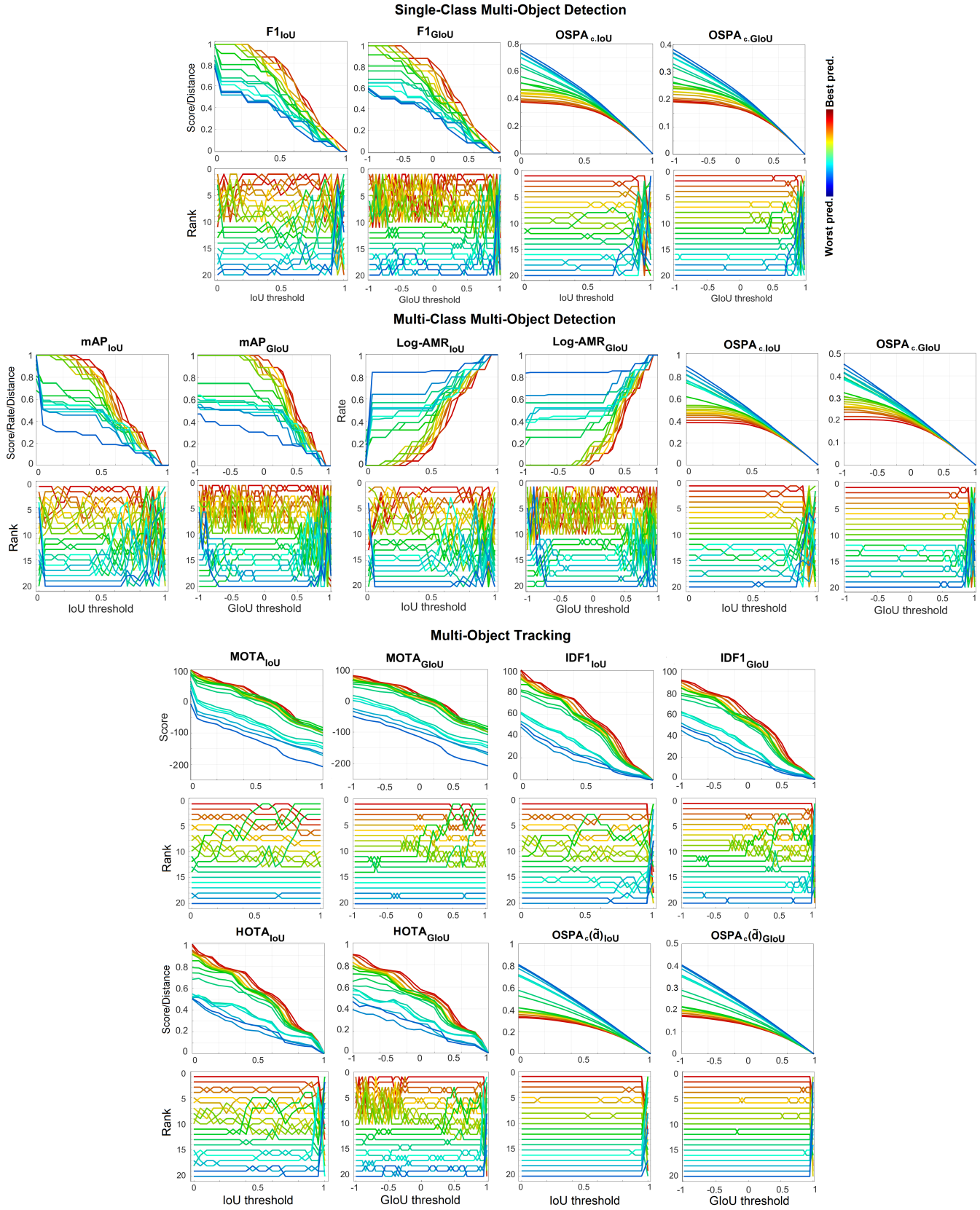

Fig. 19: Scores/rates and corresponding ranks of prediction sets (for a sample reference set) according to various criteria over a range of IoU/GIoU thresholds. The pre-determined ranks are color-coded from worst (blue) to best (red).

extra track with the same dislocation magnitude from the truth for each time step of the predicted track. At time t, we simulate missed

$$s(I^{(i,j)}, P_{id}[k]) = \begin{cases} 0 & I^{(i,j)} \leq 15 \\ 2 \times \left(\frac{I^{(i,j)}-15}{P_{id}[k]-15}\right)^2 & 15 \leq I^{(i,j)} \leq \frac{15+P_{id}[k]}{2} \\ 1 - 2 \times \left(\frac{I^{(i,j)}-15}{P_{id}[k]-15}\right)^2 & \frac{15+P_{id}[k]}{2} \leq I^{(i,j)} \leq P_{id}[k] \\ 1 & I^{(i,j)} \geq P_{id}[k] \end{cases}. \tag{8}$$

track instances by discarding the $N_{fr}^{(t,k)} = N^{(t,k)} P_{fr}[k]$ instances of tracks with the highest enumerated labels, *i.e.* worst prediction in terms of dislocation. To simulate false tracks, we introduce $P_{rft}[k]$ additional tracks with a fixed length of 10 time steps. The initial times are chosen randomly, and the initialization of these false tracks is carried out as per the reference tracks. During their active time, the false tracks appear randomly in the tracking region while their sizes vary within the range $[20, 40]$ without any dynamics.

To simulate identities swapping of detected tracks at each time step, we first calculate the mutual IoU for all pairs of tracks. For a mutual IoU of $I^{(i,j)}$ between tracks instances labeled i and j, their likelihood of swapping identities is given by Eq. 8. The tracks labels (at current time t) are swapped if this likelihood is above 0.5 and the swapping is performed in order from the pair with highest mutual IoU to the lowest.

4.3 Further Discussion and Results

For completeness, both the scores and corresponding ranks produced by criteria studied in our sanity tests are shown in Fig. 19. Further, we also include OSPA distances and corresponding ranks with different cut-off values. To distinguish this version of OSPA from the one without parameter, we use an additional subscript c, *i.e.* OSPA$_c$.

For traditional criteria, the ranking plots show a high number ranking switches. Conversely, the score plots demonstrate the rough changes of the score values across the range of thresholds in one trial of the sanity test. In contrast, we observe less ranking switches for OSPA$_c$ metric. Further, OSPA$_c$ distance decreases smoothly when the IoU/GIoU threshold increases (cut-off value decreases). This predictable behavior of OSPA$_c$ allows the users to reliably choose a cut-off threshold that reflects their evaluation intents.

In Tab. 3, we show the meaningfulness of OSPA$_c$ (evaluated using Kendall-tau metric) with different threshold settings. Note that a measure results from averaging OSPA$_c$ distances over some thresholds may not be a mathematical metric. Nonetheless, in general OSPA$_c$ is more meaningful than other measures at different threshold settings. Further, the results show that the meaningfulness of OSPA$_c$ metric at the maximum threshold is almost optimal. This demonstrates the advantage of using maximum cut-off value of 1 for generic evaluation tasks discussed in the main text.

The ranking reliability indicators shown in Tab. 4 confirm the plots in Fig. 19 which show OSPA$_c$ metric is more reliable than the traditional criteria. This is because OSPA$_c$ metric can also penalize the localization error for matches with base-distances below the cut-off values.

5 Further on Real Benchmark Datasets Ranking Comparisons

In this section, we detail the results on real dataset experiments, *i.e.* COCO detection with bounding box, COCO instance-level segmentation and MOTChallenge multi-object tracking to supplement Section 6 of the main text. In addition to the ranking plots provided in the main text, we also show the scores and variation of ranks across different thresholds in Figs. 20, 21, and 22. Evaluation results with OSPA$_c$ metric are also included for completeness.

For the COCO bounding box detection experiment, we observe that the ranks gradually change over thresholds ranges. For example, the "SSD Mobilenet v1 - 75% depth" performs relative well at low threshold but gradually gets worse when the threshold increases or the "FRCNN Inception Resnet Atrous - Low Proposals" performs worse at low thresholds but gets better at higher thresholds. In general, at low thresholds, we observe the ranks are quite stable but from value of 0.6 onward (both IoU and GIoU) the ranks start to switch more frequently. This observation is also confirmed in the log-AMR plot in Fig. 7 of [13]. Conversely, less ranking variation is observed for OSPA$_c$ metric. For COCO instance-level segmentation experiment, Fig. 21 also shows less drastic changes in rankings for OSPA$_c$ metric compared to traditional criteria.

In the MOTChallenge experiment, the ranks switch frequently across different thresholds. Especially, it is noticeable that the "jCC" method changes the rank dramatically after threshold of 0.5 on MOTA (IoU) measure. In general, we observe higher number of ranking switches at the high extreme of the thresholds ranges which indicates the criteria are more unreliable at high thresholds. Conversely, we observe the ranking orders are relatively stable across different cut-off values of OSPA$_c$ metric.

From the results shown in Tab. 5 we observe that OSPA$_c$ metric is more reliable than the traditional criteria, which quantitatively confirms the observations in Figs. 20, 21, and 22.

6 Optimal Assignment for mAP and Log-AMR

In this experiment, we construct the sanity test in the likewise manner to the mentioned multi-class multi-object detection experiment (Section 4.1). We then evaluate the predictions sets on the standard mAP, log-AMR criteria (with greedy assignment) and their corresponding optimal assignment approach. In Figs. 23 and 24, by visual inspection, it is observed that the ranks switch severely for both greedy and optimal assignments approaches in a particular trial. However, in Tab. 6 it is confirmed that the optimal assignment approach is more reliable than the greedy counterpart. In terms of the meaningfulness of the ranks, the optimal assignment approach is better than the greedy one in terms of ranking accuracy as shown in Fig. 25. For the proposed approach, while it is competitive to the Hausdorff metric, it is still less meaningful than the EMD and OSPA metrics. Tab. 7 further confirms that optimal is better than greedy assignment approach as it produces more meaningful ranking order. For both greedy and optimal assignment approaches, the partial marginalization of thresholds does not always produce more meaningful ranking order compared to the optimal threshold. However, marginalizing over the whole range of thresholds seems to improve the ranking performance overall.

TABLE 3: Monte Carlo means (and standard deviations) of normalized Kendall-tau ranking errors of various criteria at certain thresholds. The subscripts of IoU/GIoU indicate the threshold values; "optimal" threshold is the one with best ranking accuracy; "M-partial" indicates that the evaluation is done via averaging the score/rate over the range 0.5 to 0.95 in steps of 0.05. "M-full" indicates that the evaluation is done via averaging the score/rate over the entire range of the base-measure (excluded two extreme thresholds).

	$\text{IoU}_{0.5}$	$\text{IoU}_{\text{optimal}}$	$\text{IoU}_{\text{M-partial}}$	$\text{IoU}_{\text{M-full}}$	GIoU_{0}	$\text{GIoU}_{\text{optimal}}$	$\text{GIoU}_{\text{M-partial}}$	$\text{GIoU}_{\text{M-full}}$
Single-Class Multi-Object Detection: Normalized Kendall-tau ranking error (in units of 10^{-2})								
F1	10.0 (8.82)	7.33 (5.17)	6.68 (9.39)	2.15 (1.51)	7.89 (3.05)	7.69 (4.90)	7.49 (9.36)	**2.17** (**1.36**)
OSPA_c	5.19 (6.12)	**1.97** (**1.48**)	5.69 (5.81)	2.37 (1.67)	2.38 (1.67)	2.18 (1.51)	6.18 (5.52)	2.18 (1.48)
Hausdorff	17.8 (9.87)				22.4 (11.1)			
EMD	3.88 (1.96)				5.16 (3.03)			
OSPA	**1.97** (**1.48**)				2.22 (1.43)			
Multi-Object Multi-Class Detection: Normalized Kendall-tau ranking error (in units of 10^{-2})								
mAP	10.0 (8.90)	7.08 (5.56)	7.52 (8.71)	3.62 (2.65)	9.41 (3.81)	7.39 (5.51)	8.27 (8.71)	4.86 (3.00)
Log-AMR	9.91 (5.97)	8.42 (5.29)	6.75 (3.42)	4.31 (2.30)	16.5 (6.57)	8.80 (5.46)	7.33 (3.69)	4.95 (2.83)
OSPA_c	4.38 (5.72)	**1.84** (**1.60**)	4.56 (5.37)	2.06 (1.64)	2.27 (2.65)	**1.86** (**1.63**)	4.78 (5.21)	1.97 (1.68)
Hausdorff	5.43 (2.71)				6.39 (2.88)			
EMD	2.80 (1.83)				3.50 (2.26)			
OSPA	1.86 (1.64)				2.41 (1.90)			
Multi-Object Tracking: Normalized Kendall-tau ranking error (in units of 10^{-2})								
MOTA	5.18 (5.51)	1.42 (1.60)	7.64 (7.74)	3.26 (3.90)	2.20 (2.28)	1.00 (1.39)	8.46 (8.22)	0.872 (0.806)
IDF1	3.47 (3.51)	1.84 (1.63)	3.04 (3.00)	1.38 (1.54)	2.82 (2.47)	1.24 (1.68)	3.93 (3.62)	0.676 (0.920)
HOTA	4.11 (4.17)	2.95 (2.57)	3.56 (3.70)	1.45 (1.63)	4.34 (3.21)	4.24 (3.19)	4.19 (4.18)	1.02 (1.26)
$\text{OSPA}_c(\tilde{\underline{d}})$	0.869 (0.800)	**0.518** (**0.580**)	0.942 (0.843)	0.660 (0.682)	0.675 (0.670)	**0.529** (**0.567**)	1.00 (0.872)	0.558 (0.585)
$\text{Hausdorff}(\tilde{\underline{d}})$	12.0 (9.64)				10.6 (5.27)			
$\text{EMD}(\tilde{\underline{d}})$	3.53 (2.38)				5.80 (3.29)			
$\text{OSPA}(\tilde{\underline{d}})$	**0.518** (**0.580**)				0.539 (0.577)			

TABLE 4: Monte Carlo means (and standard deviations) of ranking consistency indicators over the entire range of IoU/GIoU threshold in the sanity tests.

Single-Class Multi-Object Detection

	F1_{IoU}	$\text{OSPA}_{c,\text{IoU}}$	F1_{GIoU}	$\text{OSPA}_{c,\text{GIoU}}$
$\overline{R_S}$	7.73 (1.08)	**5.17** (**1.44**)	8.98 (1.10)	5.39 (1.56)
$\overline{R_{std}}$	3.43 (0.670)	2.94 (0.745)	2.91 (0.499)	**2.22** (**0.630**)
$\overline{R_{Sen}}$	4.33 (1.22)	2.14 (1.46)	3.69 (0.665)	**1.12** (**0.793**)

Multi-Class Multi-Object Detection

	mAP_{IoU}	$\text{Log-AMR}_{\text{IoU}}$	$\text{OSPA}_{c,\text{IoU}}$	mAP_{GIoU}	$\text{Log-AMR}_{\text{GIoU}}$	$\text{OSPA}_{c,\text{GIoU}}$
$\overline{R_S}$	8.13 (1.20)	7.89 (1.09)	**4.28** (**1.70**)	10.0 (1.31)	10.4 (1.48)	4.51 (1.97)
$\overline{R_{std}}$	3.53 (0.664)	3.25 (0.602)	2.54 (0.855)	3.19 (0.534)	3.27 (0.588)	**1.93** (**0.739**)
$\overline{R_{Sen}}$	4.38 (1.42)	4.18 (1.48)	1.74 (1.59)	3.86 (1.13)	4.12 (1.40)	**0.952** (**0.980**)

Multi-Object Tracking

	MOTA_{IoU}	IDF1_{IoU}	HOTA_{IoU}	$\text{OSPA}_c(\tilde{\underline{d}})_{\text{IoU}}$	$\text{MOTA}_{\text{GIoU}}$	$\text{IDF1}_{\text{GIoU}}$	$\text{HOTA}_{\text{GIoU}}$	$\text{OSPA}_c(\tilde{\underline{d}})_{\text{GIoU}}$
$\overline{R_S}$	2.65 (1.23)	3.87 (1.28)	4.44 (1.28)	**1.64** (**0.485**)	3.31 (1.41)	4.53 (1.37)	7.14 (1.18)	1.68 (0.501)
$\overline{R_{std}}$	1.14 (0.719)	2.15 (0.633)	2.27 (0.641)	1.56 (0.232)	1.13 (0.664)	1.72 (0.492)	2.22 (0.442)	**1.13** (**0.169**)
$\overline{R_{Sen}}$	0.733 (0.290)	1.46 (0.718)	1.73 (0.745)	0.479 (0.162)	0.546 (0.183)	0.902 (0.380)	2.11 (0.439)	**0.247** (**0.0834**)

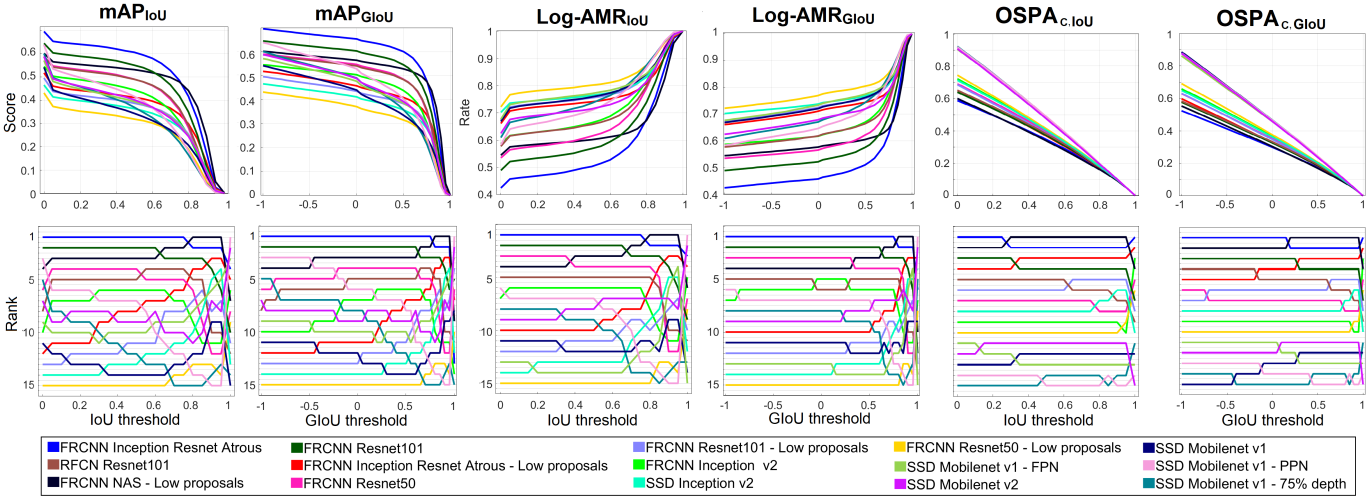

Fig. 20: Score/rate and ranks of predictions sets according to mAP, log-AMR over range of IoU/GIoU thresholds in COCO bounding box detection experiment.

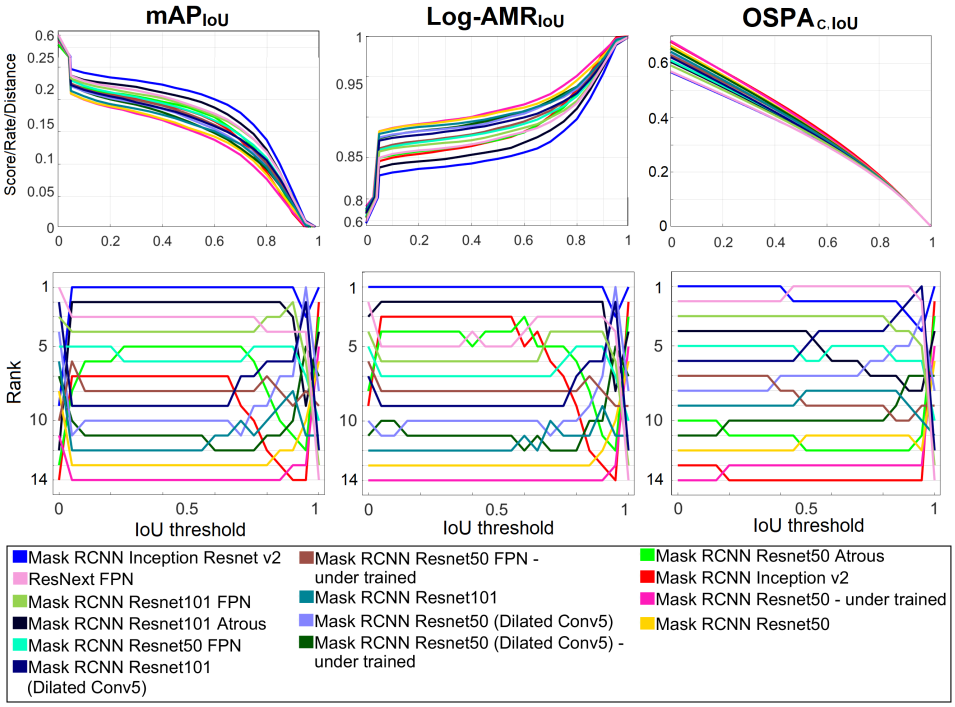

Fig. 21: Score/rate and ranks of predictions sets according to mAP, log-AMR over range of IoU thresholds in COCO instance level segmentation experiment.

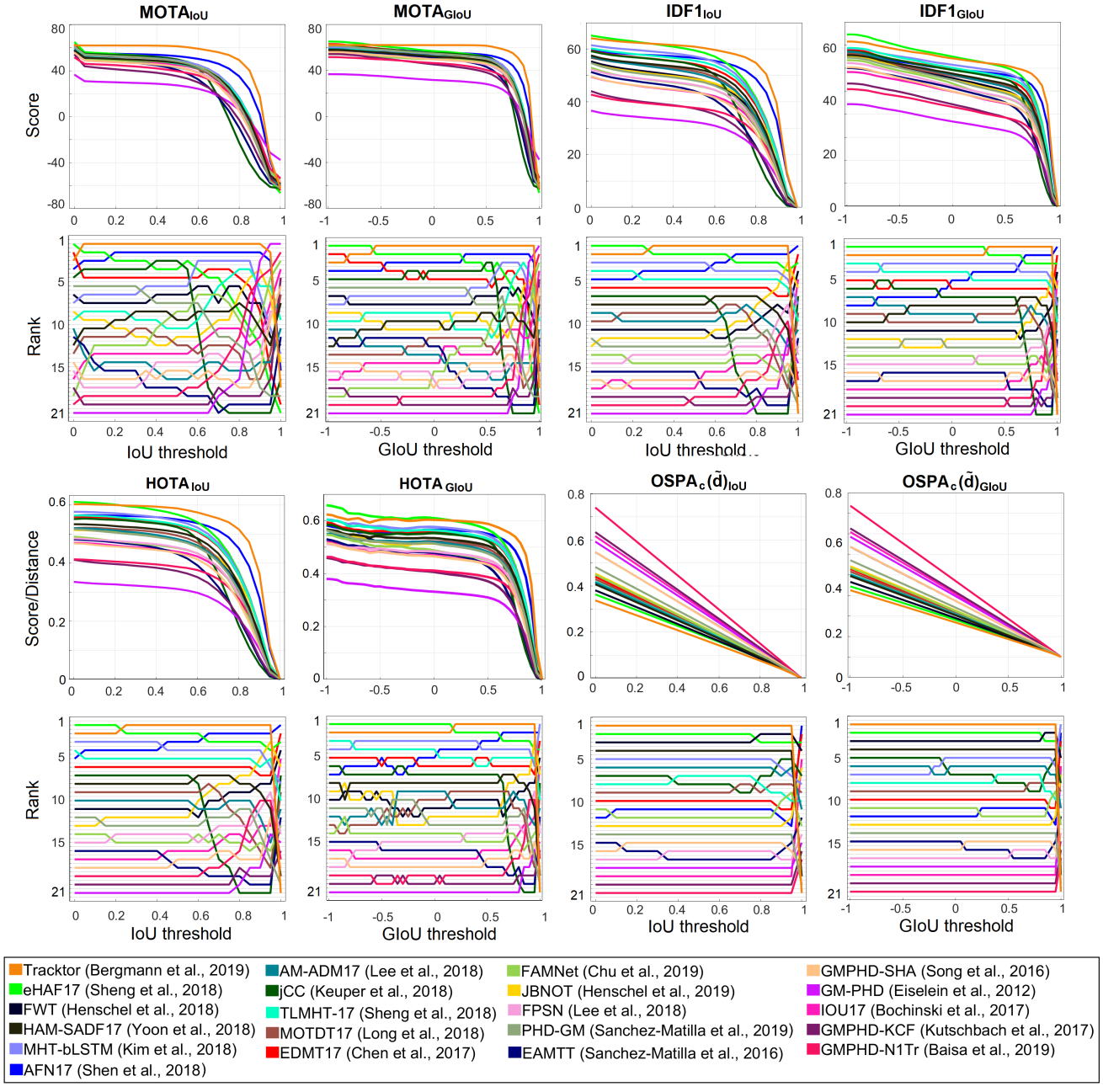

Fig. 22: Score/rate and ranks of predictions sets according to MOTA, IDF1, and HOTA over range of IoU/GIoU thresholds in MOT17 tracking experiment.

TABLE 5: Ranking consistency indicators over the entire range of IoU/GIoU threshold in real benchmark experiments.

COCO Bounding Box Detection						
	mAP_{IoU}	$\text{Log-AMR}_{\text{IoU}}$	$\text{OSPA}_{c,\text{IoU}}$	mAP_{GIoU}	$\text{Log-AMR}_{\text{GIoU}}$	$\text{OSPA}_{c,\text{GIoU}}$
$\overline{R_S}$	5.33	4.53	**1.93**	5.73	4.27	2.27
$\overline{R_{std}}$	2.14	1.98	0.775	1.76	1.45	**0.774**
$\overline{R_{Sen}}$	0.827	0.673	0.267	0.387	0.373	**0.143**

COCO Instance-Level Segmentation			
	mAP_{IoU}	$\text{Log-AMR}_{\text{IoU}}$	$\text{OSPA}_{c,\text{IoU}}$
$\overline{R_S}$	4.21	3.71	**3.00**
$\overline{R_{std}}$	2.06	1.72	**1.62**
$\overline{R_{Sen}}$	0.950	0.836	**0.514**

MOT17 Multi-Object Tracking								
	MOTA_{IoU}	IDF1_{IoU}	$\text{HOTA}^{(\alpha)}_{\text{IoU}}$	$\text{OSPA}_c(\tilde{d})_{\text{IoU}}$	$\text{MOTA}_{\text{GIoU}}$	$\text{IDF1}_{\text{GIoU}}$	$\text{HOTA}^{(\alpha)}_{\text{GIoU}}$	$\text{OSPA}_c(\tilde{d})_{\text{GIoU}}$
$\overline{R_S}$	6.05	3.90	4.05	**1.90**	6.48	4.10	4.52	2.00
$\overline{R_{std}}$	3.52	2.00	2.04	1.37	2.80	1.57	1.65	**1.05**
$\overline{R_{Sen}}$	0.938	0.695	0.695	0.405	0.526	0.352	0.526	**0.210**

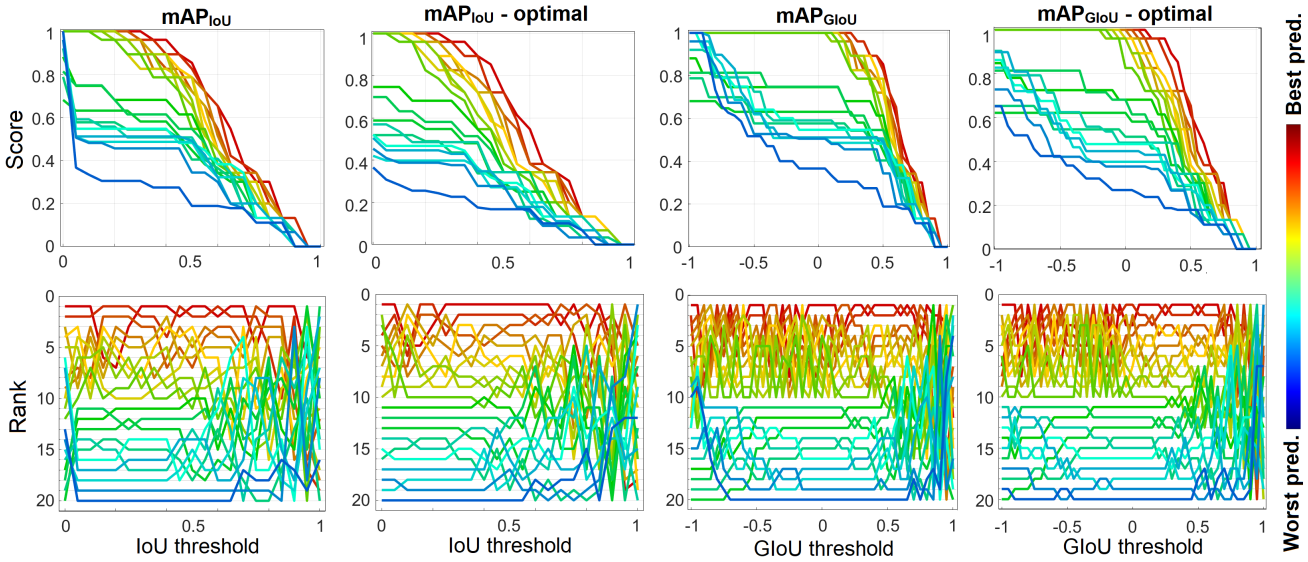

Fig. 23: mAP scores with greedy and optimal assignment approaches (top row) and the corresponding ranks of predictions (bottom row) over a range of IoU/GIoU thresholds in one trial of the multi-class multi-object detection sanity test. The pre-determined ranks are color-coded from worst (blue) to best (red).

TABLE 6: Monte Carlo means (and standard deviations) of the ranking consistency indicators for mAP and log-AMR with greedy and optimal assignment approaches.

mAP and Log-AMR with Optimal Assignment				
	mAP_{IoU}	$\text{mAP}_{\text{IoU-optimal}}$	mAP_{GIoU}	$\text{mAP}_{\text{GIoU-optimal}}$
$\overline{R_S}$	8.13 (1.20)	**7.37** (**1.52**)	10.0 (1.31)	9.01 (1.66)
$\overline{R_{std}}$	3.53 (0.664)	3.18 (0.816)	3.19 (0.534)	**2.82** (**0.657**)
$\overline{R_{Sen}}$	4.38 (1.42)	3.88 (1.65)	3.86 (1.13)	**3.39** (**1.20**)
	$\text{Log-AMR}_{\text{IoU}}$	$\text{Log-AMR}_{\text{IoU-optimal}}$	$\text{Log-AMR}_{\text{GIoU}}$	$\text{Log-AMR}_{\text{GIoU-optimal}}$
$\overline{R_S}$	7.89 (1.09)	**6.96** (**1.30**)	10.4 (1.48)	9.07 (1.66)
$\overline{R_{std}}$	3.25 (0.602)	2.73 (0.698)	3.27 (0.588)	**2.70** (**0.650**)
$\overline{R_{Sen}}$	4.18 (1.48)	3.52 (1.64)	4.12 (1.40)	**3.47** (**1.40**)

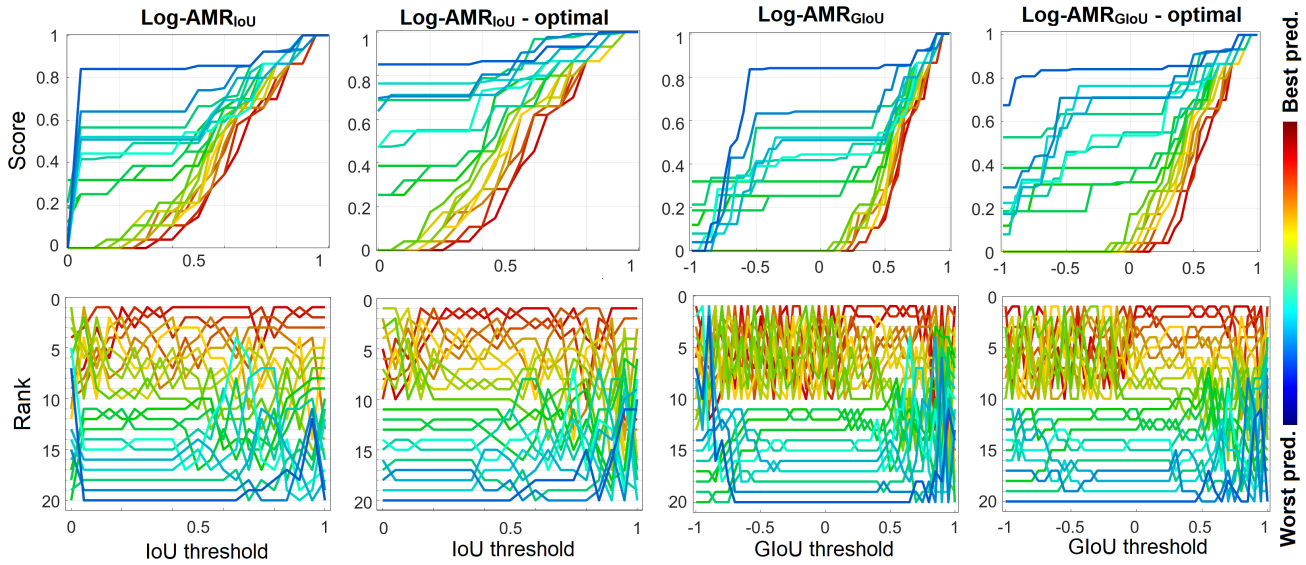

Fig. 24: Log-AMR with greedy and optimal assignment approaches (top row) and the corresponding ranks of predictions (bottom row) over a range of IoU/GIoU thresholds in one trial of the multi-class multi-object detection sanity test. The pre-determined ranks are color-coded from worst (blue) to best (red).

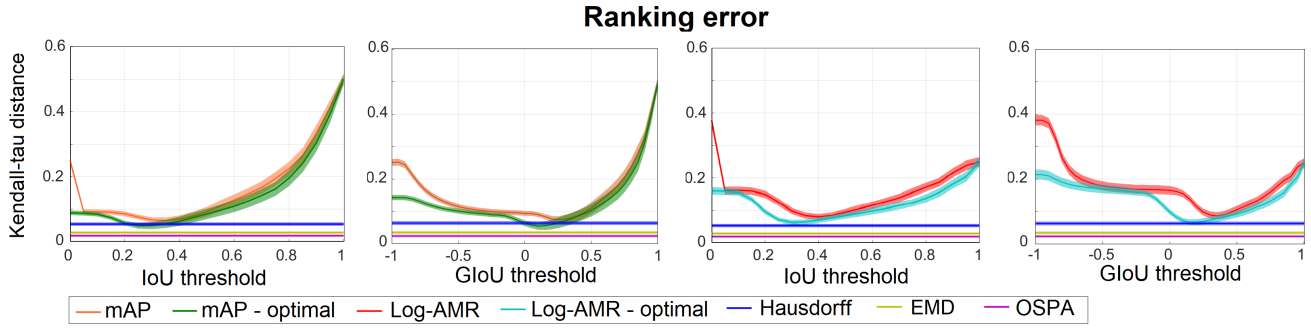

Fig. 25: Monte Carlo mean normalized Kendall-tau ranking errors (from the true ranking) of various criteria at different thresholds, in multi-class multi-object detection test with greedy and optimal assignment approaches. Shaded area around each curve indicates 0.2-sigma bound. We also show the results for metric criteria for reference.

TABLE 7: Monte Carlo means (and standard deviations) of normalized Kendall-tau ranking errors of mAP and log-AMR with greedy and optimal assignment approaches at certain thresholds. The subscripts of IoU/GIoU indicate the threshold values; "optimal" threshold is the one with best ranking accuracy; "M-partial" indicates that the evaluation is done via averaging the score/rate over the range 0.5 to 0.95 in steps of 0.05. "M-full" indicates that the evaluation is done via averaging the score/rate over the entire range of the base-measure (excluded two extreme thresholds). We also show the results for metric criteria for reference.

mAP and Log-AMR with Optimal Assignment: Normalized Kendall-tau ranking error (in units of 10^{-2})								
	IoU$_{0.5}$	IoU$_{optimal}$	IoU$_{M\text{-}partial}$	IoU$_{M\text{-}full}$	**GIoU$_{0}$**	GIoU$_{optimal}$	GIoU$_{M\text{-}partial}$	GIoU$_{M\text{-}full}$
mAP	10.0 (8.90)	7.08 (5.56)	7.52 (8.71)	3.62 (2.65)	9.41 (3.81)	7.39 (5.51)	8.27 (8.71)	4.86 (3.00)
mAP-optimal	9.16 (9.23)	5.36 (5.76)	6.34 (8.80)	**2.31 (1.75)**	10.4 (9.50)	5.45 (5.52)	6.87 (9.17)	**3.23 (4.56)**
Log-AMR	9.91 (5.97)	8.42 (5.29)	6.75 (3.42)	4.31 (2.30)	16.5 (6.57)	8.80 (5.46)	7.33 (3.69)	4.95 (2.83)
Log-AMR-optimal	8.64 (5.31)	6.72 (4.55)	5.64 (2.61)	3.36 (1.55)	9.30 (5.33)	6.78 (4.58)	5.89 (2.70)	3.73 (1.81)
Hausdorff	5.43 (2.71)				6.39 (2.88)			
EMD	2.80 (1.83)				3.50 (2.26)			
OSPA	1.86 (1.64)				2.41 (1.90)			

REFERENCES

[1] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[2] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *European Conf. on Computer Vision*, 2014.

[3] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Conf. on Computer Vision and Pattern Recognition*, 2016.

[4] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," *arXiv e-prints*, p. arXiv:1504.01942, 2015.

[5] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A Benchmark fors Multi-Object Tracking," *arXiv e-prints*, p. arXiv:1603.00831, 2016.

[6] P. Dendorfer *et al.*, "CVPR19 tracking and detection challenge: How crowded can it get?" *arXiv e-prints*, p. arXiv:1906.04567, 2019.

[7] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI Vision Benchmark Suite," in *Conf. on Computer Vision and Pattern Recognition*, 2012.

[8] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conf. on Computer Vision*, 2016.

[9] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *Journal on Image and Video Processing*, vol. 2008, p. 1, 2008.

[10] K. Smith, D. Gatica-Perez, J. Odobez, and Sileye Ba, "Evaluating multi-object tracking," in *Conf. on Computer Vision and Pattern Recognition*, 2005.

[11] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, 2015.

[12] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Conf. on Computer Vision and Pattern Recognition*, 2019.

[13] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, 2011.

[14] H. Rezatofighi *et al.*, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Conf. on Computer Vision and Pattern Recognition*, 2019.

[15] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybridboosted multi-target tracker for crowded scene," in *Conf. on Computer Vision and Pattern Recognition*, 2009.

[16] H. Ben Shitrit, J. Berclaz, F. Fleuret, and P. Fua, "Tracking multiple people under global appearance constraints," in *Int. Conf. on Computer Vision*, 2011.

[17] J. Luiten *et al.*, "HOTA: A higher order metric for evaluating multi-object tracking," *Int. Journal of Computer Vision*, 2020.

[18] L. Leal-Taixé *et al.*, "Tracking the trackers: An analysis of the state of the art in multiple object tracking," *arXiv e-prints*, p. 1704.02781, 2017.

[19] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[20] L. Maier-Hein *et al.*, "Why rankings of biomedical image analysis competitions should be interpreted with care," *Nature Communications*, vol. 9, no. 1, pp. 1–13, 2018.

[21] Z. Liu *et al.*, "Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: A comparative study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 94–109, 2012.

[22] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, "A comparative study of image retargeting," *ACM Trans. Graph.*, vol. 29, no. 6, Dec. 2010.

[23] C.-C. Hsu, C.-W. Lin, Y. Fang, and W. Lin, "Objective quality assessment for image retargeting based on perceptual geometric distortion and information loss," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 3, pp. 377–389, 2014.

[24] L. Ma, L. Xu, Y. Zhang, Y. Yan, and K. N. Ngan, "No-reference retargeted image quality assessment based on pairwise rank learning," *IEEE Trans. Multimedia*, vol. 18, no. 11, pp. 2228–2237, 2016.

[25] H. Barrow, J. Tenenbaum, R. Bolles, and H. Wolf, "Parametric correspondence and chamfer matching: two new techniques for image matching," in *Int. Joint Conf. on Articial Intelligence*, 1977.

[26] B. Grunbaum, *Convex polytopes*. Interscience, 1967.

[27] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Int. Conf. on Computer Vision*, 1998.

[28] R. L. Dobrushin, "Prescribing a system of random variables by conditional distributions," *Theory of Probability & Its Applications*, vol. 15, no. 3, pp. 458–486, 1970.

[29] J. R. Hoffman and R. P. S. Mahler, "Multitarget miss distance via optimal assignment," *IEEE Trans. Syst., Man, Cybern. A*, vol. 34, no. 3, pp. 327–336, 2004.

[30] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3447–3457, 2008.

[31] K. Oksuz and A. T. Cemgil, "Multitarget tracking performance metric: deficiency aware subpattern assignment," *IET Radar, Sonar & Navigation*, vol. 12, no. 3, pp. 373–381, 2018.

[32] K. Oksuz, B. Cam, E. Akbas, and S. Kalkan, "Localization recall precision (lrp): A new performance metric for object detection," in *European Conf. on Computer Vision*, 2018.

[33] K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas, "One metric to measure them all: Localisation recall precision (lrp) for evaluating visual detection tasks," *arXiv e-prints*, p. arXiv:2011.10772, 2021.

[34] M. Beard, B.-T. Vo, and B.-N. Vo, "A solution for large-scale multi-object tracking," *IEEE Trans. Signal Process.*, vol. 68, pp. 2754–2769, 2020.

[35] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[36] W. Liu *et al.*, "SSD: Single shot multibox detector," *Lecture Notes in Computer Science*, pp. 21–37, 2016.

[37] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Conf. on Neural Information Processing Systems*, 2016.

[38] C. Szegedy *et al.*, "Going deeper with convolutions," in *Conf. on Computer Vision and Pattern Recognition*, 2015.

[39] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Conf. on Computer Vision and Pattern Recognition*, 2016.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Conf. on Computer Vision and Pattern Recognition*, 2016.

[41] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI Conf. on Artificial Intelligence*, 2017.

[42] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2018.

[43] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," in *Int. Conf. on Learning Representations*, 2017.

[44] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv e-prints*, p. arXiv:1704.04861, 2017.

[45] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[46] T. Lin *et al.*, "Feature pyramid networks for object detection," in *Conf. on Computer Vision and Pattern Recognition*, 2017.

[47] P. Jin, V. Rathod, and X. Zhu, "Pooling pyramid network for object detection," *arXiv e-prints*, p. arXiv:1807.03284, 2018.

[48] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Int. Conf. on Computer Vision*, 2017.

[49] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Conf. on Computer Vision and Pattern Recognition*, 2017.

[50] H. Shen, L. Huang, C. Huang, and W. Xu, "Tracklet association tracker: An end-to-end learning-based association approach for multi-object tracking," *arXiv e-prints*, p. arXiv:1808.01562, 2018.

[51] S. Lee, M. Kim, and S. Bae, "Learning discriminative appearance models for online multi-object tracking with appearance discriminability measures," *IEEE Access*, vol. 6, pp. 67 316–67 328, 2018.

[52] R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro, "Online multi-target tracking with strong and weak detections," in *European Conf. on Computer Vision*, 2016.

[53] J. Chen, H. Sheng, Y. Zhang, and Z. Xiong, "Enhancing detection model for multiple hypothesis tracking," in *Conf. on Computer Vision and Pattern Recognition*, 2017.

[54] H. Sheng, Y. Zhang, J. Chen, Z. Xiong, and J. Zhang, "Heterogeneous association graph fusion for target association in multiple object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 11, pp. 3269–3280, 2019.

[55] P. Chu and H. Ling, "FAMNet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking," in *Int. Conf. on Computer Vision*, 2019.

[56] S. Lee and E. Kim, "Multiple object tracking via feature pyramid siamese networks," *IEEE Access*, vol. 7, pp. 8181–8194, 2019.
[57] R. Henschel, L. Leal-Taixé, D. Cremers, and B. Rosenhahn, "Fusion of head and full-body detectors for multi-object tracking," in *Conf. on Computer Vision and Pattern Recognition*, 2018.
[58] V. Eiselein, D. Arp, M. Patzold, and T. Sikora, "Real-time multi-human tracking using a probability hypothesis density filter and multiple detectors," in *Int. Conf. on Advanced Video Signal-based Surveillance*, 2012.
[59] T. Kutschbach, E. Bochinski, V. Eiselein, and T. Sikora, "Sequential sensor fusion combining probability hypothesis density and kernelized correlation filters for multi-object tracking in video data," in *Int. Conf. on Advanced Video Signal-based Surveillance*, 2017.
[60] N. L. Baisa and A. Wallace, "Development of a N-type GM-PHD filter for multiple target, multiple type visual tracking," *Journal of Visual Communication and Image Representation*, vol. 59, pp. 257 – 271, 2019.
[61] Y. Song and M. Jeon, "Online multiple object tracking with the hierarchically adopted GM-PHD filter using motion and appearance," in *Int. Conf. on Consumer Electronics-Asia*, 2016.
[62] Y. Yoon, A. Boragule, Y. Song, K. Yoon, and M. Jeon, "Online multi-object tracking with historical appearance matching and scene adaptive detection filtering," in *Int. Conf. on Advanced Video Signal-based Surveillance*, 2018.
[63] E. Bochinski, V. Eiselein, and T. Sikora, "High-speed tracking-by-detection without using image information," in *Int. Conf. on Advanced Video Signal-based Surveillance*, 2017.
[64] R. Henschel, Y. Zou, and B. Rosenhahn, "Multiple people tracking using body and joint detections," in *Conf. on Computer Vision and Pattern Recognition*, 2019.
[65] M. Keuper, S. Tang, B. Andres, T. Brox, and B. Schiele, "Motion segmentation multiple object tracking by correlation co-clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 140–153, 2020.
[66] C. Kim, F. Li, and J. M. Rehg, "Multi-object tracking with neural gating using bilinear LSTM," in *European Conf. on Computer Vision*, 2018.
[67] L. Chen, H. Ai, Z. Zhuang, and C. Shang, "Real-time multiple people tracking with deeply learned candidate selection and person re-identification," in *Int. Conf. on Multimedia and Expo*, 2018.
[68] R. Sanchez-Matilla and A. Cavallaro, "A predictor of moving objects for first-person vision," in *Int. Conf. on Image Processing*, 2019.
[69] H. Sheng *et al.*, "Iterative multiple hypothesis tracking with tracklet-level association," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3660–3672, 2019.
[70] P. Bergmann, T. Meinhardt, and L. Leal-Taixé, "Tracking without bells and whistles," in *Int. Conf. on Computer Vision*, 2019.