Revisiting the COCO Panoptic Metric to Enable Visual and Qualitative Analysis of Historical Map Instance Segmentation

E. Carlinet and J. Chazalon
EPITA Research & Development Laboratory (LRDE)

Motivation

- Evaluate historical map (dense instance) segmentation (ICDAR21 MapSeg competition)
- Identify uses for major document segmentation metrics, and compare them.
- Propose a unifying framework for these metrics.
- Provide new qualitative and quantitative tools to leverage these metrics.

```
https://github.com/icdar21-mapseg/coco-pano-ext-demo
```

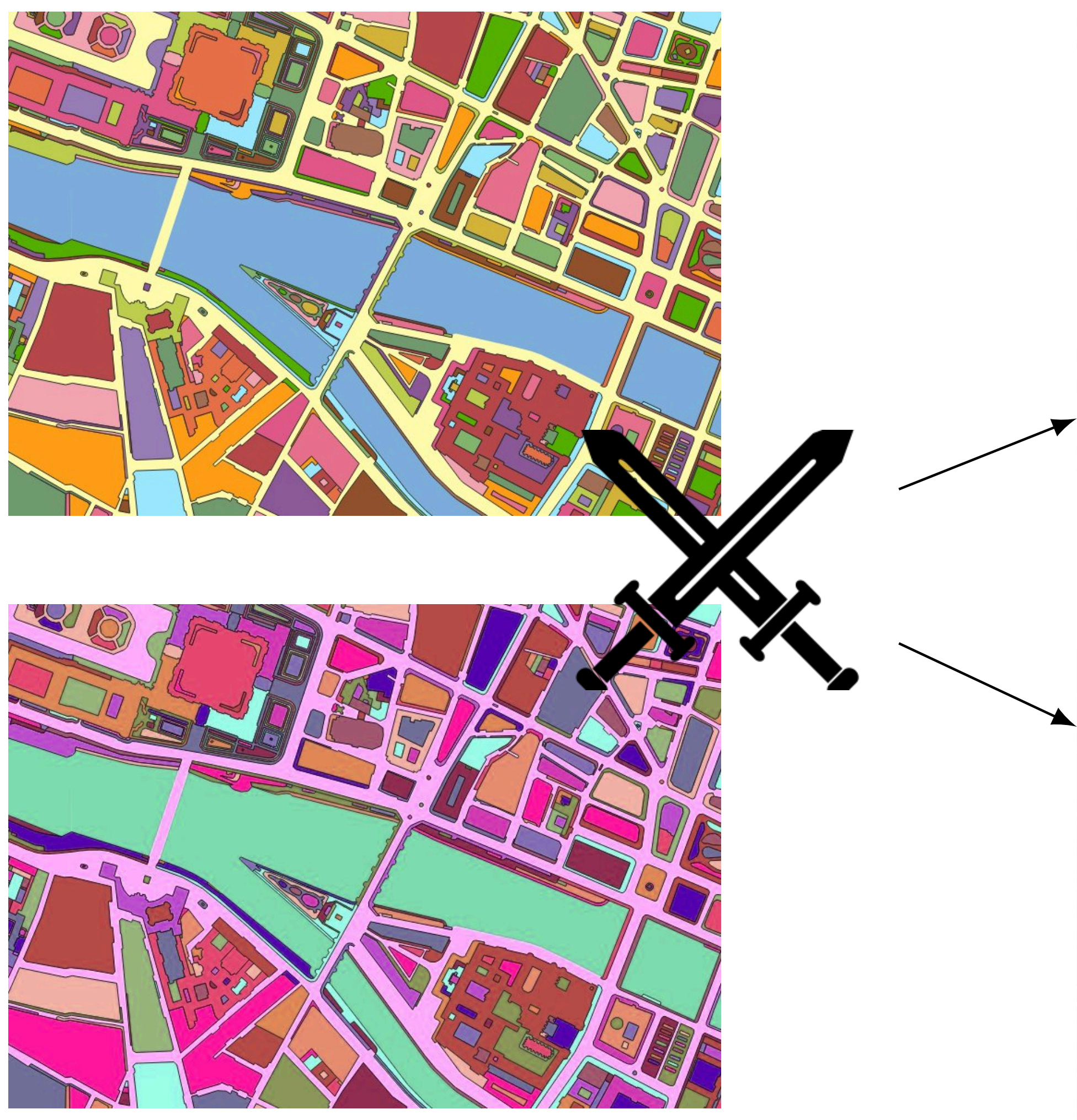
```
pip install coco-pano-ext-demo
```

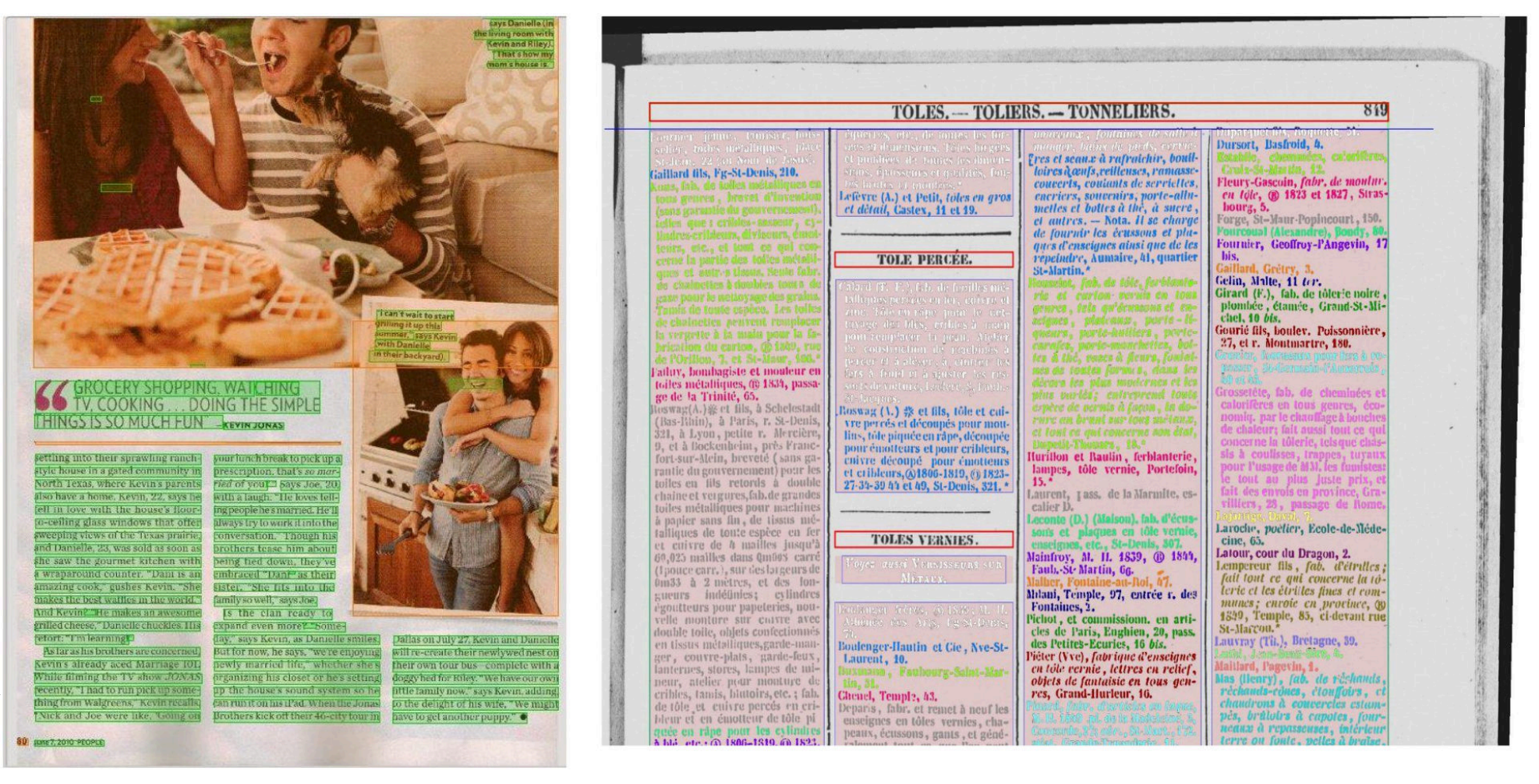
This work was partially funded by the French National Research Agency (ANR): Project SoDuCo, grant ANR-18-CE38-0013. Thanks to the City of Paris for granting us with the permission to use and reproduce the atlases used in this work.

Segmentation Metrics

What makes a metric bad ?

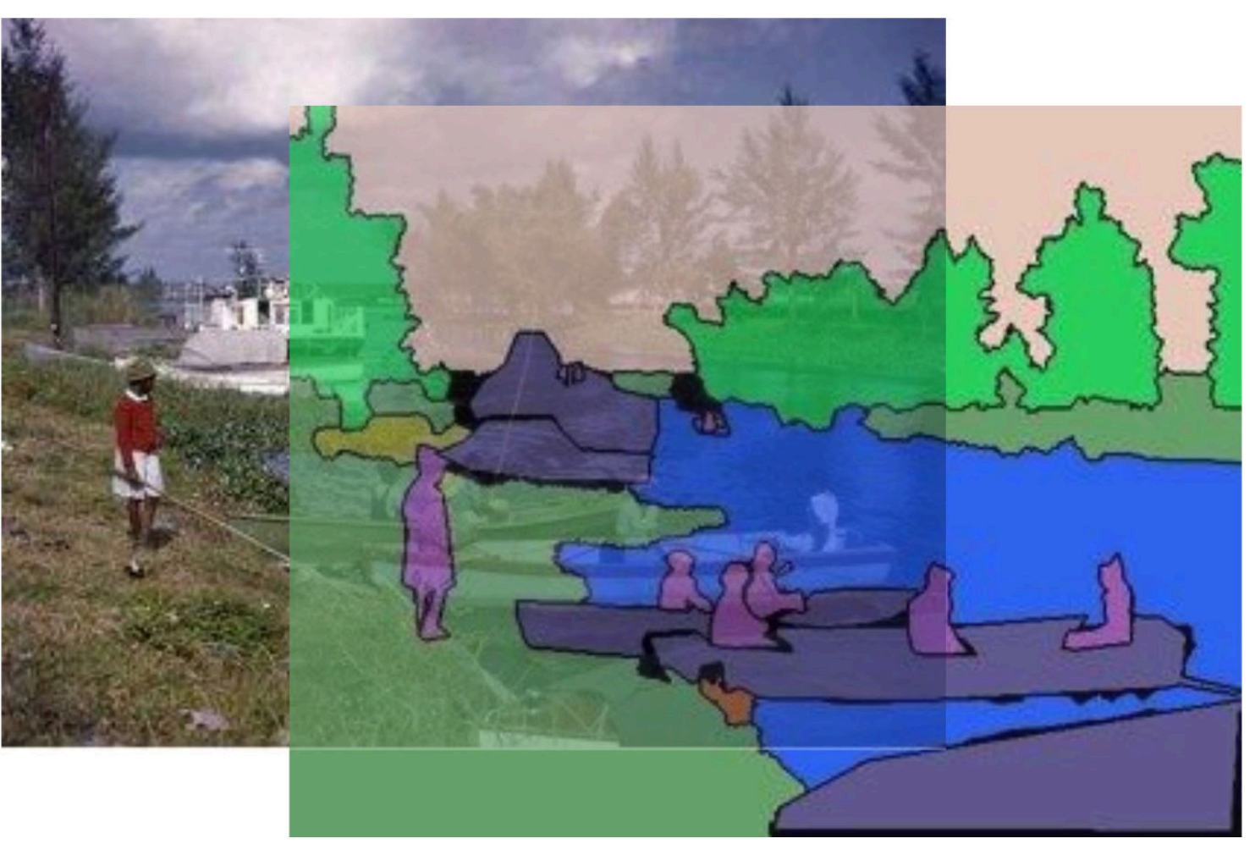
1. Not normalized/with no bounds (*"Congrats you've got 641! All contenders are between 111 and 4747"*)
2. Does not fit the human perception of "good" (*"The method B totally messed up and got the score 0.951"*)
3. Many thresholds/parameters (*"for our experiments, we use $\alpha_1 = 0.51$ and $\zeta_{26} = 3.29$"*)
4. Hard to explain (*"because a* log *makes it look smarter"*)





Document-oriented: DetEval, ZoneMap

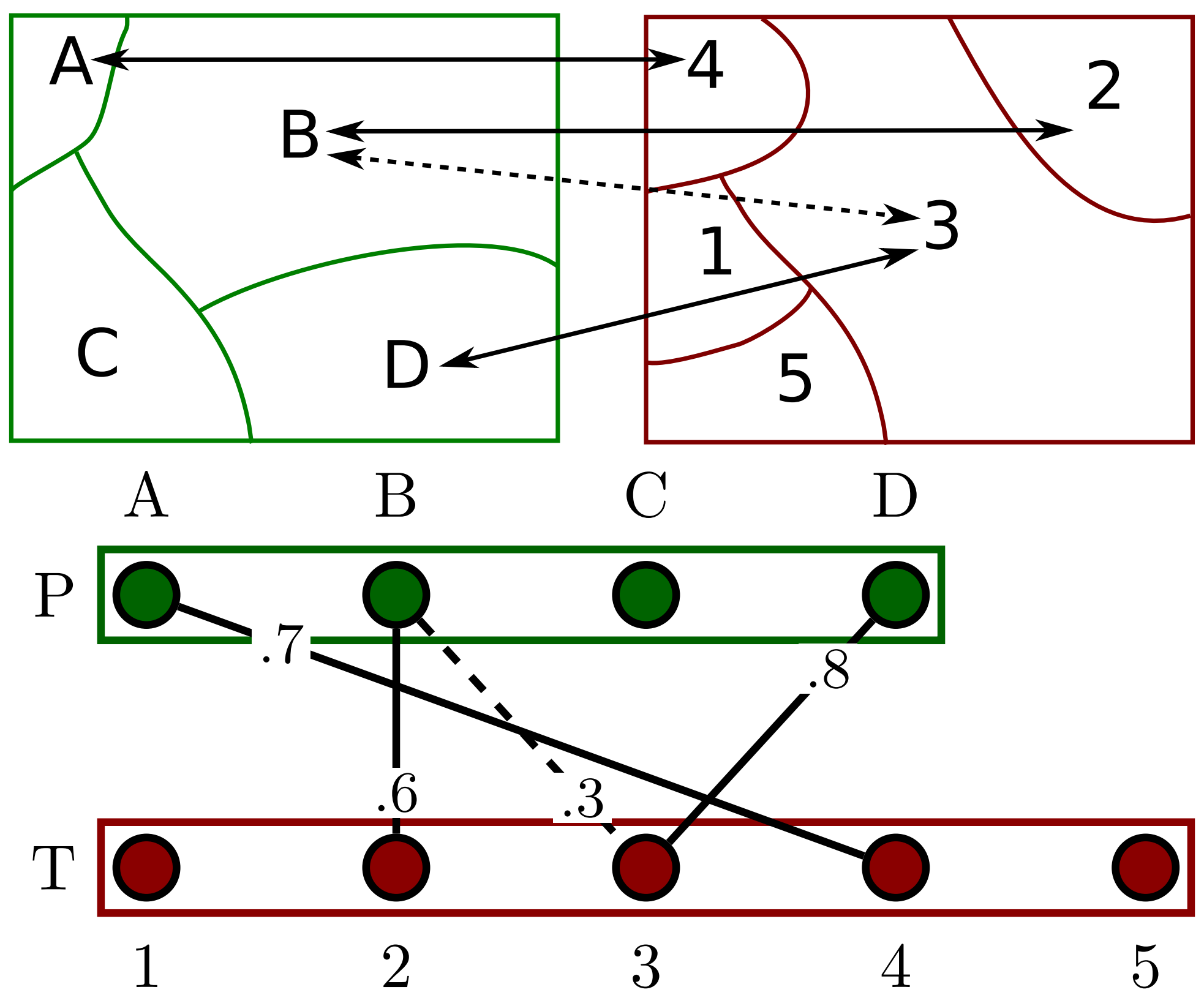
- ✓ Consider fragmentation (1-to-many relations)
- ✗ Under/over-segmentation is somehow over-scoring
- ✗ ZoneMap: Non-normalized scores
- ✗ DetEval: calibration of several thresholds

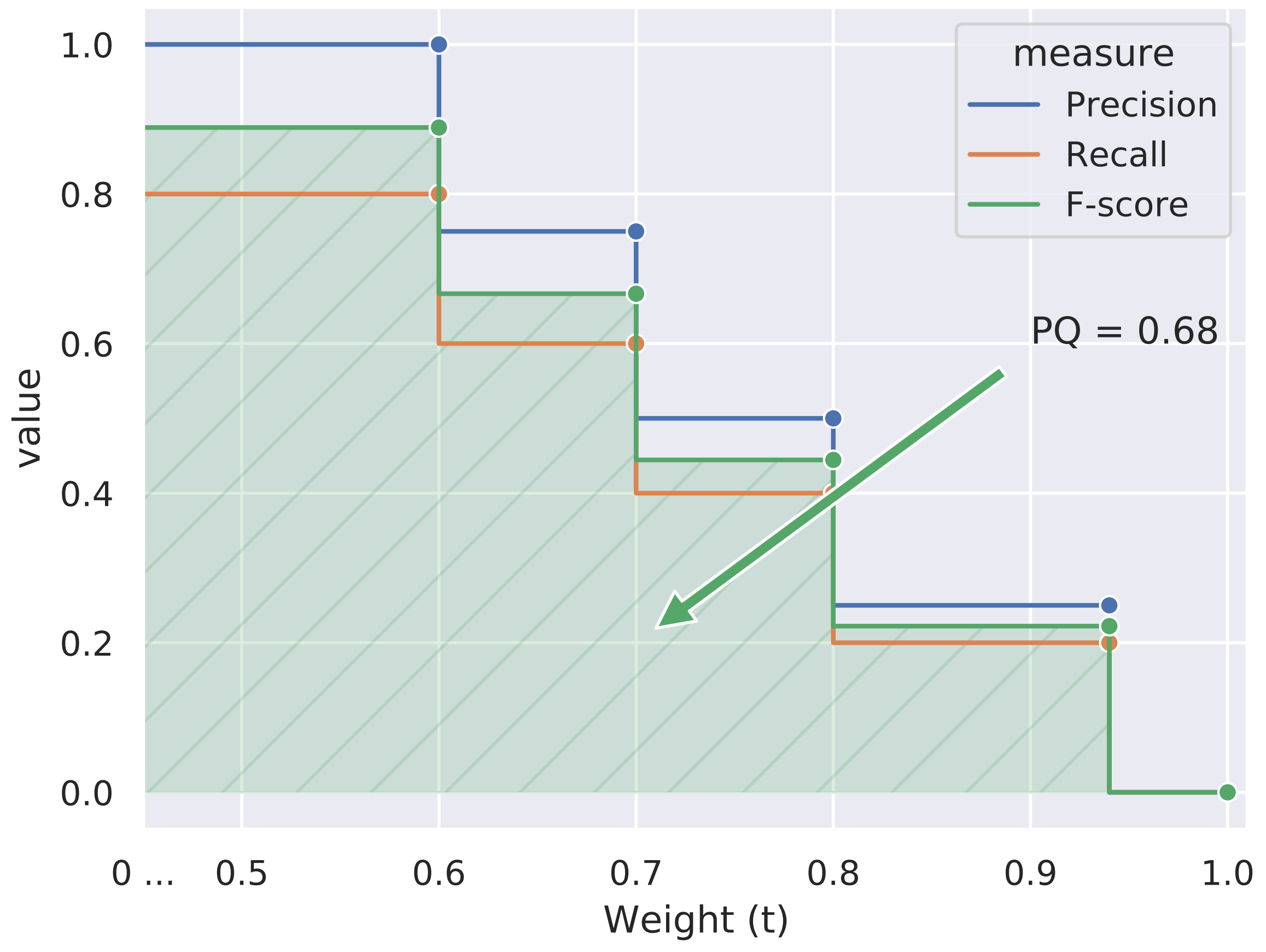


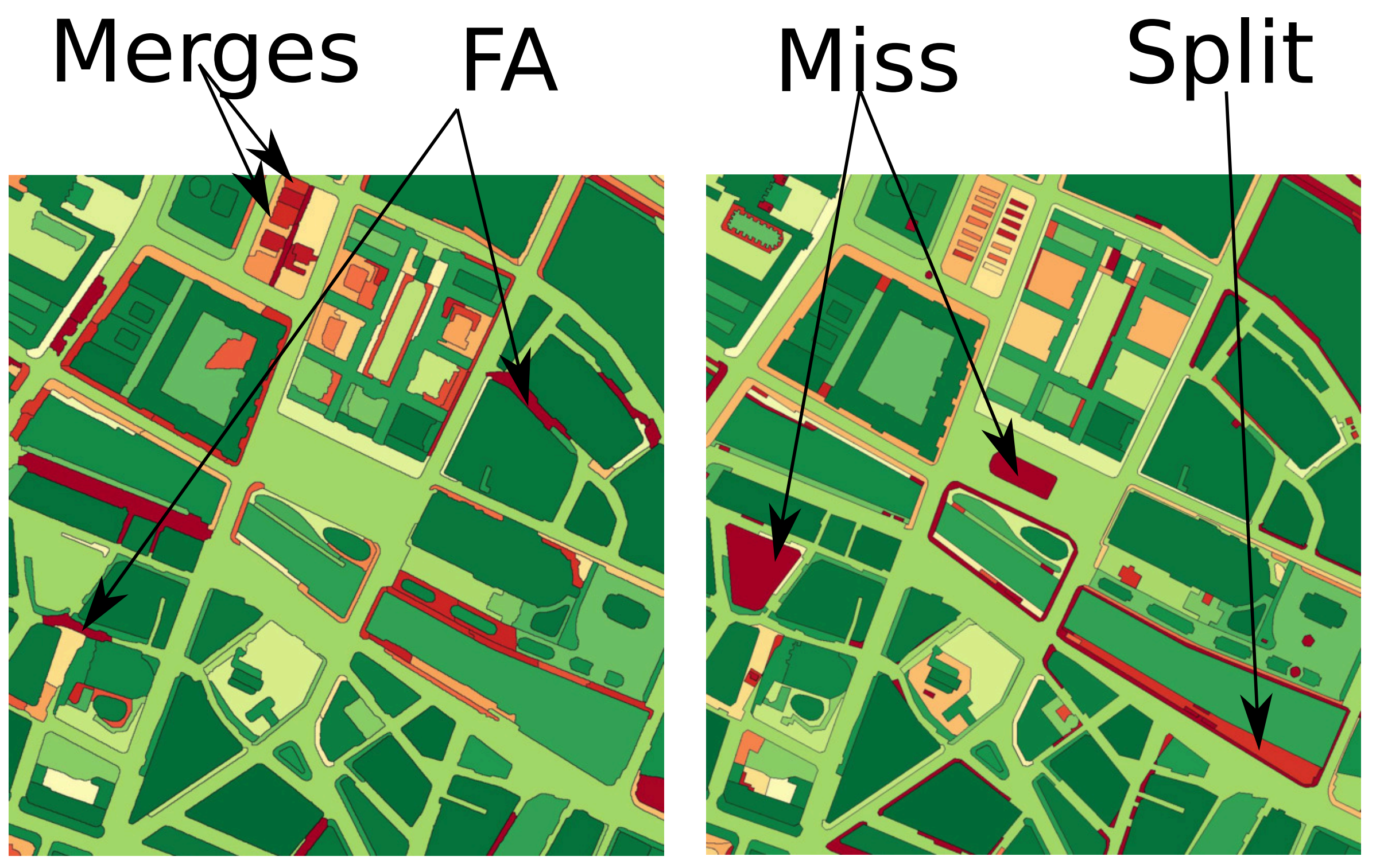
Computer Vision-related: COCO Panoptic

- ✓ Simple: a combination of "segmentation accuracy" and "detection performance"
- ✓ A single parameter, easy to tune and sensible (the IoU min)
- ✗ Single global score (ok for ranking but bad for a system performance analysis)

A framework for weighted pairing evaluation







↑ Precision Map Recall map ↑

1-1 Pairing

- Edges weighted with IoU (or other)
- A single and meaningful parameter α
  - With IoU, $\alpha = 0.5 \to$ pairing
  - **Tolerance Threshold**: the level over which no human correction is required
- Compute the number of remaining pairing T_c as the IoU increases.

Toward COCO PQ - Quantitative Analysis

$$prec. = \frac{T_c}{|P|} \quad recall = \frac{T_c}{|T|} \quad F\text{-}score = \frac{2.T_c}{|P| + |T|}$$

- Precision/Recall/F-Score Curves to analyze systems performances New!
- The COCO PQ is the AUC of the F-Score curve ! Bang!

Prec./Recall Maps - Qualitative Analysis

New! A region is valued by the best match score in the pairing.

- Precision Map. Assess the quality of detections (objects from the system)
- Recall Map. Detect the missed components (objects from the GT)

16th International Conference on Document Analysis and Recognition (ICDAR 2021), September 5-10, 2021, Lausanne, Switzerland.