How Trustworthy are Performance Evaluations for Basic Vision Tasks?

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

Abstract—This article 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 a 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

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.

- First, 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].

- Second, 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.

- Third, 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, and 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 defined, respectively, as

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

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 $r_1, \ldots, r_N$ such that $r_n < r_{n+1}, \forall n < N$, the approximate AP score is:

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

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

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.
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

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

where \( f_1, \ldots, f_N \) are the sampled FPPI rates.

Performance evaluations for multi-object tracking consider the similarity/dissimilarity 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 \( \alpha \),

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

where:

\[
A(c) = \frac{|TPA(c)|}{|TPA(c)| + |FNA(c)| + |FP(c)|};
\]

\( TP, FN, \) 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 \( FP(c) \) 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 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?

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
First Second 3 etc.

Fers from a number of drawbacks. First, human evaluation meaningful humans to evaluate whether the performance criteria are their meaningfulness when applied to real data.

The better the criteria fare, and the more extensive the test scenarios, the more trust we have in its ability to provide meaningful evaluation. On the other hand, the bet-ter 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.

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 exercise. 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. First, 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. Second, 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 systemati-cally 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.

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 perturba-tions. 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-

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.

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 one 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.
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.

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 perturbation (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, false objects, 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. 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.

---

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.
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. $1 - \text{F1}$ and $1 - \text{mAP}$ (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 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.

One way to ensure mathematical consistency is to consider (mathematical) metrics—dissimilarity measures with certain mathematical properties. Specifically, a function $d: S \times S \to [0, \infty]$ is called a metric (or distance function) on the space $S$, if for all $x, y, z \in S$ 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, $d(\{x\}, \{y\}) = D(1 - \theta)$, where $D$ is a function such that: $D(0) = 1$ (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 $d(\{x\}, \{x + 0.6\}) = D(0) > 0$, but $d(\{x\}, \{x + 0.6\}) + d(\{x\}, \{x + 0.6\}) = D(1) + D(1) = 0$. It also violates the Identity property because $\{x\} \neq \{x + 0.6\}$ but $d(\{x\}, \{x + 0.6\}) = D(1) = 0$. 

Fig. 5. For an IoU threshold of 0.5, the Prediction is “closest” to the Approximate truth ($F_1 = 1$), which is “closest” to Ground truth ($F_1 = 1$). Thus, the Prediction should be “close” to Ground truth, but it is as “far” as possible from Ground truth ($F_1 = 0$).

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).
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
\[
\text{IoU}(x, y) = \frac{|x \cap y|}{|x \cup y|} \in [0,1],
\]
where \( |\cdot| \) denotes hyper-volume.

For convex shapes, the Generalized IoU index is given by
\[
\text{GIoU}(x, y) = \text{IoU}(x, y) - \frac{C(x \cup y \setminus (x \cup y))}{C(x \cup y)},
\]
where \( C(x \cup y) \) is the convex hull of \( x \cup y \). Note that unlike \( \text{IoU}(x, y) \), \( \text{GIoU}(x, y) \in [-1, 1] \). For arbitrary shapes, the definition of GIoU is given in the supplementary section [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
\[
d_{\text{IoU}}(x, y) = \frac{1}{|C(x \cap y)|} \quad \text{and} \quad d_{\text{GIoU}}(x, y) = \frac{1}{|C(x \cap y)|^{1/2}},
\]
which are indeed metrics bounded by 1.

For convex shapes, the Generalized IoU index is given by
\[
\text{GIoU}(x, y) = \text{IoU}(x, y) - \frac{C(x \cup y \setminus (x \cup y))}{C(x \cup y)},
\]
where \( C(x \cup y) \) is the convex hull of \( x \cup y \). Note that unlike \( \text{IoU}(x, y) \), \( \text{GIoU}(x, y) \in [-1, 1] \). For arbitrary shapes, the definition of GIoU is given in the supplementary section [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
\[
d_{\text{IoU}}(x, y) = \frac{1}{|C(x \cap y)|} \quad \text{and} \quad d_{\text{GIoU}}(x, y) = \frac{1}{|C(x \cap y)|^{1/2}},
\]
which are indeed metrics bounded by 1.

4.2 Metrics for Sets of Shapes

Our interest is the distance between two point patterns (or finite subsets) of a metric space \( (W, d) \), where \( d : W \times W \to [0,1] \) denotes the base-distance between the elements of \( W \). Specifically, \( W \) is the space of arbitrary/convex shapes and the base-distance \( d \) 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 \( W \) is defined by [26, 29]
\[
d_H(X, Y) = \max_{x \in X} \min_{y \in Y} d(x, y), \quad \max_{y \in Y} \min_{x \in X} d(x, y).
\]

For convex shapes, the Generalized IoU index is given by
\[
\text{GIoU}(x, y) = \text{IoU}(x, y) - \frac{C(x \cup y \setminus (x \cup y))}{C(x \cup y)},
\]
where \( C(x \cup y) \) is the convex hull of \( x \cup y \). Note that unlike \( \text{IoU}(x, y) \), \( \text{GIoU}(x, y) \in [-1, 1] \). For arbitrary shapes, the definition of GIoU is given in the supplementary section [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
\[
d_{\text{IoU}}(x, y) = \frac{1}{|C(x \cap y)|} \quad \text{and} \quad d_{\text{GIoU}}(x, y) = \frac{1}{|C(x \cap y)|^{1/2}},
\]
which are indeed metrics bounded by 1.

\[
\text{GIoU}(x, y) = \text{IoU}(x, y) - \frac{C(x \cup y \setminus (x \cup y))}{C(x \cup y)},
\]
where \( C(x \cup y) \) is the convex hull of \( x \cup y \). Note that unlike \( \text{IoU}(x, y) \), \( \text{GIoU}(x, y) \in [-1, 1] \). For arbitrary shapes, the definition of GIoU is given in the supplementary section [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
\[
d_{\text{IoU}}(x, y) = \frac{1}{|C(x \cap y)|} \quad \text{and} \quad d_{\text{GIoU}}(x, y) = \frac{1}{|C(x \cap y)|^{1/2}},
\]
which are indeed metrics bounded by 1.

For convex shapes, the Generalized IoU index is given by
\[
\text{GIoU}(x, y) = \text{IoU}(x, y) - \frac{C(x \cup y \setminus (x \cup y))}{C(x \cup y)},
\]
where \( C(x \cup y) \) is the convex hull of \( x \cup y \). Note that unlike \( \text{IoU}(x, y) \), \( \text{GIoU}(x, y) \in [-1, 1] \). For arbitrary shapes, the definition of GIoU is given in the supplementary section [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
\[
d_{\text{IoU}}(x, y) = \frac{1}{|C(x \cap y)|} \quad \text{and} \quad d_{\text{GIoU}}(x, y) = \frac{1}{|C(x \cap y)|^{1/2}},
\]
which are indeed metrics bounded by 1.
This metric was traditionally used as a measure of dissimilarity between binary images. It gives a good indication of the dissimilarity 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 \( X = \{x_1, \ldots, x_m\} \) and \( Y = \{y_1, \ldots, y_n\} \) is defined by [28], [29]

\[
d_q(X, Y) = \min_C \left( \sum_{i=1}^{m} \sum_{j=1}^{n} c_{i,j} d(x_i, y_j)^p \right)^{\frac{1}{p}},
\]

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 \( \ell_p \)-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 \( d \) is bounded by 1 (as per the IoU/GIoU distance), this problem can be resolved (while observing the metric properties) by defining \( d_X(X, Y) = d_Q(X, Y) = 1 \) if one of the set is empty, and \( d_X(\emptyset, \emptyset) = d_Q(\emptyset, \emptyset) = 0 \).

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

\[
d_0(X, Y) = \frac{1}{n} \left( \sum_{x_i \neq \emptyset} \min_{y_j \neq \emptyset} \sum_{x_i} d(x_i, y_j) + (n - m) \right),
\]

if \( n \geq m > 0 \), where \( H_n \) is the set of all permutations of \( \{1, 2, \ldots, n\} \), additionally: \( d_0(X, Y) = d_0(Y, X) \), if \( m > n > 0 \); \( d_0(X, Y) = 1 \), if one of the set is empty; and \( d_0(\emptyset, \emptyset) = 0 \). 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 supplementary materials Section 3.2, which can be found on the Computer Society Digital Library at online available) 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 \((\mathbb{W}, d)\) and discrete-time window \(T\), is defined as a mapping \(f: \mathbb{T} \mapsto \mathbb{W}\) [34]. Its domain \(D_f \subseteq T\) 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 \(g\)

\[
\overline{d}(f, g) = \sum_{t \in D_f \cup D_g} \frac{d_0(\{f(t)\}, \{g(t)\})}{|D_f \cup D_g|},
\]

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

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

Remark: The Hausdorff, EMD, and OSPA metrics (with both base-distances \( d \) and \( \bar{d} \)) 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 supplementary materials, available online.

In contrast to the inconsistencies of various criteria shown in Fig. 10, 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, Table 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 un-normalized 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 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 \( 2^{0.5 \times k} \) 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 \( F_{1,\text{IoU}} \) prediction error (1 – \( F_{1,\text{IoU}} \)), \( \text{OSPA}_{\text{IoU}} \), un-normalized \( \text{OSPA}_{\text{IoU}} \), \( \text{EMD}_{\text{IoU}} \), and \( \text{Hausdorff}_{\text{IoU}} \) distances for each scenario. Note that the \( \text{EMD}_{\text{IoU}}, \text{Hausdorff}_{\text{IoU}} \) and \( \text{OSPA}_{\text{IoU}} \) distances exhibit identical behavior that corroborate with physical intuition as they decrease with better performance. The \( F_{1,\text{IoU}} \) 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 \( \text{OSPA}_{\text{IoU}} \) metric produces non-sensical prediction error that increases drastically with unequivocally better performance.

2. This distance takes same the form as Eq. (3) but without the normalizing constant \( 1/n \).
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 supplementary materials, available online.

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,
objects with incorrect classes and the lower the detection confidence scores for objects with correct class. In the multiclass 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, \ldots, 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.

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. Table 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

![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.](image)

The subscripts of IoU/GIoU indicate the threshold values; “optimal” threshold is the one with best ranking accuracy; “M-full” 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).

| TABLE 2 |

| Monte Carlo Means (and Standard Deviations) of Normalized Kendall-Tau Ranking Errors of Various Criteria at Certain Thresholds |
| --- |
| **Single-Class Multi-Object Detection:** Normalized Kendall-tau ranking error (in units of 10^{-2}) |
| IoU_{\text{optimal}} | IoU_{\text{M-full}} | GIoU_{\text{M-full}} |
| --- | --- | --- |
| 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) | 9.39 (5.51) | 8.27 (8.71) | 4.86 (3.00) |
| Log-AMR | 9.91 (5.97) | 9.82 (5.29) | 6.75 (3.42) | 4.31 (2.30) | 16.5 (0.57) | 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) | 3.39 (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 | 12.0 (9.64) | 10.6 (5.27) |
| EMD | 3.53 (2.38) | 5.80 (3.29) |
| OSPA | 0.518 (0.580) | 0.539 (0.577) |

The subscripts of IoU/GIoU indicate the threshold values; “optimal” threshold is the one with best ranking accuracy; “M-full” 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).
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 Supplementary Materials Section 4, available online.

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 (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 supplementary materials (Section 5), available online. 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 Table 2).

In the detection and segmentation tasks, the difference between EMD and OSPA rankings is smaller than that

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.
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 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 GloU 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 Supplementary Materials Section 5, available online.

7 CONCLUSION

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.

ACKNOWLEDGMENTS

Tran Thien Dat Nguyen and Hamid Rezatofighi are equal contribution.

REFERENCES

[1] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (VOC) challenge,” Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, 2010.
[2] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 740–755.
[3] M. Cordts et al., “The cityscapes dataset for semantic urban scene understanding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 3213–3223.
[4] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, “MOTChallenge 2015: Towards a benchmark for multi-target tracking,” 2015, arXiv:1504.01942.
[5] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, “MOT16: A Benchmark for Multi-Object Tracking,” 2016, arXiv:1603.00831.
[6] P. Dendorfer et al., “CVPR19 tracking and detection challenge: How crowded can it get?,” 2019, arXiv:1906.04567.
[7] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The KITTI vision benchmark suite,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2012, pp. 3354–3361.
[8] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-target, multicamera tracking,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 17–35.
K. Bernardin and R. Stiefelhagen, “Evaluating multiple object tracking performance: The CLEAR MOT metrics,” J. Image Video Process., vol. 2008, 2008, Art. no. 1.

K. Smith, D. Gatica-Perez, J. Odobez, and S. Ba, “Evaluating multi-object tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2005, pp. 36–43.

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, Dec. 2015.

A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, “Panoptic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 9396–9405.

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, Apr. 2012.

H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 658–666.

Y. Li, C. Huang, and K. Nevatia, “Learning to associate: Hybridized multi-target tracker for crowded scene,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 2953–2960.

H. B. Shitrit, J. Berczal, F. Fleuret, and P. Fua, “Tracking multiple people under global appearance constraints,” in Proc. IEEE Int. Conf. Comput. Vis., 2011, pp. 137–144.

J. Luiten et al., “HOTA: A higher order metric for evaluating multi-object tracking,” Int. J. Comput. Vis., vol. 129, pp. 3447–3457, Aug. 2008.

C. Szegedy et al., “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1–9.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 2818–2826.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

K. Oksuz, B. Cam, and E. Akbas, “Inception-v4, inception-ResNet and the impact of residual connections on learning,” in Proc. AAAI Conf. Artif. Intell., 2017, pp. 4278–4284.

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, Apr. 2018.

B. Wu, K. Xu, and Q. V. Le, “A benchmark,” in Proc. Int. Conf. Learn. Representations, 2017, pp. 1–16.

A. G. Howard et al., “MobileNets: Efficient convolutional neural networks for mobile vision applications,” 2017, arXiv:1704.04861.

M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 4510–4520.

T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 936–944.

P. Jin, V. Rathod, and X. Zhu, “Pooling pyramid network for object detection,” 2018, arXiv:1807.03284.

K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2980–2988.

S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5987–5995.

H. Shen, L. Huang, C. Huang, and W. Xu, “Tracklet association tracker: An end-to-end learning-based association approach for multi-object tracking,” 2018, arXiv:1808.01562.

S. Lee, M. Kim, and S. Bae, “Learning discriminative appearance models for online multi-object tracking with appearance discrimination,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 627–632.

R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro, “Online multi-target tracking with strong and weak detections,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 84–99.

J. Chen, H. Sheng, Y. Zhang, and Z. Xiong, “Enhancing detection model for multiple hypothesis tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2143–2152.

H. Sheng, S. I. Zhao, Z. 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, Nov. 2019.

P. Chu and H. Ling, “FAMNet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 6171–6180.

S. Lee and E. Kim, “Multiple object tracking via feature pyramid siamese networks,” IEEE Access, vol. 7, pp. 8181–8194, 2019.

R. Henschel, L. Leal-Taixé, D. Cremers, and B. Rosenhahn, “Fusion of head and full-body detectors for multi-object tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 1509–150909.

V. Eiselein, D. Arp, M. Patzold, and T. Sikora, “Real-time multi-human tracking using a probability hypothesis density filter and multiple detector hypotheses,” in Proc. IEEE Int. Conf. Adv. Video Signal-Based Surveillance, 2012, pp. 325–330.

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 Proc. IEEE 14th Int. Conf. Adv. Video Signal-Based Surveillance, 2017, pp. 1–5.
Ba-Ngu Vo (Fellow, IEEE) received the bachelor’s (first class honors) degrees in mathematics and electrical engineering, in 1994, and the PhD degree, in 1997. Currently, he is professor of Signals and Systems at Curtin University. He is best known as a pioneer in the stochastic geometric approach to multi-object system. His research interests include signal processing, systems theory, and stochastic geometry.

Ba-Tuong Vo received the BSc degree in applied mathematics and the BE (with first-class honors) degree in electrical and electronic engineering, in 2004, and the PhD (with distinction) degree in engineering in 2008, all from the University of Western Australia. He is currently a professor of Signal Processing at Curtin University and has primary research interests in random sets, filtering and estimation, multiple object systems.

Silvio Savarese received the PhD degree in electrical engineering from the California Institute of Technology, in 2005. He is an associate professor of computer science with Stanford University and the inaugural Mindtree faculty scholar. He was a Beckman Institute fellow with the University of Illinois at Urbana-Champaign from 2005-2008. He joined Stanford in 2013 after being assistant and then associate professor of electrical and computer engineering with the University of Michigan, Ann Arbor, from 2008 to 2013. His research interests include computer vision, robotic perception, and machine learning. He is recipient of several awards including a Best Student Paper Award at CVPR 2016, the James R. Croes Medal in 2013, a TRW Automotive Endowed Research Award in 2012, an NSF Career Award in 2011 and Google Research Award in 2010. In 2002, he was awarded the Walker von Brimer Award for outstanding research initiative.

Ian Reid received the BSc (first class honours) degree in computer science and mathematics from the University of Western Australia, in 1987, and the DPhil degree from the University of Oxford, in 1992. He is the head with the School of Computer Science, University of Adelaide, and the senior researcher with the Australian Institute for Machine Learning. He is a fellow of the Academy of Technological Sciences and Engineering and held an ARC Australian Laureate Fellowship there 2013-2018. He is also deputy director of the ARC Centre of Excellence in Robotic Vision, in which capacity he also co-chaired sub-committees of the National Roadmap for Robotics produced and published by the Centre in 2018. Between 2000 and 2012, he was a professor of engineering science with the University of Oxford, where, together with long-time colleague Prof. David Murray, he ran the Active Vision Group which is part of the wider Robotics Research Group. He was awarded a Rhodes Scholarship in 1988 to study at the University of Oxford.

Tran Thien Dat Nguyen received the BE degree in electrical power engineering, in 2018, and the PhD degree in electrical and computer engineering, in 2022. He is currently a post-doctoral researcher with the Intelligent Sensing & Perception Laboratory (ISPL), Curtin University. His research interests include multi-object systems, Bayesian filtering, and random sets.

Hamid Rezatofighi is a lecturer with the Faculty of Information Technology, Monash University, Australia. Before that, he was an endeavour research fellow with the Stanford Vision Lab (SVL), Stanford University and a senior research fellow with the Australian Institute for Machine Learning (AIML), University of Adelaide. His main research interest focuses on computer vision and vision-based perception for robotics, including object detection, multi-object tracking, human trajectory and pose forecasting and human collective activity recognition. He has also research expertise in Bayesian filtering, estimation and learning using point process and finite set statistics.

For more information on this or any other computing topic, please visit our Digital Library at [www.computer.org/csdl](http://www.computer.org/csdl).