Counting the Corner Cases

Revisiting **Robust Reading Challenge** Data Sets, Evaluation Protocols, and Metrics

Jerod Weinman, Amelia Gómez Grabowski, and Dimosthenis Karatzas

Grinnell College
Computer Science Department

Computer Vision Center

Problem and Motivation

Challenges have driven and measured progress for two decades. Evaluation protocols have not kept pace with task complexity; even ground truth receives sub-perfect scores. We propose protocol updates that boost both recall and precision of competition submissions and evaluate the *parameter-free* potential of Panoptic Quality (PQ) as a tightness-aware metric.

Background: Limitations

- **Correspondence:** Like VOC and COCO, prior RRCs used greedy matching to pair ground truth words with detections.
- **Ignored "Don't Cares":** Any detections sufficiently overlapping ignored ground truth regions are prevented from matching other valid ground truth items, lowering recall.
- **Sequencing:** End-to-end challenges (detection + recognition) establish correspondences first by geometry (*i.e.*, IoU>0.5), and *then* assess recognition, allowing potential mismatches.

Prior work explores limitations of one-to-one matching in RRC protocols (Wolf & Jolion, IJDAR 2006; Calarasanu *et* al., IVC 2016; Dangla *et* al., DAS 2018; Lee *et* al., ICDAR 2019; Baek *et* al., CVPRW 2020); yet *within* that assumption, we find greater nuance can increase fairness and completeness.

Test Data Statistics: Quantitative Impact of Protocol

Challenge		Year	Words	Ignores	Ignored Valid Words		Overlapping Valid Words	
Focused Scene Text	(FST)	2015	1,455	698	3	0.40%	2	0.26%
Incidental Scene Text	(IST)	2015	11,886	7,418	3	0.07%	18	0.40%
COCOText		2017	145,862	58,742	1,382	1.59%	4,848	5.56%
Multi-Lingual Scene Text	(MLT)	2019	111,998	22,562	71	0.08%	78	0.11%
Arbitrary-Shaped Text	(ArT)	2019	62,990	12,899	132	0.26%	108	0.22%
Large-scale Street View Text	(LSVT)	2019	382,606	138,969	262	0.11%	220	0.09%
Out of Vocabulary	(OOV)	2022	3,676,250	89,420	1,459	3.73%	98,088	2.73%
Hierarchical Text	(HierText)	2023	1,014,142	151,565	582	0.07%	82	0.01%

Competition Metrics

Given correspondences, many RRCs evaluate the average ratios of true positives (TP) to the ground truth set G and detections D. To account for localization tightness, HierText (Long *et al.*, CVPR 2022) uses Panoptic Quality (PQ) (Kirillov *et al.*, CVPR 2016)

Precision $P \triangleq \frac{|\mathrm{TP}|}{|D|}$ ***Recall*** $R \triangleq \frac{|\mathrm{TP}|}{|G|}$ ***H-Mean*** $F \triangleq \frac{2PR}{P+R}$

Panoptic Quality

$$\mathrm{PQ} \triangleq F \times \underbrace{\left(\frac{1}{|\mathrm{TP}|} \sum_{(g,d)\in \mathrm{TP}} \mathrm{IoU}\,(g,d) \right)}_{\text{Tightness } T}$$

Revised Evaluation Protocol

We recast the correspondence stage as **bipartite matching**: a polynomial-time linear sum assignment algorithm optimizes the metric for a given set of predictions using metric-specific match scores ψ between ground truth items $g \in G$ and detections $d \in D$.

Detection: H-Mean

$$\psi(g,d) = \begin{cases} +1 & \text{if } \mathrm{IoU}\,(g,d) > \tau \\ -1 & \text{otherwise,} \end{cases}$$

↑ Basic optimization

All improvements →

End-to-End: Panoptic Quality

(Joint Geometry+Text Criteria, Post-Filter "Don't Cares")

$$\psi(g,d) = \begin{cases} \mathrm{IoU}\,(g,d) & \text{if } \mathrm{IoU}\,(g,d) > \tau \\ & \wedge\, g_{\text{text}} = d_{\text{text}} \\ & \wedge\, \mathrm{Valid}\,(g) \\ -1 & \text{otherwise} \end{cases}$$

Jointly including **transcription** and **"don't care"** checks during optimal matching allows greater flexibility. See the paper for additional variants incorporating character edit distance (ED).

Empirical Analyses

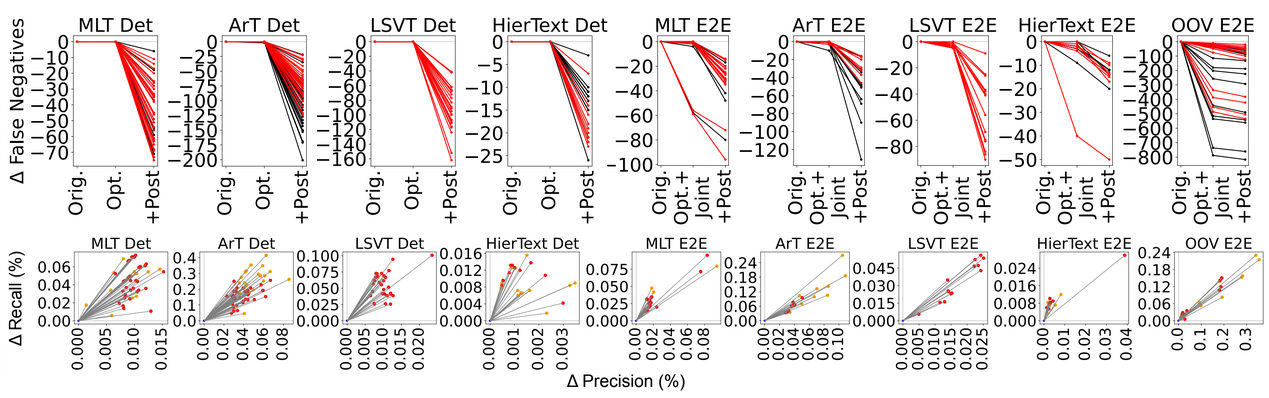
Using ground truths as predictions, existing protocols result in false negatives; our protocol revisions eliminate them.

Test Data False Negatives: Ground Truth as Prediction

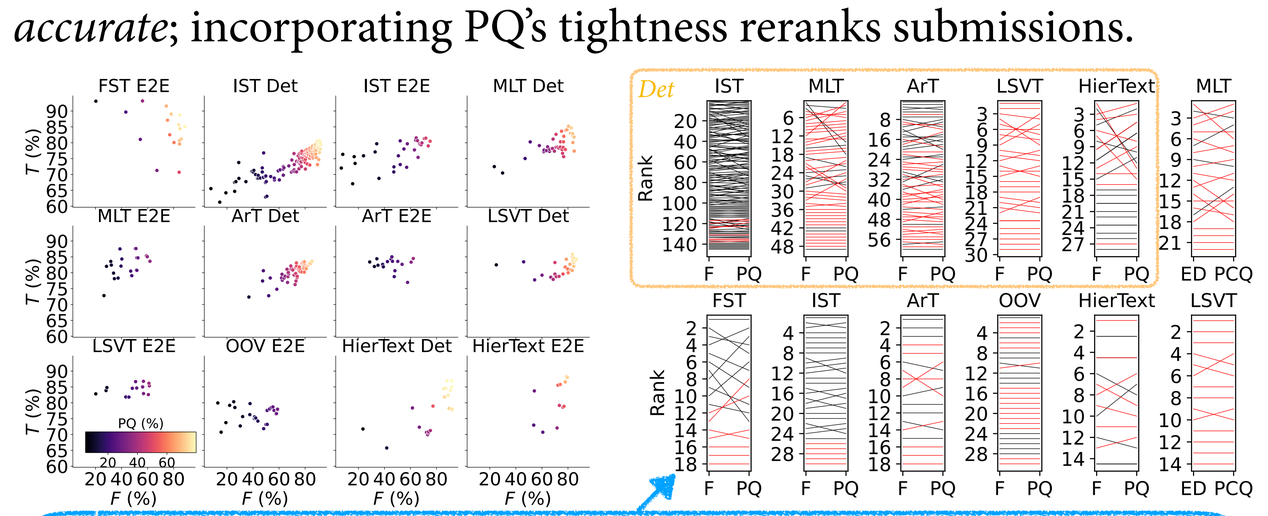
Rev-erse	Opti-mize	Joint	Post-Filt.	ArT		MLT		LSVT		OOV	HierText	
				Det	E2E	Det	E2E	Det	E2E	E2E	Det	E2E
				306	306	125	125	258	258	226	28	28
✓				306	306	125	141	258	258	448	28	28
✓	✓			306	306	125	137	258	258	230	28	28
✓	✓	✓		-	306	-	135	-	258	226	-	28
✓			✓	0	0	*2	14	*1	*1	273	0	0
✓	✓		✓	0	0	*2	5	*1	*1	7	-	0
✓	✓	✓	✓	-	0	-	*2	-	*1	*2		

Columns for FST and IST are all zeros. *Ground truth annotation have un-matchable area of zero; IoU undefined.

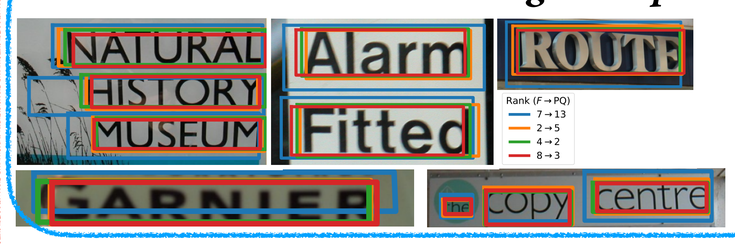
Applying protocol revisions reduces false negatives on all public RRC submissions, improving competition scores by more than the difference between many (10%) submissions.



Acceptable correspondences (IoU>0.5) may not be equally *accurate*; incorporating PQ's tightness reranks submissions.

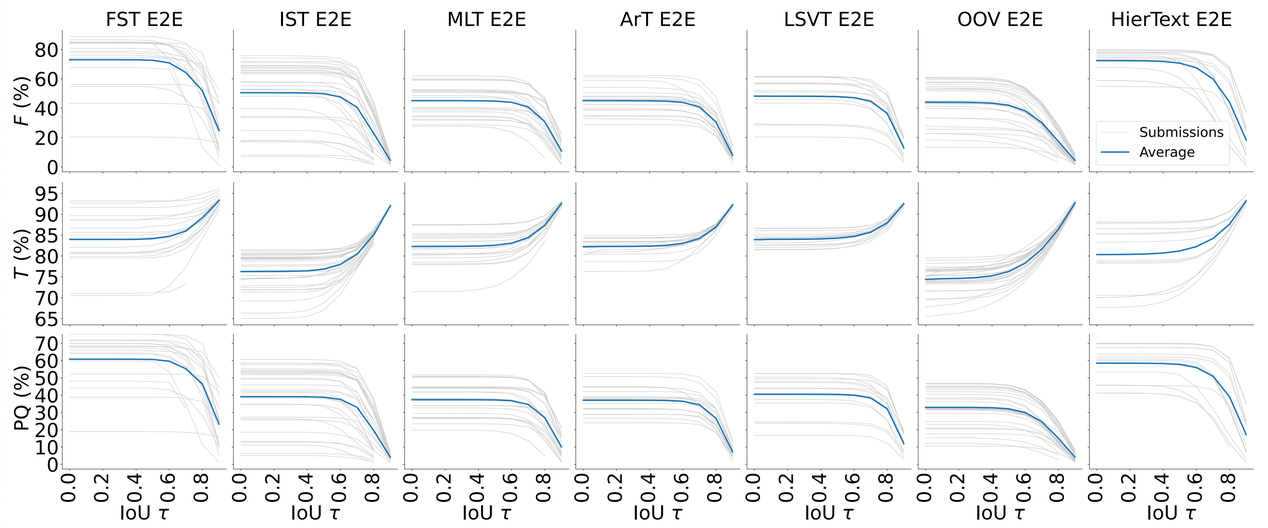


Re-Ranking Examples: FST E2E



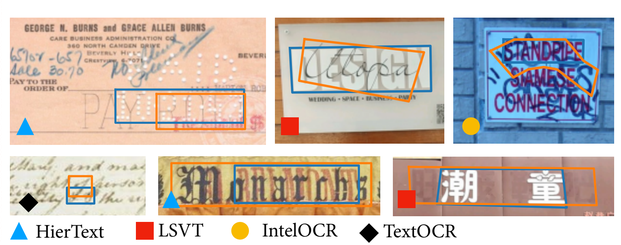
Rank (*F*)	*F*	*T*	PQ	Rank (PQ)
7	84.20	70.69	59.53	13
2	85.78	81.17	69.62	5
4	84.65	88.78	75.15	2
8	83.89	85.70	71.90	3

Optimal matching stabilizes scores, even as IoU threshold τ approaches zero, possibly obviating the parameter.



Code available: `github.com/weinman/rrc-evaluation`

Conclusion



Complex competition scenarios may require more nuanced evaluation protocols. Consider these for your next competition!