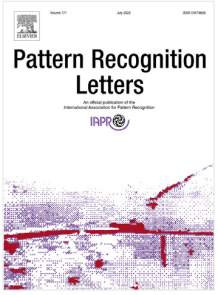
Journal Pre-proof



A sharper definition of alignment for Panoptic Quality

Ruben van Heusden, Maarten Marx

PII: S0167-8655(24)00208-3
DOI: https://doi.org/10.1016/j.patrec.2024.07.005
Reference: PATREC 9224

To appear in: *Pattern Recognition Letters*

Received date : 21 December 2023
Revised date : 30 June 2024
Accepted date : 3 July 2024

Please cite this article as: R. van Heusden and M. Marx, A sharper definition of alignment for Panoptic Quality, *Pattern Recognition Letters* (2024), doi: https://doi.org/10.1016/j.patrec.2024.07.005.

ELSEVIER

Pattern Recognition Letters
journal homepage: www.elsevier.com

A sharper definition of alignment for Panoptic Quality

Ruben van Heusden[a,**], Maarten Marx[a]

[a] *Information Retrieval Lab, University of Amsterdam, Science Park 900, Amsterdam, 1098XH, The Netherlands*

ABSTRACT

The Panoptic Quality metric, developed by Kirillov et al. in 2019, makes object-level precision, recall and F1 measures available for evaluating image segmentation, and more generally any partitioning task, against a gold standard. Panoptic Quality is based on partial isomorphisms between hypothesized and true segmentations. Kirillov et al. desire that functions defining these one-to-one matchings should be simple, interpretable and effectively computable. They show that for t and h, true and hypothesized segments, the condition stating that there are more correct than wrongly predicted pixels, formalized as $IoU(t,h) > .5$ or equivalently as $|t \cap h| > .5|t \cup h|$ has these properties. We show that a weaker function, requiring that more than half of the pixels in the hypothesized segment are in the true segment and vice-versa, formalized as $|t \cap h| > .5|t|$ and $|t \cap h| > .5|h|$, is not only sufficient but also necessary. With a small proviso, every function defining a partial isomorphism satisfies this condition. We theoretically and empirically compare the two conditions.

1. Introduction

Kirillov et al. [14] have developed *Panoptic Quality (PQ)*, a metric that can be used to evaluate image segmentation methods by comparing predicted and true segmentations. PQ is specifically developed for segmentation problems in which *exact matches* are unfeasible and not even needed for successful applications. Although originally developed for the image domain, PQ can also be used for text, and even for any partitioning problem. The only requirement is that there is an underlying set of elements (in images the pixels, in text segmentation typically tokens) which are *partially partitioned* (i.e., elements are combined into non overlapping segments, but not all elements need to be assigned to a segment).

PQ makes the well known F1 measure (the harmonic mean between precision and recall) available for the partial match segmentation setting. This is done by generalizing the definition of True Positives from a set of items to a set of *pairs of matched items*. If this set is a partial bijection (for every predicted segment h there is at most one true segment t and vice-versa), the false positives and false negatives and thus all contingency table based metrics are also defined.

Kirillov et al. suggest to match a predicted segment h to a true segment t iff $IoU(h,t) > .5$, where IoU denotes the intersection-over-union operation, defined as $\frac{t \cap h}{t \cup h}$ [1]. They show that this condition guarantees that the resulting matching is a partial bijection. This condition is simple, effectively computable and interpretable, as desired by Kirillov et al. However, even though it is very natural (as it requires that there are strictly more overlapping than missed and spurious pixels) one could ask for a more foundational reason to choose this threshold. This led to the following research question.

> Are there other useful [2], interpretable, simple and effective matching definitions which imply the partial bijection property? And if so, is there a most general one?

Indeed there is a strictly weaker, most general and thus sufficient and necessary condition. Let h and t be two segments (thus subsets) of the same set. Now $IoU(h,t) > .5$ is equivalent

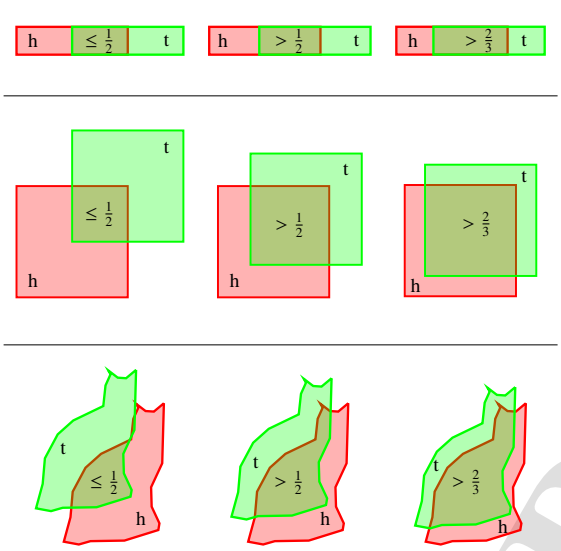
**Corresponding author at: Information Retrieval Lab, University of Amsterdam, Science Park 900, 1098 XH Amsterdam, The Netherlands.
e-mail: r.j.vanheusden@uva.nl (Ruben van Heusden), maartenmarx@uva.nl (Maarten Marx)

[1] For the definition of IoU used in the paper we assume both operands are 2D objects, unless stated otherwise.

[2] We added the property *useful*, because the identity matching satisfies all other properties, but obviously this is often too strict and thus not (very) useful.

Table 1: Examples for three objects of matches with too little overlap for both θ^+ and $\theta^\&$ (left column), only enough overlap for $\theta^\&$ (center column) and enough overlap for both θ^+ and $\theta^\&$ (right column). Green indicates the ground truth object , red indicates the predicted object, ✓ means the matching satisfies the condition and ✗ means it fails the condition.

Not enough overlap	Necessary and sufficient overlap	Sufficient overlap
$\theta^\&$ ✗, θ^+ ✗	$\theta^\&$ ✓, θ^+ ✗	$\theta^\&$ ✓, θ^+ ✓



to $|h \cap t| > |h \oplus t|$, where $\oplus$ denotes the symmetric difference between h and t. In turn this is equivalent to

$$|h \cap t| > |h \setminus t| + |t \setminus h|. \qquad (\theta^+)$$

The weaker most general matching definition distributes this and requires with each conjunct implying the injectivity of one side of the matching.

$$|h \cap t| > |h \setminus t| \ \textbf{and} \ |h \cap t| > |t \setminus h|, \qquad (\theta^\&)$$

Our main result is that every "fair" matching is a partial bijection if and only if it satisfies $(\theta^\&)$. Below we will develop what "fair" means in this context. Clearly this alternative condition is also simple, natural and interpretable. Table 1 shows the difference between $\theta^\&$ and θ^+.

As a sidenote, non-trivial alternative matching rules exist, such as the one introduced by Chen et al. [3], as an alternative to the θ^+ matching. Instead of enforcing a single threshold that ensures a one-to-one mapping between predictions and ground truth, a mapping is created by using the Hungarian algorithm to obtain an one-on-one mapping without enforcing a single threshold. However, this method still requires an IoU threshold to allow for false positives and false negatives, and is arguably more involved than the θ^+ or $\theta^\&$ matchings.

The rest of the paper is structured as follows. In the next section we prove this result, and compare the two ways of matching

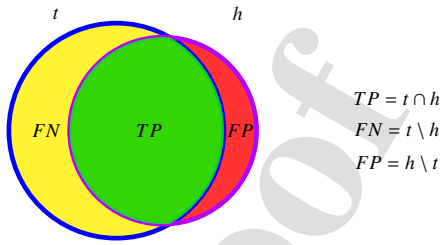


Fig. 1: Illustration of a ground truth object t and a predicted object h, outlined by blue and purple respectively. The green area represents the overlap, yellow represents the area only in t and red represents the area only in h. TP, FN and FP can then be expressed as set operations on t and h.

true and predicted objects. We then empirically look at the differences between the two versions of PQ, and finish with related work.

But before we dive into the technicalities, let us compare the two ways of matching from another perspective. The matching defined by $(\theta^\&)$ can equivalently be stated as

$$\frac{|h \cap t|}{|h|} > .5 \ \textbf{and} \ \frac{|h \cap t|}{|t|} > .5, \qquad (\theta^\&)$$

simply stating that the overlap of the two segments should cover more than half of the segments. All objects pairings in the center column of Table 1 match by this criterion. They do however not satisfy $IoU(h,t) > .5$, and thus do not match by criterion (θ^+). In fact, when the two segments have the same size ($|h| = |t|$), then $IoU(h,t) > .5$ is equivalent to asking that the overlap covers more than two-third of the segments, like in the top right image in Table 1.

This can be derived from the formula for IoU, where we let $|t \cap h|$ be equal to a, and $|t| = |h| = b$.

$$\frac{a}{a + (b-a) + (b-a)} > \frac{1}{2}$$

Through simple algebraic steps this is equal to requiring $a > \frac{2}{3}b$. We believe this shows that the weaker matching condition is at least as natural as $IoU(h,t) > .5$, but less arbitrary.

2. Theoretical results

We first recall the definitions of segmentation, recognition and panoptic quality. Then we show the following basic results about $(\theta^\&)$ and its relation with (θ^+):

- $(\theta^\&)$ is effectively computable, and defines a partial bijection.
- $(\theta^\&)$ is strictly weaker than (θ^+).
- Precision, recall, recognition quality and panoptic quality defined with $(\theta^\&)$ are all larger than or equal to the metrics defined with (θ^+).
- The threshold of .5 is optimal when defining a matching using IoU.

Then in section 2.1 we prove our main result.

Let E be a set (in images this would be a set of pixels), and both H and T partial partitions of E. Thus both H and T consist of pairwise disjoint subsets of E. We call such H and T segmentations and often ignore the underlying set. Let $b \subseteq H \times T$ be a partial bijection. We will use $dom(b)$ and $ran(b)$ to denote the domain and range of b respectively. We often view b as a (partial) function from H into T, and b^{-1} as one from T into H. We call the triple H, T, b an *alignment*.

We recall the definition of the Panoptic Quality metric from [14]. It is given relative to an alignment. Let

$$\begin{aligned} TP &= b \\ FP &= H \setminus dom(b) \\ FN &= T \setminus ran(b). \end{aligned}$$

Figure 1 contains an example and the way we present these three different parts in this paper. Now the *recognition quality* RQ is simply the harmonic mean between precision and recall, known as the F1 measure:

$$RQ = \frac{|TP|}{|TP| + .5(|FP| + |FN|)}.$$

The Segmentation Quality (SQ) is the mean IoU of the True Positives, and the Panoptic Quality PQ then is the product of SQ and RQ. In a direct definition the relation with $F1$ is even closer. Let wTP denote $\Sigma\{IoU(h,t) \mid (h,t) \in TP\}$. Then

$$PQ = \frac{wTP}{|TP| + .5(|FP| + |FN|)}.$$

This is the reason why one can refer to PQ as a weighted version of F1. Similarly, one can define weighted and unweighted versions of precision and recall by dividing wTP or $|TP|$ by $|TP| + |FP|$ for precision and $|TP| + |FN|$ for recall, respectively. Here the metrics are defined relative to one alignment, thus to one example. The PQ, RQ and SQ of a set of examples is simply the mean PQ, RQ and SQ over them, respectively.

Theorem 1. *Let H and T be segmentations of the same set and let $B = \{(h,t) \in H \times T \mid (h,t)$ satisfies $(\theta^{\&})\}$. Then B is effectively computable, a partial bijection, and for each $(h,t) \in B$, $IoU(h,t) > \frac{1}{3}$.*

Proof. A trivial nested for loop over H and T finds the alignment defined by $(\theta^{\&})$. This can be optimized using the order on the elements of the underlying domain. That for each $(h,t) \in B$, $IoU(h,t) > \frac{1}{3}$, follows from the fact that $|t \cup h|$ is equal to $|t \cap h| + |t \setminus h| + |h \setminus t|$. $(\theta^{\&})$ states that $|t \cap h|$ is strictly larger than these last two summands.

To show that B is a partial bijection, suppose to the contrary that it is not. We will derive a contradiction. There are two possibilities. We treat one: there is a $t \in T$ and two different (and thus disjoint) $h_1, h_2 \in H$ satisfying $(\theta^{\&})$. Thus by assumption, both $\frac{|h_1 \cap t|}{|t|} > .5$ and $\frac{|h_2 \cap t|}{|t|} > .5$. From $\frac{|h_1 \cap t|}{|t|} > .5$ it follows that $|t \setminus h_1| \leq .5|t|$. Because h_1 and h_2 are disjoint, $h_2 \cap t \subseteq t \setminus h_1$, and thus by transitivity, $|h_2 \cap t| \leq .5|t|$, which contradicts our assumption. □

We now establish the relation between the two ways of matching. We have seen that both (θ^{+}) and $(\theta^{\&})$ define alignments, and thus give alternative definitions of RQ, SQ and PQ. Given two segmentations H and T of the same set, we speak about their θ^{+}- and $\theta^{\&}$-alignment, and distinguish the corresponding metrics using the same superscripts.

Theorem 2. *$(\theta^{\&})$ is strictly weaker than (θ^{+}). That is, every θ^{+}-alignment is a $\theta^{\&}$-alignment, and there are segmentations H, T with an $\theta^{\&}$-alignment but no θ^{+}-alignment.*

Proof. Because $|h \setminus t|$ and $|t \setminus h|$ are disjoint, (θ^{+}) implies $(\theta^{\&})$. For strictness let $H = \{\{1\}, \{2,3,4\}\}$ and $T = \{\{1,2,3\}, \{4\}\}$. Then with $(\theta^{\&})$, $\{2,3,4\}$ is aligned to $\{1,2,3\}$ because it has 2 elements in the overlap, and it has one missing and one spurious element. But the IoU of these two segments is equal to $\frac{2}{4}$ and thus not strictly larger than .5, and thus the alignment defined by (θ^{+}) is empty. □

We now investigate what happens to the three panoptic quality metrics when they are defined using $(\theta^{\&})$ and (θ^{+}).

Theorem 3. *Consider the θ^{+}- and the $\theta^{\&}$-alignment of the same two segmentations H and T of the same underlying set. Then*

$$\begin{aligned} SQ^{\&} &\leq SQ^{+} \\ P^{\&} &\geq P^{+} \\ R^{\&} &\geq R^{+} \\ RQ^{\&} &\geq RQ^{+} \\ PQ^{\&} &\geq PQ^{+}. \end{aligned}$$

Proof. By the previous theorem, $b^{+} \subseteq b^{\&}$, and thus the number of True Positives remains the same or grows with $(\theta^{\&})$, since the alignment condition $\theta^{\&}$ is less strict, and as shown in the example in Theorem 2, allows pairings with an IoU of less than .5 to be considered true positives. The extra true positives have an IoU between $\frac{1}{3}$ and $\frac{1}{2}$, and so these extra TPs will bring the mean IoU, which is SQ, down. Every extra True Positive reduces both the number of FPs and FNs by one. So extra TPs lead to a higher numerator but an equal denominator in the definitions of precision, recall and RQ (which is after all F1) and thus it will go up with the weaker matching condition $\theta^{\&}$. The numerator in the definition of PQ is the sum of all IoU of all TPs. So that also goes up when the number of TPs increases, and thus also $PQ^{\&} \geq PQ^{+}$. Indeed, the same holds for the weighted versions of Precision and Recall. □

We will now establish that the IoU threshold of 0.5 in θ^{+} is optimal in the sense that any lower value would not guarantee a partial bijection for all images. We will show this in a more general setting. Both IoU and RQ are instances of a general schema, known as the Tversky index [20]. Given a true and a predicted object t and h and the corresponding TP, FP and FN as defined in the beginning of this section, the Tversky index S is defined as follows.

$$S_{\alpha,\beta}(t,h) = \frac{|TP|}{|TP| + \alpha|FP| + \beta|FN|}, \text{ where } \alpha, \beta \geq 0.$$

The RQ (or $F1$) score corresponds to $\alpha = \beta = .5$, and the IoU to $\alpha = \beta = 1$.

Theorem 4. *Let B denote $\{(h,t) \in H \times T \mid S_{\alpha,\beta}(h,t) > \gamma\}$, for some α, β, γ all between 0 and 1 and H and T segmentations of the same set. Then the following are equivalent.*

1. *B is a partial bijection for all segmentations H and T;*
2. *both $\frac{\gamma}{1-\gamma}\alpha$ and $\frac{\gamma}{1-\gamma}\beta$ are larger than or equal to 1.*

The theorem immediately implies that B defined by $IoU(h,t) > \gamma$ is guaranteed to be a partial bijection for all segmentations if and only if $\gamma \geq .5$.

Proof. Let B be defined as stated in the theorem. A little algebra shows that $(t,h) \in B$ if and only if

$$|TP| > \frac{\gamma}{1-\gamma}\alpha|FP| + \frac{\gamma}{1-\gamma}\beta|FN|. \tag{1}$$

We start with the easy direction. Assuming both $\frac{\gamma}{1-\gamma}\alpha$ and $\frac{\gamma}{1-\gamma}\beta$ are at least 1, (1) implies that $|TP| > |FP|$ and $|TP| > |FN|$ and thus $\theta^{\&}(h,t)$ holds and by Theorem 1, B is always a partial bijection.

We prove the other direction by contraposition. Suppose $\frac{\gamma}{1-\gamma}\alpha < 1$. The case for β is shown similarly. We abbreviate $\frac{\gamma}{1-\gamma}\alpha$ by w for ease of notation. Define two total partitions H and T of a set E where $H = \{h\}$ and $T = \{t_1, t_2\}$, satisfying $|t_1| > |t_2| > w|t_1|$. This is possible as $w < 1$. We prove that both (h,t_1) and (h,t_2) are in B and thus B is not a partial bijection. Now $(h,t_1) \in B$ iff (1) holds. But we have

- $TP_h^{t_1} = t_1 \cap h = t_1$, as $t_1 \subseteq h$;
- $FP_h^{t_1} = h \setminus t_1 = t_2$, as $t_1 \cup t_2 = h$;
- $FN_h^{t_1} = t_1 \setminus h = \emptyset$, as $t_1 \subseteq h$.

Thus (1) reduces to $|t_1| > w|t_2|$, which holds because $w < 1$ and we have constructed t_1 and t_2 such that $|t_1| > |t_2|$. We can similarly show that $(h,t_2) \in B$ using the fact that $|t_2| > w|t_1|$. □

2.1. Every fair alignment satisfies ($\theta^{\&}$)

We will now prove our main result stating that every reasonable alignment is a partial bijection if and only if it satisfies ($\theta^{\&}$). We first develop what are reasonable (we call them "fair") alignments.

Definition 1. *Let H, T, b be an alignment.*

1. *We call $(h,t) \in H \times T$ a* mismatch *in H, T, b if $h \notin dom(b)$ but $|h \cap t| \geq |b^{-1}(t) \cap t|$.*
2. *We say that H, T, b is* not fair *if either H, T, b or T, H, b^{-1} contains a mismatch.*

Definition 2. 1. *The alignment H, T', b' is an* improvement *of the alignment H, T, b if $dom(b) = dom(b')$ and $h \cap b'(h) \supseteq h \cap b(h)$ holds for all $h \in dom(b)$.*
2. *H', T, b' is an* improvement *of H, T, b if $ran(b) = ran(b')$ and $t \cap b'^{-1}(t) \supseteq t \cap b^{-1}(t)$ holds for all $t \in ran(b)$.*

Definition 3. *We call an alignment H, T, b* fair *if every improvement of H, T, b is fair.*

Thus with an improvement, we may change one of the segmentations and either the range or the domain of the alignment b, but only if the IoU of each aligned pair remains the same or increases. Let's see an example:

T	1,2,3		4,5,6
H	1,2,3,4,5		6
H'	1,2,3	4,5	6

We have two alignments H, T, b and H', T, b', with b and b' as indicated in the mapping. H', T, b' is an improvement of H, T, b as the overlap of the matched pairs remains the same for both segments in $ran(b)$. The pair $(\{4,5\}, \{4,5,6\})$ is a mismatch of H', T, b', and thus H', T, b' is not fair. And thus is H, T, b not fair, because it has an unfair improvement.

Theorem 5. *Let H and T be two segmentations of the same set and $B \subseteq H \times T$. Then the following are equivalent:*

- *all $(h,t) \in B$ satisfy ($\theta^{\&}$);*
- *B is a partial bijection and H, T, B is a fair alignment.*

Proof ($\Downarrow$) Assume that all $(h,t) \in B$ satisfy ($\theta^{\&}$). By Claim 2, B is a partial bijection, so we will write B as the function b. Now suppose to the contrary that H, T, b is not a fair alignment. Then there is an improvement of H, T, b which is not fair. There are two cases. We do one, and let the improvement be H, T', b' with an $h \in H$ and a $t \in T'$ such that $|h \cap t| \geq |h \cap b'(h)|$. Because H, T', b' is an improvement, it holds that $|h \cap b'(h)| \geq |h \cap b(h)|$. Thus we have that $|h \cap t| \geq |h \cap b(h)|$.

Now t and $b'(h)$ are disjoint, so $h \setminus b'(h) \supseteq h \cap t$. Because b' is an improvement, $h \cap b'(h) \supseteq h \cap b(h)$ and thus $h \setminus b(h) \supseteq h \setminus b'(h)$, and by transitivity, $h \setminus b(h) \supseteq h \cap t$. Using $|h \cap t| \geq |h \cap b(h)|$ we obtain $|h \setminus b(h)| \geq |h \cap b(h)|$ which contradicts ($\theta^{\&}$).

($\Uparrow$) Assume H, T, b is a fair alignment and b a partial bijection. Suppose to the contrary that ($\theta^{\&}$) does not hold. Then one of the two conjuncts fails. Suppose the first. Thus there is an $h \in H$ such that $|h \setminus b(h)| \geq |h \cap b(h)|$. Let $z = h \setminus b(h)$. Now create H, T', b' as follows.

$$T' = \{z\} \cup \{t \in T \mid t \cap z = \emptyset\} \cup \{t \setminus z \mid t \in T \text{ and } t \cap z \neq \emptyset\},$$

and for all $h \in dom(b)$ set $b'(h) = b(h)$ if $b(h) \in T$, and $b(h) \setminus z$ otherwise.

We will show that H, T', b' is an improvement which is not fair because of h, our required contradiction. Because T and b have these properties, also T' is a partial partition, and b' a partial bijection. To show that H, T', b' is an improvement of H, T, b, we must show that for all $\bar{h} \in H$, the overlap with its match remained the same or increased, i.e., $\bar{h} \cap b'(\bar{h}) \supseteq \bar{h} \cap b(\bar{h})$. This holds by definition of b' when $b(\bar{h})$ and z do not overlap. Thus in particular for the segment h. If they do overlap, then as $z = h \setminus b(h) \subseteq h$, for all $\bar{h} \neq h$, the overlap will increase because the elements in z are disjoint from $\bar{h}$ and thus taking them out of $b(\bar{h})$ reduces the number of errors.

Now we show that H, T', b' is not fair for h, precisely because of the set $z \in T'$ which is not in $ran(b')$. By definition $z = h \setminus b(h)$ and thus $z = h \cap z$. By assumption on h, $|h \setminus b(h)| \geq |h \cap b(h)|$, and thus $|h \cap z| \geq |h \setminus b(h)|$. We found our desired contradiction.

3. Empirical Evaluation

In this section, θ^+ and $\theta^\&$ are compared on three instance segmentation benchmarks and their differences evaluated: How many additional True Positives does $\theta^\&$ yield? What is their IoU? Are certain classes or visual properties over-represented in the additional TPs? And are they indeed acceptable as correct predictions?

In the experiments, PQ is calculated over the prediction and ground truths sets for all three datasets, where only the matching condition is altered to be either θ^+ or $\theta^\&$. Recall that, since the $\theta^\&$ matching is less strict than the θ^+ matching, ground truth and predicted pairs with an IoU of less than .5 can be considered true positives under $\theta^\&$ but not under θ^+, and thus the number of true positive matchings under $\theta^\&$ will be larger or equal to that of θ^+.

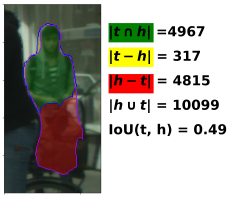
Figure 2 shows two examples of additional true positives yielded by $\theta^\&$. In the top image, the model makes a recall error and misses a substantial portion of the ground truth object, indicated by the large yellow region. In the bottom image, the model makes a precision error and erroneously predicts a large portion of the image to belong to the ground truth object, indicated by the red region. In both cases, the IoU equals .49, and thus not enough to be a TP according to θ^+.

For the comparison, Mask-R-CNN [12] was run on the CityScapes [6], MS COCO [17] and LVIS(version 0.5) [11] datasets. CityScapes consists of images from streets in several German cities, and MS COCO and LVIS both consist of Flickr images. As is common practice, we use the validation sets to evaluate the performance of the image detection method. It should be noted that CityScapes has much more objects per image annotated ($\mu = 20.4$) than the two other sets. We used the pre-trained Mask-R-CNN models made available by Meta through their Detectron2 library[23] for all three datasets. The code and data to reproduce the results of the empirical evaluation are publicly available on Github.[3] The main findings can be summarized as follows.

- Evaluating by $\theta^\&$ instead of θ^+ yielded 1,250 additional TPs on the three datasets in total. Per dataset, this meant one to two percent point higher recall.
- The IoU of the additional TPs is close to the 0.5 threshold of θ^+ ($\mu = .46$, $\sigma = .03$).
- The number of objects in an image and the additional number of true positives in an image are positively correlated, with a Pearson correlation of .35.
- additional TPs are heterogeneous in size, their visual properties and their missed and spurious pixels.
- Five percent of the additional TPs can be considered as being incorrectly classified as detected objects. The majority of these *false hits* are very small objects.

[3] `https://github.com/RubenvanHeusden/GeneralizedPanopticQuality`

(a) **Partial**: The predicted object only contains a part of the ground truth object.



(b) **Extra**: The predicted object is larger than the ground truth object, and the spurious pixels are not assigned to another object.

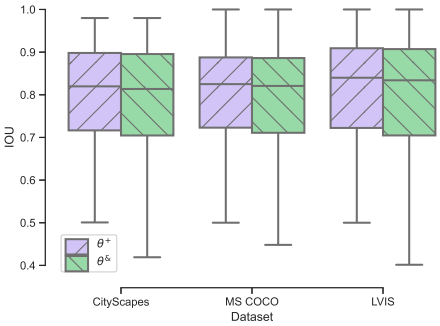
Fig. 2: Two examples of TPs according to $\theta^\&$, but not according to θ^+. The blue and purple contours refer to t and h, respectively. The green area indicates the intersection between t and h, red signifies pixels only present in h, and yellow signifies pixels only present in t. Cardinalities of sets denote number of pixels.

We first describe the distribution of the IoU values of the additional TPs. Table 2 shows that $\theta^\&$ yields 1,250 additional true positives counted over all three datasets, between 1 and 2 percent more true positives than θ^+ depending on the dataset. Figure 3a shows the distributions of the IoU for both θ^+ and $\theta^\&$ for all three datasets. For all three, the distributions differ significantly when measured using the Kolmogorov-Smirnov test ($D_{6944,7166} = .03, p = .001$, $D_{18957,19483} = .027, p < .001$ and $D_{14053,14545} = .03, p < .001$), for CityScapes, MS COCO and LVIS respectively, where the subscripts indicate the sample sizes of θ^+ and $\theta^\&$ respectively). Figure 3b shows the distributions of the IoU of the additional TPs. For all datasets, over 75% of these TPs has an IoU higher than .43.

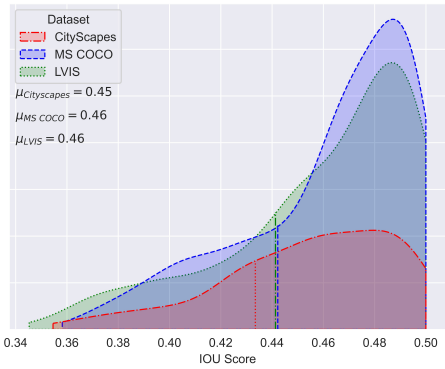
Table 2: Number of additional TPs and the fraction of the additional TPs within all TPs.

Dataset	Additional TPs	Fraction
CityScapes	232	.02
MS COCO	536	.01
LVIS	492	.01

We now investigate whether the true objects of the additional TPs have particular characteristics. The number of objects in an image and the additional number of true positives in an image are positively correlated ($r(9036) = .35, p < .001$)). In all three datasets, the objects are also assigned to classes. The distributions of the classes over all objects and over the additional TPs do not differ significantly for CityScapes and COCO when mea-



(a)



(b)

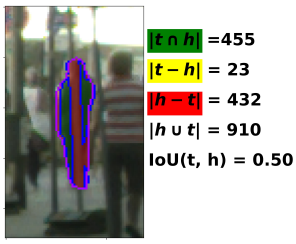
Fig. 3: (a): Boxplot of the distributions of the IoU of true positives under both θ^+ and $\theta^\&$ and (b): KDE plot of the IoU of the additional TPs with the vertical lines representing the first quartile.

sured using the Kolmogorov-Smirnov test ($D_{10} = .43, p = .37$, $D_{65} = .03, p = .07$. For the LVIS dataset there is a significant difference ($D_{149} = .23, p < .001$), which can be explained by the fact that the dataset has many, 830, fine-grained labels and that there are is only a small amount of additional true positives, so many classes with a few objects in the gold standard do not occur in the additional true positives.

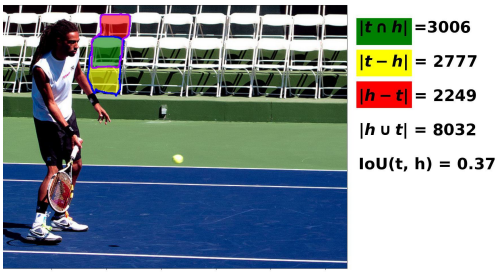
We manually inspected all 232 additional TPs in the CityScapes dataset and classified them into the best fitting error-types, shown in Figures 3 and 4 ($N_{partial} = 48$, $N_{extra} = 45$, $N_{occluded} = 24$, $N_{crowd} = 48$, $N_{small} = 45$, $N_{annotation\ error} = 12$).

The classification of error types is partly inspired by previous work by for example Bernhard et al. [2], with similar definitions for the *partial* and *extra* classes, with the work by Bernhard et al. also providing error categories for cases where the predicted and ground truth objects are disjoint. For all error types, the mean IoU is close to 0.5, meaning there is not one error type that accounts for most of the low IoU TPs.

We now investigate whether the additional TPs can really be considered True Positives. The $\theta^\&$ criterion can yield true positive pairs with an IoU of just over one-third, in which case the



(a) **Partially Occluded**: The ground truth object is partially occluded by another object.



(b) **Crowd**: The predicted object includes parts of other objects close to it.

Fig. 4: Examples of the *Partially occluded* and *Crowd* classes

overlap is only a little bit more than both the missed part and the wrongly assigned part, like in the middle column of Table 1. Figure 3b shows that in all three datasets there are TPs with an IoU just over a third. If this wrongly assigned part (the FP pixels) overlaps almost completely with another true object, we consider the TP a *false hit*. Figure 4b contains an example of such a false hit: the wrongly assigned yellow and red parts are almost as large as the green overlap *and* the red part is for a large part contained in another object. We make this notion of false hit precise as follows, using a parameter $0 < \pi < 1$ to capture "almost as large". A TP (t, h) is a *false hit* if there is a true object $t' \neq t$ such that $\pi \cdot |t \cap h| \leq |t' \cap h|$ and $\pi \cdot |t \cap h| \leq |t \setminus h|$. Note that the first conjunct also implies that $\pi \cdot |t \cap h| \leq |h \setminus t|$ because t and t' are disjoint. It is easy to see that only additional TPs can be false hits. We set $\pi = .75$. Only 61 of the 1,250 additional TPs are false hits, (0.5%, 5% and 6% for CityScapes, COCO and LVIS, respectively). These false hits tend to be small objects (within the COCO dataset an object is called small if it has less than 1024 pixels). For CityScapes, MS COCO and LVIS the median pixel size of the true object of the false hits is 359, 422 and 930 pixels, respectively. For CityScapes, MS COCO and LVIS, 100, 71 and 65 percent of the false hits are small objects. Foucart et al. [9] discuss the PQ metric for medical imaging for cell nuclei, and argue against the usage of PQ for small objects, as the small size of the objects means small perturbations in the predictions can have a large effect on whether or not a cell is considered a true positive. Since the majority of

the additional True Positives can be considered correct, we believe that using the $\theta^{\&}$ alignment condition thus yields a more accurate way of measuring model performance.

4. Related Work

Traditionally, the performance of image segmentation models has been evaluated using a variety of pixel- and object-level metrics, such as pixel-level precision and recall, and object-level metrics such as Average Precision and average IoU scores[18, 10, 6]. Both Average Precision and average IoU make usage of the *IoU* score to produce a mapping between predicted and ground truth objects. In particular, Average Precision is calculated by using the *IoU* threshold of .5 and defining precision and recall in a similar manner to PQ. Although these metrics are still widely reported, the emergence of the panoptic segmentation task has resulted in the more widespread adaptation of the PQ metric [15, 5, 12]. Examples include the MaskFormer and OneFormer architectures [5, 12], both of which use the Transformer architecture at the basis of the prediction model, and present unified frameworks for tackling semantic, instance and panoptic segmentation simultaneously. Li et al. provide a detailed overview of the usage of Transformer-based models in their survey paper [16], reporting their results using the PQ metric for panoptic segmentation datasets. Several image segmentation challenges and benchmarks have also adopted the Panoptic Quality metric as part of their evaluation setup, with examples in diverse fields such as the detection of different types of crops, detection of cell nuclei in the medical domain, the segmentation of 3D point clouds from LiDAR data and other domains such as for modelling attacks on network systems [22, 21, 8, 25, 7, 19]. The PQ metric has also been extended for usage in the video domain, where it is referred to as *Video Panoptic Quality (VPQ)* [13]. In this setting, IoU is used to obtain TP, FP and FN values over a collection of video frames, after which VPQ is calculated in the same way as the original PQ metric. The VPQ metric includes a hyperparameter k that determines how many frames are considered for the calculation of PQ. Cheng et al [4] propose a change to the calculation of the IoU metric to make it more sensitive to mistakes in object contours. Since the number of pixels around the edge of an object does not scale proportionally with the area of the object, IoU tends to under-penalize boundary mistakes in larger objects. To remedy this problem, they propose *Boundary-IoU*, which only considers pixels up to d pixels away from the object contour and is formalized as follows. Given a distance parameter d , ground truth object $t \in T$ and predicted object $h \in H$.

$$IoU_{Boundary} = \frac{t_d \cap h_d}{t_d \cup h_d} \tag{2}$$

Where t_d and t_h are the portions of the ground truth and predicted objects that are up to d pixels away (measured in Euclidean distance) from their respective object contours. The choice of the distance parameter d controls the sensitivity of the metric towards mistakes in the object boundaries: the lower d, the more the object contour decides whether a pair (t, h) is a true positive. Choosing the appropriate value for d is an important consideration, as setting the value of d too low can result in very small perturbations (for example stemming from contour ambiguity) having a large effect on the final score. Cheng et al. experimented with this by comparing the annotations from two annotators on the LVIS dataset, treating one as the gold standard and one as the prediction, varying the value of d and counting the number of TPs. They found that, for the LVIS dataset, setting d to roughly two percent of the image diagonal, and so d would be potentially different from image to image, and could possibly even depend on the ground object size, to avoid the aforementioned problem. For both the VPQ and Boundary-PQ variations, substituting the θ^{+} matching by the $\theta^{\&}$ matching still results in partial isomorphism, as both variants constrain the predicted objects to be non-overlapping.

5. Discussion & Future Work

Although the paper explores the effect of the altered alignment condition $\theta^{\&}$ on the PQ metric specifically, the fact that the matching concerns the IoU metric means that the altered condition can be used in any setting where IoU is used in determining the correctness of a prediction against a gold standard. Examples of this include the Average Precision and average IoU metrics discussed previously, as well as metrics in different fields, such as Named Entity Recognition and text segmentation. Depending on the specific application, the fact that the $\theta^{\&}$ matching is less strict than the original formulation means that in general more True Positives will be yielded with this matching.

If predicted and ground truth segments are particularly small (such as cell nuclei in medical images), the use of Panoptic Quality as a reliable evaluation metric can become problematic, as outlined by Foucart et al. [9]. As in the *Boundary-IoU* paper, one of the problems is that mistakes around the boundaries of an object are penalized more heavily for small objects compared to larger objects. Experiments on nuclei segmentation datasets in which artificial distortions are applied to the predictions (dilation by 1 pixel, erosion by 1 pixel and 1 pixel vertical shift), show these small perturbations can lead to a significant amount of predictions receiving an IoU of less than 0.5, in turn resulting in low PQ scores, that do not reflect the actual quality of the predictions when manually inspected. One of the causes of this behaviour is the fact that *IoU* inherently does not weight spurious and mixed pixels equally, with predictions with missed pixels being scored lower than spurious pixels errors of the same size. Although the proposed matching rule would alleviate this problem in some regards as the threshold is relaxed, the inherent problem is still present. Although this problem is usually solved in practice by discarding predictions below a certain size (dependent on the dataset)[24, 1]. Future work can be done in adapting the metric to use cases where small predictions are common. Although this work explores the implications of a new matching rule for Panoptic Quality in the image domain, the practical implication has only been measured for a handful of datasets in the image domain, and the difference between the two matching rules might be more pronounced for different fields such as usage in Named Entity Recognition, something that could also be explored in future work.

6. Conclusion

We found a useful, simple, interpretable and effectively computable definition for aligning true and predicted segments which is both a necessary and sufficient condition for the alignment being a partial bijection. If, given a predicted and true segment, we let TP, FP and FN stand for the pixels in the overlap, the spurious and the missed pixels, respectively, then the necessary condition aligns the two segments if |TP| > |FN| and |TP| > |FP|. This in contrast to the stronger $IoU > .5$ condition which is equivalent to |TP| > |FN|+|FP|. The effect of the weaker condition was small but not negligible; on three instance segmentation datasets it led to a 1-2% increase in recall. Our empirical analysis of these additional true positives shows that 95% of them are indeed valuable correctly identified objects. The few misses were mostly very small objects. As the new condition is the most general effectively computable alignment that guarantees a partial bijection, we recommend it will be used in future implementations of PQ.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported in part by the Netherlands Organization for Scientific Research (NWO) through the ACCESS project grant CISC.CC.016.

References

[1] Abbas, A., & Swoboda, P. (2021). Combinatorial optimization for panoptic segmentation: A fully differentiable approach. *Advances in Neural Information Processing Systems*, *34*, 15635–15649.

[2] Bernhard, M., Amoroso, R., Kindermann, Y., Baraldi, L., Cucchiara, R., Tresp, V., & Schubert, M. (2024). What's outside the intersection? fine-grained error analysis for semantic segmentation beyond iou. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 968–977).

[3] Chen, L., Wu, Y., Stegmaier, J., & Merhof, D. (2023). Sortedap: Rethinking evaluation metrics for instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 3923–3929).

[4] Cheng, B., Girshick, R., Dollár, P., Berg, A. C., & Kirillov, A. (2021). Boundary iou: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CCVPR)* (pp. 15334–15342).

[5] Cheng, B., Schwing, A., & Kirillov, A. (2021). Per-pixel classification is not all you need for semantic segmentation. *Adv. in Neural Inf. Syst.*, *34*, 17864–17875.

[6] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The Cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CCVPR)* (pp. 3213–3223).

[7] Du, Z., Xie, X., Qu, Z., Hu, Y., & Stojanovic, V. (2024). Dynamic event-triggered consensus control for interval type-2 fuzzy multi-agent systems. *IEEE Transactions on Circuits and Systems I: Regular Papers*, .

[8] Fong, W. K., Mohan, R., Hurtado, J. V., Zhou, L., Caesar, H., Beijbom, O., & Valada, A. (2022). Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *IEEE Robot. and Automation Lett.*, *7*, 3795–3802. URL: `https://doi.org/10.1109/LRA.2022.3148457`. doi:`10.1109/LRA.2022.3148457`.

[9] Foucart, A., Debeir, O., & Decaestecker, C. (2023). Panoptic quality should be avoided as a metric for assessing cell nuclei segmentation and classification in digital pathology. *Sci. Rep.*, *13*, 8614. URL: `https://doi.org/10.1038/s41598-023-35605-7`. doi:`10.1038/s41598-023-35605-7`.

[10] Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 1440–1448).

[11] Gupta, A., Dollar, P., & Girshick, R. (2019). LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CCVPR)* (pp. 5356–5364).

[12] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 2961–2969).

[13] Kim, D., Woo, S., Lee, J.-Y., & Kweon, I. S. (2020). Video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CCVPR)* (pp. 9859–9868).

[14] Kirillov, A., He, K., Girshick, R., Rother, C., & Dollár, P. (2019). Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CCVPR)* (pp. 9404–9413).

[15] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y. et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4015–4026).

[16] Li, X., Ding, H., Zhang, W., Yuan, H., Pang, J., Cheng, G., Chen, K., Liu, Z., & Loy, C. C. (2023). Transformer-based visual segmentation: A survey. *arXiv preprint arXiv:2304.09854*, . URL: `https://doi.org/10.48550/arXiv.2304.09854`. doi:`10.48550/arXiv.2304.09854`.

[17] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (pp. 740–755). Cham: Springer International Publishing. URL: `https://doi.org/10.1007/978-3-319-10602-1_48`. doi:`10.1007/978-3-319-10602-1_48`.

[18] Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CCVPR)* (pp. 8759–8768).

[19] Song, X., Song, Y., Stojanovic, V., & Song, S. (2023). Improved dynamic event-triggered security control for t–s fuzzy lpv-pde systems via pointwise measurements and point control. *International Journal of Fuzzy Systems*, *25*, 3177–3192.

[20] Tversky, A. (1977). Features of similarity. *Psychol. Rev.*, *84*, 327–352.

[21] Verma, R., Kumar, N., Patil, A., Kurian, N. C., Rane, S., & Sethi, A. (2020). Multi-organ nuclei segmentation and classification challenge 2020. *IEEE Trans. on Med. Imaging*, *39*, 3413–3423. URL: `https://doi.org/10.1109/TMI.2021.3085712`. doi:`10.1109/TMI.2021.3085712`.

[22] Weyler, J., Magistri, F., Marks, E., Chong, Y. L., Sodano, M., Roggiolani, G., Chebrolu, N., Stachniss, C., & Behley, J. (2023). Phenobench–a large dataset and benchmarks for semantic image interpretation in the agricultural domain. *arXiv preprint arXiv:2306.04557*, .

[23] Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). Detectron2. `https://github.com/facebookresearch/detectron2`.

[24] Xiong, Y., Liao, R., Zhao, H., Hu, R., Bai, M., Yumer, E., & Urtasun, R. (2019). Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CCVPR)* (pp. 8818–8826).

[25] Zhang, Z., Song, X., Sun, X., & Stojanovic, V. (2023). Hybrid-driven-based fuzzy secure filtering for nonlinear parabolic partial differential equation systems with cyber attacks. *International Journal of Adaptive Control and Signal Processing*, *37*, 380–398.

Research Highlights

- We present $\theta^{\&}$, a sharper definition of the object alignment in Panoptic Quality.
- We provide an extensive theoretical evaluation of both matchings.
- We empirically evaluate both matchings on three image segmentation datasets.

Confirmation of Authorship

Pattern Recognition Letters

Authorship Confirmation

Please save a copy of this file, complete and upload as the "Confirmation of Authorship" file.

As corresponding author I, Ruben van Heusden, hereby confirm on behalf of all authors that:

1. This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.
2. If presented at or submitted to or published at a conference(s), the conference(s) is (are) identified and substantial justification for re-publication is presented below. A copy of conference paper(s) is(are) uploaded with the manuscript.
3. If the manuscript appears as a preprint anywhere on the web, e.g. arXiv, etc., it is identified below. The preprint should include a statement that the paper is under consideration at Pattern Recognition Letters.
4. All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.
5. All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript.

Signature _____ Date 21-12-2023

List any pre-prints:
The most general manner to injectively align true and predicted segments `https://arxiv.org/abs/2212.13445`

Relevant Conference publication(s) (submitted, accepted, or published):

Justification for re-publication:

Declaration of Interest Statement

Declaration of interests

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: