ICDAR 2024 Competition on Historical Map Text Detection, Recognition, and Linking

Zekun Li[1], Yijun Lin[1], Yao-Yi Chiang[1], Jerod Weinman[2], Solenn Tual[3,4], Joseph Chazalon[4], Julien Perret[3,5], Bertrand Duménieu[5], and Nathalie Abadie[3]

[1] University of Minnesota - Twin Cities, United States
[2] Grinnell College, United States
[3] Univ Gustave Eiffel, ENSG, IGN, LASTIG, France
[4] EPITA, France
[5] CRH, EHESS, France
{li002666, lin00786, yaoyi}@umn.edu, jerod@acm.org, {solenn.tual, julien.perret, nathalie-f.abadie}@ign.fr, joseph.chazalon@epita.fr, bertrand.dumenieu@ehess.fr

Abstract. Text on digitized historical maps contains valuable information, e.g., providing georeferenced political and cultural context. The goal of the ICDAR 2024 MapText Competition is to benchmark methods that automatically extract textual content on historical maps (e.g., place names) and connect words to form location phrases. The competition features two primary tasks—text detection and end-to-end text recognition—each with a secondary task of linking words into phrase blocks. Submissions are evaluated on two data sets: 1) *David Rumsey Historical Map Collection* which contains **936** map images covering 80 regions and 183 distinct publication years (from 1623 to 2012); 2) *French Land Registers* (created during the 19th century) which contains **145** map images of 50 French cities and towns. The competition received **44** submissions among all tasks. This report presents the motivation for the competition, the tasks, the evaluation metrics, and the submission analysis.

Keywords: Text detection · Text recognition · Historical maps.

1 Introduction: Motivation and Challenges

Maps tell stories of places, cultures, resources, and history. Digitized collections of historical maps contain a wealth of information often locked in an unsearchable raster format. This competition aims to raise awareness in the document analysis community of some unique and difficult challenges in extracting useful textual information from these cartographic artifacts (e.g., Figure 1).

With a significant, if limited, body of classical approaches to text/graphics separation applied to maps [9], only a few recent works have addressed the specific problem of map text detection and recognition with modern ML techniques [43,26,38,3,28]. The problem is similar to recent robust reading challenges

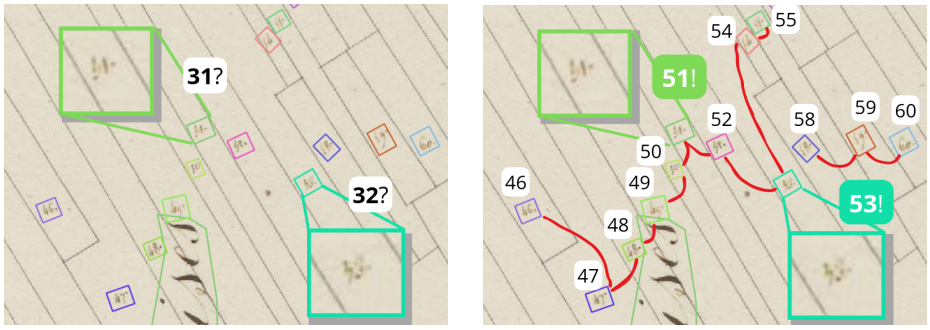
Appears in *Proceedings of the 18th International Conference on Document Analysis and Recognition (ICDAR 2024)*. Elisa Barney Smith, Marcus Liwicki, and Liangrui Peng (Eds.). Springer Nature LNCS. 2024.

Fig. 1. Querying word "Funen" from David Rumsey Map Collection website [4]. The results reveal the complexities of map text: curved, rotated, widely-spaced, interleaved with complex backgrounds, engraved, and handwritten.

in scene text [39,19,18,41,14,36]. However, the problem is also sufficiently different from prior competitions; recognizing and linking text labels in maps presents unique challenges, such as complex backgrounds, various font styles, and extremely wide character spacing. Successful solutions will significantly improve the indexing and searchability of ever-growing digital map archives [21,4].

Previously, the ICDAR 2021 Competition on Historical Map Segmentation [5] spurred methods that can identify regions corresponding to the principal cartographic area(s). Nevertheless, substantial advancement is still required to achieve the overarching objective of effectively searching and indexing historical maps. These tasks encompass steps such as word detection and recognition, as well as the process of linking individual words into coherent phrases. The **detection** and **recognition** tasks share similarities with the long line of prior robust reading competitions. However, map text detection also presents distinctive challenges. Rotated and strongly curved text is the norm on maps, rather than the exception, which only the more recent robust reading competitions have targeted [10,47]. Moreover, words can be highly spaced with complicated text-like distractors—even other words—appearing between the characters, a scenario infrequent in other reading domains.

The **word linking** task is similar to document layout analysis in that it requires grouping words together to form higher-level structures such as phrases. However, it is distinct in maps because semantics plays an unusually strong role: the multiple words of a single place name (i.e., "New York City") may not fall on a single line. The ICDAR 2023 Competition on Hierarchical Text Detection and Recognition [35] pushed the field forward by requiring scene text processing systems to link words into lines and lines into paragraphs, yet the association cues were generally more visual and geometric than semantic. Finally, in some early modern maps such as French land registers (illustrated in Figure 2), hand-



Unordered transcription Numerically-ordered transcription

Fig. 2. Excerpt of French land register map with handwritten text. Even for a human, unordered transcription is challenging (left), while a numerically-ordered transcription is more straightforward (right).

written text is intrinsically ambiguous and may require the use of the context (surrounding text) to be correctly transcribed.

In addition to the curved and rotated text prominent on many maps, Figure 3 highlights two of the additional primary challenges described above. The red boxes with very widely spaced characters overlap other words and have text-like graphical structures within. Moreover, the boxes are linked because they correspond to a single toponym ("SOUTH CAROLINA"). However, although the purple boxes are similarly situated geometrically, they should not be linked because they correspond to two separate county names ("FLORENCE" and "WILLIAMSBURG"). Although they do not form a place name, the cyan boxes are linked as the label on the rail line ("ATL. COAST LINE").

In sum, this competition report features: a rigorous definition of the tasks of text detection, recognition, and linking on historical maps, along with proposed metrics and evaluation protocols (Section 2); two new public data sets of historical maps with ground truth annotations: a large English one, covering a wide variety of years, scales, locations, and graphical styles, and a smaller French one, focusing on early modern land registers and covering a narrower style, location, and time (Section 3); a description of participants' methods (Section 4); and an analysis of the results (Section 5). This competition continues running live on the Robust Reading Competition platform at https://rrc.cvc.uab.es/?ch=28, and the data sets are available for download at https://zenodo.org/communities/icdar-maptext.

2 Tasks and Evaluation

The competition consists of multiple inter-related tasks on historical maps involving text detection and recognition at both the word and phrase levels. The four primary competition tasks are 1) word detection 2) phrase detection (word linking) 3) end-to-end word recognition and 4) end-to-end phrase recognition.

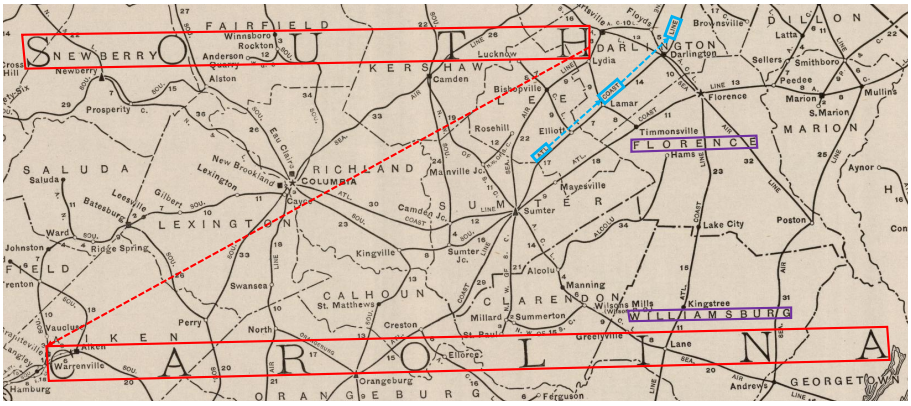


Fig. 3. Word detection and linking (red, cyan) and non-linking (purple) examples. Image credit: Rumsey Collection [4] Image 5028.054 (Rand McNally and Company, *South Carolina*, 1924).

Each of these primary tasks consists of two evaluations, which differ only in the data's target scope. One evaluation covers a very wide range of maps (using the "Rumsey" data set), while the other involves data more narrowly tailored to a specific place, time, and map style (using the "French land registers" data set). In this way, the competition assesses general system performance, as well as the ability to target a particular map collection. Section 3 details that evaluation data; in this section, we elaborate on the four general tasks and corresponding evaluation metrics. See the supplementary material for complete details on the evaluation protocol, which puts ground truth polygons in correspondence with detected polygons. The evaluation code is publicly available at https://github.com/icdar-maptext/evaluation.

Task 1: Word Detection

Although the end goal is recognizing text to enable searching and processing, many systems begin with a stage that localizes the textual elements in the images. In robust reading for scene images, this task has been called text detection, but it is sometimes known as text/graphics separation in the document analysis community. Because map text is different from traditional documents (including even engineering drawings) and scene text in some key ways, it is important to measure the performance of state-of-the-art text detection systems in this context and encourage new methodologies.

The goal of this task is to detect individual words on map images, i.e., generating bounding polygons that enclose text instances at the word level. Many text instances may be straightforward to detect. However, due to the extreme intra-word character spacing and other cartographic or printer artifacts, the task of drawing a single polygon around a word remains challenging (see Figure 4).

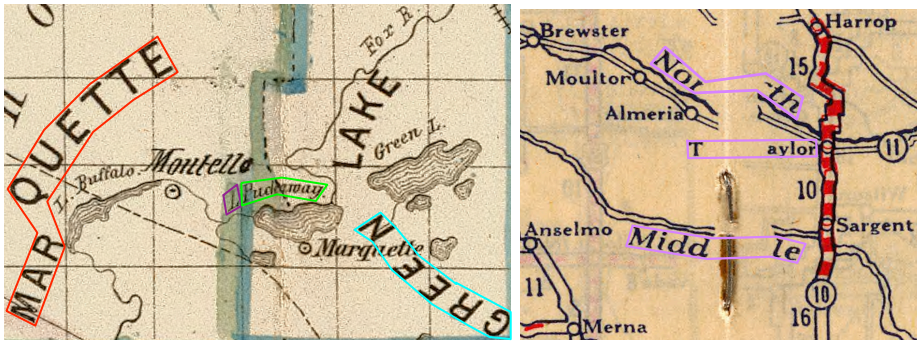


Fig. 4. Word detection challenges. LEFT: Highly irregular baselines (red), very tight spacing (purple and green), and interposed text (cyan). RIGHT: Printing and binding artifacts (lavender). Image credits: Rumsey Collection [4] Image 1070.005 (U.S. General Land Office, *Wisconsin*, 1866); Image 5755.025 (Rand McNally and Company, *Nebraska*, 1927).

As Long et al. [34] point out, the Panoptic Quality (PQ) metric [20] is attractive for text detection tasks because it combines the familiar F-measure (a.k.a, F1 or H-mean) often used in detection evaluation with an average of IoU scores that can promote methods with more precise localizations/segmentations. Although a COCO-style mAP evaluation (averaging over several IoU thresholds) has the same effect of rewarding methods with better localization, AP requires prediction confidence scores to calculate the precision-recall curve.

As in many prior RRCs, the set of true positives in a one-to-one matching protocol must have sufficient overlap between the ground truth word and detection polygon,

$$\mathrm{TP}_{\mathrm{Det}} \subset \{(g,d) \in G \times D \mid \mathrm{IoU}\,(g,d) > 0.5\}\,, \tag{1}$$

where G is the set of ground truth regions and D is the set of detected regions.

Following HierText [34,35], we use **Panoptic Detection Quality** (PDQ) to evaluate word detection,

$$\mathrm{PDQ} \triangleq T \times F, \tag{2}$$

where the tightness T is the average IoU among true positive regions

$$T \triangleq \frac{1}{|\mathrm{TP}|} \sum_{(g,d)\in \mathrm{TP}} \mathrm{IoU}\,(g,d)\,, \tag{3}$$

and F represents the F-score, the harmonic mean between precision and recall:

$$P \triangleq \frac{|\mathrm{TP}|}{|\mathrm{TP}| + |\mathrm{FP}|} \qquad R \triangleq \frac{|\mathrm{TP}|}{|\mathrm{TP}| + |\mathrm{FN}|} \qquad F \triangleq \frac{2PR}{P+R}.$$

Task 2: Phrase Detection (Word Grouping)

Downstream tasks typically involve processing over toponyms, rather than individual words. For this reason, and because phrases represent a semantically coherent unit of text on the page, we want systems that can perform the layout task of grouping together semantically associated text fragments. Nearly all such groups on a map are the multiple parts of a single place name. However, other groups might correspond to phrases labeling or explaining other map elements (see Figure 3 and accompanying discussion in Section 1).

Much like the HierText competition [35], the words that must be linked into a single group are treated as one unit for (joint) detection. Competition entries give words (with their polygon boundaries) in a list belonging to a phrase.

The unions of the word polygons forming predicted and ground truth phrase groups are used to calculate the IoU. The PDQ score at the phrase level is calculated among corresponding unions; thus word order is ignored. This protocol also allows some flexibility in word segmentation errors because the matching is essentially many-to-many among words within the group. Unlike HierText, we do not combine the scores across multiple levels, but instead evaluate tasks 1 and 2 independently.

Task 3: Word Detection and Recognition

The goal of this task is to jointly locate and recognize words on maps, as in many previous robust reading competitions [19,14,10,13,35]. End-to-end map text detection and recognition builds on the text detection task by attaching a text transcription to each detected word polygon.

Given a map image, participants were expected to produce word-level text detection and recognition results: a set of word bounding polygons and corresponding transcriptions. As in prior RRCs, true positives must have matching transcriptions in addition to meeting the IoU threshold:

$$\mathrm{TP}_{\mathrm{Rec}} \subset \{(g, d) \in G \times D \mid \mathrm{IoU}\,(g, d) > 0.5 \wedge g_{\mathrm{text}} = d_{\mathrm{text}}\}\,. \tag{4}$$

We introduce the **Panoptic Word Quality** (PWQ) metric. It is identical to PDQ, except that it uses $\mathrm{TP}_{\mathrm{Rec}}$; in this regard, PWQ is likewise useful as the metric because it effectively combines a) localization accuracy (tightness), b) detection quality (polygon presence/absence), and c) word-level recognition accuracy in a single measure. As the competition metric, PWQ gives strong preference for well-localized, textually accurate detections.

Task 4: Phrase Detection and Recognition

While detection with phrase grouping is evaluated in Task 2, we measure the overall end-to-end performance with a "grand challenge" of localized place name search, i.e., end-to-end joint detection, linking, and recognition.

This task requires the detection and recognition of the entire label phrase. Word polygons and transcriptions are grouped as an ordered list into phrases.

Word recognition accuracy can be quite stringent, as measured by the PWQ calculated with $\mathrm{TP}_{\mathrm{Rec}}$ (4). Therefore we also desire a metric that accounts for character-level recognition accuracy by comparing the edit distance between ground truth and predicted text. To this end, we propose the **Panoptic Character Quality** (PCQ) metric:

$$\mathrm{PCQ} \triangleq T \times F \times C, \tag{5}$$

where $T \times F$ are as in the original PDQ metric, and C represents the average complementary normalized edit distance [48] of each word's text among the matched true positive detections $\mathrm{TP}_{\mathrm{Det}}$, as originally defined in Eq. (1),

$$C \triangleq 1 - \frac{1}{|\mathrm{TP}|} \sum_{(g,d)\in \mathrm{TP}} \mathrm{NED}\,(g, d)\,. \tag{6}$$

Here NED is the standard normalized edit distance between the detected and ground truth regions' strings, each of which is taken as the space-separated concatenation of the words in the group.

With its three factors $T, F, C \in [0, 1]$, the product PCQ falls into the same range. Thus, the PCQ combines a) localization accuracy (tightness), b) detection quality, and c) character-level recognition accuracy into a single measure, free of additional parameters.

We use PCQ rather than PWQ to evaluate this challenging task. First, as is often the case, we expect character error rates to be much lower than word error rates. Using character errors at this level helps assess the overall degree of recognition accuracy, particularly in the open-vocabulary setting. Second, errors in word segmentation that are otherwise grouped correctly are not as heavily penalized because the edit distance only counts the spaces when missing (an under-segmentation) or extra (an over-segmentation).

The supplementary material contains additional important details of the evaluation protocol for this and all tasks.

3 Data Sets and Annotations

The competition comprises the combination of two data sources: human-annotated selections from the David Rumsey Historical Map Collection and a series of French land registers. This section provides an overview of these distinct data sets. See the supplementary material for details on map selection, annotation, and version history. All data sets are archived at Zenodo [29,33,6,7].

David Rumsey Historical Map Collection This archive hosts an extensive set of over 126,000 maps accessible online [4]. The catalog spans maps from the 16th to the 21st century, encompassing regions from every continent, the Pacific, the Arctic, and the entirety of the World. From this rich assortment, we select **936** representative maps for human annotation using style clustering. The sampled maps cover 80 regions and 183 distinct publication years (from 1623 to 2012).

Table 1. Data set statistics.

	Rumsey			**French Land Registers**		
	Train	Validation	Test	Train	Validation	Test
Tiles	200	40	700	80	15	50
Map Sheets	196	40	700	37	9	49
Words	34 518	5544	128 457	8096	1801	7346
Label Groups	21 205	3502	78 582	7449	1661	6814
Illegible Words	1870	313	8116	563	217	450
Truncated Words	3582	628	14 566	371	91	300
Valid Words	30 563	4860	111 821	7533	1584	6896
Average words per group	1.63	1.58	1.63	1.09	1.08	1.08
Fraction of valid words	0.89	0.88	0.87	0.93	0.88	0.94

Digitized map images can be on the order of 6K–15K pixels per dimension. For the competition, we crop each map into 2K × 2K pixel tiles and select 1–2 tiles for annotation. The total number of annotated cropped tiles is 940. We split the map tiles into 200 for training, 40 for validation, and 700 for testing. Four maps have multiple tiles in training; all test tiles are from distinct maps and the splits are also disjoint. Table 1 provides statistics for the data sets.

French Land Registers To complement the broad selection of maps from the previous collection—which covers a diversity of scales, styles, geographical region and historical period—we also provide another subset covering a very narrow region and time. This second subset is composed of 19th century land registers from approximately 50 French cities and towns, at a very large scale. (In a cartographic context, a "large scale" map means it covers a relatively small area in great detail.) These cadastral plans contain an important quantity of parcel numbers, now-forgotten place names, and many other local details (in French) relevant to the accurate delineation of parcels in order to identify owners and compute taxes. The entire online collection consists of over 800 map sheets [1].

For the competition, map sheets were selected according to two criteria: we used stratification to ensure a good representation of the different map types (first or second campaign, geographic area), and we grouped the maps by the city they represent to avoid any overlap between the training, validation, and test sets in terms of geographic area. The resulting data set, summarized in Table 1, features more than 17,000 words from 145 different map sheets. Contrary to the Rumsey data set, the French land registers images have a lower quality, both in terms of resolution and contrast, and the text is mostly handwritten. However, it contains fewer groups and many words represent numbers.

4 Competition Protocol and Participants

The competition is hosted on the well-established RRC platform (https://rrc.cvc.uab.es), where competition results are standardized, archived, and future post-

competition submissions can track progress in the field. This platform enables the submission of predicted results computed by the participants themselves. Thus, participants did not need to provide code or binaries for their predictions to be accepted. However, we encouraged participants to provide links to papers, data sets, models, and public codes if available. The competition was open to all participants, with the following timeline: training and validation sets were released on Feb. 2, 2024, and the test set was released on March 4, 2024. The deadline for submitting results was May 6, 2024; participants had three months to train their models using any open data sets (disjoint from the test set).

The remainder of the section lists the primary eight teams of participants.

MapTest Hongen Liu (Tianjin University)

Winner Task 2 (Rumsey & French), Task 3 (French), Task 4 (French)

This team participates in all four tasks on both map data sets and adopts different approaches for each task. For task 1, their method uses the PP-YOLOE-R [42] model pretrained on COCOTextV2 [41] and finetuned on the target David Rumsey data set and French Land Registers data set. For task 3, the model is the ABINet [12] pretrained on the MJSynth [17] and SynthText [15] data sets and finetuned on the target data sets. The linking solution is a heuristic-based approach where text labels with heights larger than 50 pixels are selected as linking candidates, and the linking is performed by comparing the distance, bounding box height, and rotation angles.

MapText Detection Strong Pipeline Yu Xie and Ziyue Wang (Bilibili Inc.)

Winner Task 1 (Rumsey), Task 3 (Rumsey), Task 4 (Rumsey)

This team participates in all four tasks on both map data sets. The method is built upon DeepSolo [46] network with ViTAEv2-S [49] as the backbone followed by transformer-based encoder and decoders. The encoder features are used to predict the Bezier center curves for the bounding polygon, and the decoder features are used to recognize the text. This method uses existing data sets as training data: TextOCR [40], MLT17 [37] , TotalText [11], ICDAR15 [18], and ICDAR13 [19], which contain a total of 35,109 images.

DINO_MAP, DINO_MVIT, & ENSEM Rajat Kumar Singh, Himani Shrotriya, Shivshankar Reddy, and Himanshu Bhatt

Winner Task 1 (French)

This team participates in task 1 on both map data sets and attempts multiple approaches. DINO_MAP uses the MaskDINO [22] network, which is an object-detection and instance segmentation fine-tuned on both map data sets. DINO_MVIT uses the MViTv2 [25] model, which is another object detection network strong at handling objects in different scales (i.e., sizes). MViTv2 was applied to the Rumsey

data set. ENSEM is an ensemble method of MaskDINO, MViTv2 and ViTDet [24], while the team notices that DINO_MAP gives the best performance instead of the ensemble method. To handle maps with dense text labels, their methods crop the input map into four overlapping patches and merges the detected regions according to the overlapping ratio.

MapTextSpotter Jialiang Li, Canhui Xu, Cao Shi, and Yucai Qu (Qingdao University of Science and Technology)

This team participates in task 1 and 3 on the Rumsey map data set. The novel proposed model MapTextSpotter utilizes a Transformer-based decoder model to predict text Bezier curve points and character classification in parallel. The point and character queries are designed to incorporate spatial and semantic text distribution in historical maps. The approach also employs a Large Language Model to enhance the recognition precision.

DS-LP Siyuan Huang (Beijing University of Posts and Telecommunications)

This team participates in all four tasks on both map data sets. For detection and recognition, their approach uses the DeepSolo++ [45] network with a ViTAEv2-S [49] backbone. The weights are initialized from DeepSolo [46] and fine-tuned on the competition data sets. To reduce the number of overlapping polygons, their method applies Non-Maximum Suppression (NMS) post-processing on the output of DeepSolo++. For linking, the team designs a novel network called LayoutPointer, which is a relation extraction model based on LayoutLMv3 [16], to predict relationships between text boxes.

MapText Using EasyOCR/TrOCR Pengyu Chen, Xuezi Bi, Quanzhi Xiang and Junxian Li (University of South Carolina; Sun Yat-sen University; University of Science and Technology of China; Beihang University)

This team participates in task 1 and 3 on the Rumsey map data set. For the map detection problem (task 1), they employ the CRAFT [2] model embedded in EasyOCR. The method outputs a rectangle bounding box with four vertices for each text label. For the recognition problem (task 3), the method crops out the text label patches using the predicted bounding boxes and then feeds the patches to the TrOCR [23] to recognize the text.

MapDet Yize Yang, Chaolang Li, Jingyu Li, Pengwen Dai (Sun Yat-sen University)

This team participates in task 1 on both map data sets. Their text detection model is based on DBNet [27], with additional modules designed to enhance the learning of boundary regions and text. They incorporate auxiliary modules for multi-scale and multi-view learning for text recognition.

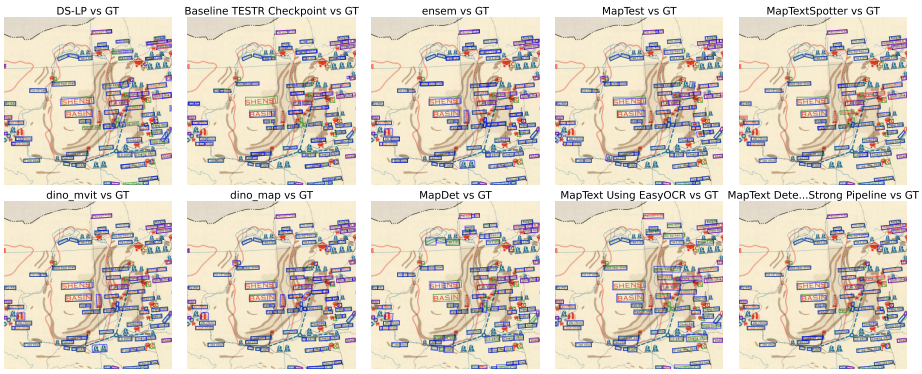


Fig. 5. Example results for Task 1 on the Rumsey data set. Blue regions are predictions for the entitled submission. Green indicates valid ground truth words, while red indicates cropped or ignored words. Consult the supplementary material for many additional examples from all tasks and data sets.

Baseline TESTR Checkpoint MapText Competition Organizers

> We adopt an existing text spotting model, TESTR [50], to detect and recognize text instances on maps. The model is built upon Deformable DETR [51] and uses dual decoders for text-box control point regression and character recognition, respectively. We use existing model weights to generate baseline results for both data sets. The model was pretrained on SynthText and multiple human-annotated scene image data sets and finetuned on the TotalText data set.

5 Results and Discussion

This section reports the quantitative results for each task of the competition on the two data sets. It also offers some insights with qualitative analysis.

5.1 Results for Task 1: Word Detection

Table 2 presents results for task 1 on both data sets. As the table shows, the best method for the Rumsey data set is the "MapText Detection Strong Pipeline", while the best method for the French Land Register data set is "DINO_MAP". In both cases, the "DINO_MAP" method is closely followed by "MapTest" which exhibits a better detection performance (higher F-score), but a weaker tightness. The ranking of "DINO_MAP" indicates that with proper fine-tuning, the general object detection models can achieve quite good performance on text detection tasks. However, the main limitation of such an approach is that it does not support end-to-end spotting and requires a separate recognition step to get the final text labels. As confirmed by qualitative results (provided in the supplementary

Table 2. Results for task 1 (isolated word detection). Values expressed in percentage. For all metrics, higher is better.

	Rank	Method name	Det. Quality	Tightness	FScore	Precision	Recall
Rumsey	1	MapText Dete...Strong Pipeline	76.1	82.7	92.0	94.2	89.9
	2	DINO_MAP	73.4	84.0	87.3	87.2	87.5
	3	MapTest	73.1	81.8	89.3	90.5	88.2
	4	DINO_MVIT	72.4	83.6	86.7	89.2	84.2
	5	MapTextSpotter	70.6	81.4	86.7	92.6	81.5
	6	ENSEM	64.3	85.6	75.1	94.4	62.3
	7	Baseline TESTR Checkpoint	55.1	79.6	69.3	71.9	66.9
	8	DS-LP	53.8	71.6	75.2	71.8	78.9
	9	MapText Using EasyOCR	42.7	73.2	58.3	69.3	50.4
	10	MapDet	32.7	69.2	47.2	53.6	42.2
French	1	DINO_MAP	64.7	72.2	89.7	88.7	90.8
	2	MapTest	64.2	69.9	91.9	90.8	93.0
	3	ENSEM	52.0	71.4	72.9	90.8	60.9
	4	DS-LP	44.1	65.0	67.8	64.8	71.0
	5	MapText Dete...Strong Pipeline	42.3	69.3	61.1	82.5	48.5
	6	MapDet	35.7	65.3	54.7	70.1	44.8
	7	Baseline TESTR Checkpoint	20.6	70.5	29.2	86.4	17.6

material), the "MapText Detection Strong Pipeline" is very sensitive to the low quality of the French Land Register data set, especially regarding small text sizes and low contrast text. This suggests that its training material may lack some challenging elements of this sort. Also, it should be noted that in terms of raw detection quality, the baseline method "Baseline TESTR Checkpoint" offers a decent performance as an annotation assistance. Indeed, for both data sets its tightness is high, indicating that little modification to the shapes is needed to match the ground truth. However, while the precision is perhaps acceptable on the French Land Register data set (86.4%), it is much lower for Rumsey (71.9%), and many false positives would need to be manually discarded.

5.2 Results for Task 2: Phrase Detection (Word Grouping)

Table 3 presents results for task 2. As expected, this task is more challenging than task 1, especially regarding the Rumsey data set, which contains a higher number of groups. We confirmed this by evaluating the scores obtained by removing the links from the ground truth, and evaluating it against the ground truth. On task 2, such "linkless" evaluation gives a Detection Quality of 56.1% for the Rumsey data set and 93.9% for the French Land Register data set, effectively showing that the Rumsey data set is more challenging for this task. On the Rumsey data set, the "MapTest" and the "MapText Detection Strong Pipeline" methods are the best, with a very close Detection Quality. We also note the resilience of the "DS-LP" approach, whose performance is remarkably stable on the French land registers. Indeed, qualitative analysis reveals a very

Table 3. Results for task 2 (grouped word detection). Values expressed in percentage. For all metrics, higher is better.

	Rank	Method name	Det. Quality	Tightness	FScore	Precision	Recall
Rumsey	1	MapTest	41.9	74.4	56.3	44.2	77.8
	2	MapText Dete...Strong Pipeline	41.5	75.4	55.1	43.2	75.8
	3	Baseline TESTR Checkpoint	35.5	75.0	47.3	37.8	63.2
	4	DS-LP	35.0	69.7	50.2	39.4	69.4
French	1	MapTest	59.8	68.6	87.1	83.7	90.9
	2	DS-LP	43.3	64.9	66.7	63.1	70.8
	3	MapText Dete...Strong Pipeline	30.7	68.8	44.6	44.2	45.0
	4	Baseline TESTR Checkpoint	14.6	68.7	21.2	58.7	13.0

promising linking performance on this data set. In the future, a precision/recall evaluation directly over links may be even more informative.

5.3 Results for Task 3: Word Detection and Recognition

Table 4 presents results for task 3. On the Rumsey data set, the best method is the "MapText Detection Strong Pipeline," apparently benefiting from both a solid detection of isolated words and a good recognition rate. The second-best method on this data set, "MapTest", is the best on the French Land Register data set by a large margin, exhibiting the strongest detection and recognition rates. The drop in performance of the "MapText Detection Strong Pipeline" on the French Land Register data set is likely due not only to its poorer detection performance, but heavily compounded by transcription inaccuracy. As mentioned in the introduction, text recognition is very challenging on both data sets because of the multiple clues required to disambiguate how a word should

Table 4. Results for task 3 (word detection with perfect transcription). Values expressed in percentage. For all metrics, higher is better.

	Rank	Method name	Det. Quality	Tightness	FScore	Precision	Recall
Rumsey	1	MapText Dete...Strong Pipeline	60.1	84.2	71.3	76.0	67.2
	2	MapTest	52.3	83.8	62.5	63.3	61.7
	3	MapTextSpotter	41.1	82.8	49.6	53.0	46.6
	4	DS-LP	37.9	72.5	52.3	49.9	54.9
	5	Baseline TESTR Checkpoint	27.8	84.6	32.9	34.1	31.8
	6	Recognition ...uned from TrOCR	12.2	78.5	15.5	18.5	13.4
French	1	MapTest	40.1	70.7	56.7	56.0	57.3
	2	DS-LP	26.1	65.7	39.7	38.0	41.6
	3	MapText Dete...Strong Pipeline	8.6	71.0	12.2	16.5	9.7
	4	Baseline TESTR Checkpoint	2.2	74.7	2.9	8.6	1.8

Table 5. Results for task 4 (joint grouped word detection and transcription). Values expressed in percentage. For all metrics, higher is better.

	Rank	Method name	Char. Quality	Char. Acc.	Tightness	FScore	Precision	Recall
Rumsey	1	MapText Dete...Strong Pipeline	33.1	79.7	75.4	55.1	43.2	75.8
	2	MapTest	32.0	76.3	74.4	56.3	44.2	77.8
	3	DS-LP	28.6	81.6	69.7	50.2	39.4	69.4
	4	Baseline TESTR Checkpoint	26.2	74.0	75.0	47.3	37.8	63.2
French	1	MapTest	51.0	85.3	68.6	87.1	83.7	90.9
	2	DS-LP	37.1	85.6	64.9	66.7	63.1	70.8
	3	MapText Dete...Strong Pipeline	17.1	55.7	68.8	44.6	44.2	45.0
	4	Baseline TESTR Checkpoint	6.0	41.5	68.7	21.2	58.7	13.0

be transcribed. While the performance of leading methods is encouraging, such integration is likely one of the next big challenges to be solved.

5.4 Results for Task 4: Phrase Detection and Recognition

Table 5 presents results for task 4. Despite decent character accuracies on the Rumsey data set, group detection strongly penalizes the methods, as shown by the low detection quality. The two best methods on this data set are "MapText Detection Strong Pipeline" and "MapTest", the former having a slight advantage due to better recognition quality. On the French Land Register data set, the "MapTest" method is benefiting from a very high recognition quality, leading to better performance overall. As for task 2, we studied the impact of the links on the evaluation, and found similar drops in performance between tasks 3 and 4.

Table 6 summarizes the results on both data sets for each task. The supplementary material illustrates example results for all methods, tasks, and data.

Table 6. A summary of the results for both data sets, Rumsey (R) and French Land Register (FLR), for each task.

	Task 1		**Task 2**		**Task 3**		**Task 4**	
Method Name	R	FLR	R	FLR	R	FLR	R	FLR
MapTest	3	2	1	1	2	1	2	1
MapText [...] Pipeline	1	5	2	3	1	3	1	3
DINO_MAP	2	1						
DINO_MVIT	4							
MapTextSpotter	5				3			
DS-LP	8	4	4	2	4	2	3	2
ENSEM	6	3						
Baseline TESTR [...]	7	7	3	4	5	4	4	4
MapText [...] EasyOCR	9				6			
MapDet	10	6						

6 Conclusion and Final Ranking

This competition featured a robust reading task on historical maps, a new challenging visual content. The nature of such documents—with multiple text sizes and orientations, the paramount importance of visual clues, surrounding context, as well as some common sense required to read the text—makes the new public data sets good targets for the most advanced architectures. We also introduced new evaluations grounded on solid theoretical and experimental studies.

Thanks to the participants of this competition, to whom we express our sincere gratitude, we were able to collate an initial set of baseline methods for the main challenges and draw some general conclusions.

First, some methods have excellent detection performance for isolated words, but performance remains highly dependent on training, as the best ranking methods are not the same across data sets. In particular, some approaches do poorly with small text while others are more stable. The best methods produce complex polygons matching any text orientation. From a visual point of view, detection and segmentation quality are very high, and it may only be a matter of tuning architectures and training on more data to reach human performance.

On the other hand, grouping distant related words is a challenging task. Capturing the faint semantic relation between these components involves advanced skills for humans, who rely on visual appearance, spatial organization, and toponomy. The link prediction task therefore remains largely unsolved, even if we must acknowledge the impressive performance of the DS-LP approach and MapTest on the French Land Registers.

Finally, transcription is far from being solved for these documents. While some images yielded a high level of automated transcriptions, many in this data set are challenging. Greater success may require advanced contextualization to read elements accurately (i.e., relating toponym fragments, or jointly recognizing a sequence of plot numbers to better disambiguate their transcriptions).

The winners of the MapText competition are MapTest, MapText Strong Pipeline, and DINO_MAP. MapTest and MapText Strong Pipeline participated in all tasks on both data sets and had the best overall performance among all performers. DINO_MAP participated in task 1 and ranked place 1 and 2 for the French Land Registers and Rumsey data set, respectively. Also, we highlight DS-LP for its strong second-place rankings for task 2, task 3, and task 4 on the French Land Registers data set.

We encourage the reuse of the public material (data sets and evaluation code), as well as browsing the supplementary material containing qualitative results from the competition, via our Zenodo community at https://zenodo.org/communities/icdar-maptext.

Acknowledgments. The authors thank David Rumsey for his generous support for the competition. We thank Sergi Robles and Dimosthenis Karatzas for support with the RRC server. This work is partially supported by the French Ministry of the Armed Forces - Defence Innovation Agency (AID). Digitized French land registers are provided by the Archives of the French *departement* of Val-de-Marne (AD94).

References

1. Archives Départementales du Val de Marne: Cadastre napoléonien. https://archives.valdemarne.fr/recherches/archives-en-ligne/cadastre-napoleonien
2. Baek, Y., Lee, B., Han, D., Yun, S., Lee, H.: Character region awareness for text detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9365–9374 (2019)
3. Can, Y.S., Erdem Kabadayi, M.: Text detection and recognition by using CNNs in the austro-hungarian historical military mapping survey. In: The 6th international workshop on historical document imaging and processing. pp. 25–30 (2021)
4. Cartography Associates: David Rumsey map collection. https://www.davidrumsey.com
5. Chazalon, J., Carlinet, E., Chen, Y., Perret, J., Duménieu, B., Mallet, C., Géraud, T., Nguyen, V., Nguyen, N., Baloun, J., Lenc, L., , Král, P.: ICDAR 2021 competition on historical map segmentation. In: Proceedings of the 16th International Conference on Document Analysis and Recognition (ICDAR'21). Lausanne, Switzerland (2021)
6. Chazalon, J., Tual, S., Abadie, N., Duménieu, B., Perret, J., Weinman, J.: IGN test data for ICDAR'24 MapText competition (Mar 2024). https://doi.org/10.5281/zenodo.10732281
7. Chazalon, J., Tual, S., Abadie, N., Duménieu, B., Perret, J., Weinman, J.: IGN Train and Validation Data for ICDAR'24 MapText Competition (Apr 2024). https://doi.org/10.5281/zenodo.10987299
8. Chazalon, J., Tual, S., Abadie, N., Duménieu, B., Perret, J., Weinman, J.: IGN Train and Validation Data for ICDAR'24 MapText Competition (Feb 2024). https://doi.org/10.5281/zenodo.10610732
9. Chiang, Y.Y., Leyk, S., Knoblock, C.A.: A survey of digital map processing techniques. ACM Computing Surveys (CSUR) **47**(1), 1–44 (2014)
10. Chng, C.K., Liu, Y., Sun, Y., Ng, C.C., Luo, C., Ni, Z., Fang, C., Zhang, S., Han, J., Ding, E., et al.: ICDAR2019 robust reading challenge on arbitrary-shaped text - RRC-ArT. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1571–1576. IEEE (2019)
11. Ch'ng, C.K., Chan, C.S., Liu, C.: Total-Text: Towards orientation robustness in scene text detection. International Journal on Document Analysis and Recognition (IJDAR) **23**, 31–52 (2020). https://doi.org/10.1007/s10032-019-00334-z
12. Fang, S., Xie, H., Wang, Y., Mao, Z., Zhang, Y.: Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
13. Garcia-Bordils, S., Mafla, A., Biten, A.F., Nuriel, O., Aberdam, A., Mazor, S., Litman, R., Karatzas, D.: Out-of-vocabulary challenge report. In: European Conference on Computer Vision. pp. 359–375. Springer (2022)
14. Gomez, R., Shi, B., Gomez, L., Numann, L., Veit, A., Matas, J., Belongie, S., Karatzas, D.: ICDAR2017 robust reading challenge on COCO-text. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 1435–1443. IEEE (2017)
15. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)

16. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: LayoutLMv3: Pre-training for document AI with unified text and image masking. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4083–4091 (2022)
17. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Reading text in the wild with convolutional neural networks. International Journal of Computer Vision **116**(1), 1–20 (jan 2016)
18. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., et al.: ICDAR 2015 competition on robust reading. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 1156–1160. IEEE (2015)
19. Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., Gomez i Bigorda, L., Robles Mestre, S., Mas, J., Mota, D.F., Almazan, J.A., De Las Heras, L.P.: ICDAR 2013 robust reading competition. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 1484–1493. IEEE (2013)
20. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9404–9413 (2019)
21. Klokan Technologies GmbH: Old maps online. https://www.oldmapsonline.org
22. Li, F., Zhang, H., xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask DINO: Towards a unified transformer-based framework for object detection and segmentation (2022)
23. Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F.: TrOCR: Transformer-based optical character recognition with pre-trained models. Proceedings of the AAAI Conference on Artificial Intelligence **37**(11), 13094–13102 (2023). https://doi.org/10.1609/aaai.v37i11.26538
24. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: European Conference on Computer Vision. pp. 280–296. Springer (2022)
25. Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C.: Mvitv2: Improved multiscale vision transformers for classification and detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4804–4814 (2022)
26. Li, Z., Chiang, Y.Y., Tavakkol, S., Shbita, B., Uhl, J.H., Leyk, S., Knoblock, C.A.: An automatic approach for generating rich, linked geo-metadata from historical map images. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 3290–3298 (2020)
27. Liao, M., Zou, Z., Wan, Z., Yao, C., Bai, X.: Real-time scene text detection with differentiable binarization and adaptive scale fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
28. Lin, Y., Chiang, Y.Y.: Hyper-local deformable transformers for text spotting on historical maps. In: Proceedings of the 30th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (accepted) (2024)
29. Lin, Y., Li, Z., Chiang, Y.Y., Weinman, J.: Rumsey Train and Validation Data for ICDAR'24 MapText Competition (Jun 2024). https://doi.org/10.5281/zenodo.11516933
30. Lin, Y., Li, Z., Chiang, Y.Y., Weinman, J.: Rumsey Train and Validation Data for ICDAR'24 MapText Competition (Feb 2024). https://doi.org/10.5281/zenodo.10608901
31. Lin, Y., Li, Z., Chiang, Y.Y., Weinman, J.: Rumsey Train and Validation Data for ICDAR'24 MapText Competition (Feb 2024). https://doi.org/10.5281/zenodo.10656556

32. Lin, Y., Li, Z., Chiang, Y.Y., Weinman, J.: Rumsey Train and Validation Data for ICDAR'24 MapText Competition (Apr 2024). https://doi.org/10.5281/zenodo.10997972
33. Lin, Y., Li, Z., Chiang, Y.Y., Weinman, J.: Rumsey test data for ICDAR'24 MapText competition (Mar 2024). https://doi.org/10.5281/zenodo.10776183
34. Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., Raptis, M.: Towards end-to-end unified scene text detection and layout analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1049–1059 (2022)
35. Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., Raptis, M.: ICDAR 2023 competition on hierarchical text detection and recognition. arXiv preprint arXiv:2305.09750 (2023)
36. Nayef, N., Patel, Y., Busta, M., Chowdhury, P.N., Karatzas, D., Khlif, W., Matas, J., Pal, U., Burie, J.C., Liu, C.l., et al.: ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition—RRC-MLT-2019. In: 2019 International conference on document analysis and recognition (ICDAR). pp. 1582–1587. IEEE (2019)
37. Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., Luo, Z., Pal, U., Rigaud, C., Chazalon, J., et al.: ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification-RRC-MLT. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR). vol. 1, pp. 1454–1459. IEEE (2017)
38. Schlegel, I.: Automated extraction of labels from large-scale historical maps. AGILE: GIScience Series **2**, 12 (2021)
39. Shahab, A., Shafait, F., Dengel, A.: Icdar 2011 robust reading competition challenge 2: Reading text in scene images. In: 2011 international conference on document analysis and recognition. pp. 1491–1496. IEEE (2011)
40. Singh, A., Pang, G., Toh, M., Huang, J., Galuba, W., Hassner, T.: TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8802–8812 (2021)
41. Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S.: COCO-Text: Dataset and benchmark for text detection and recognition in natural images. arXiv preprint arXiv:1601.07140 (2016)
42. Wang, X., Wang, G., Dang, Q., Liu, Y., Hu, X., Yu, D.: PP-YOLOE-R: An efficient anchor-free rotated object detector. arXiv preprint arXiv:2211.02386 (2022)
43. Weinman, J., Chen, Z., Gafford, B., Gifford, N., Lamsal, A., Niehus-Staab, L.: Deep neural networks for text detection and recognition in historical maps. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 902–909. IEEE (2019)
44. Weinman, J., Gómez Grabowska, A., Karatzas, D.: Counting the corner cases: Revisiting robust reading challenge data sets, evaluation protocols, and metrics. In: 18th International Conference on Document Analysis and Recognition (ICDAR 2024). Lecture Notes in Computer Science, Springer (2024)
45. Ye, M., Zhang, J., Zhao, S., Liu, J., Liu, T., Du, B., Tao, D.: DeepSolo++: Let transformer decoder with explicit points solo for multilingual text spotting. arxiv preprint arXiv:2305.19957 (2023)
46. Ye, M., Zhang, J., Zhao, S., Liu, J., Liu, T., Du, B., Tao, D.: DeepSolo: Let transformer decoder with explicit points solo for text spotting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19348–19357 (2023)

47. Yu, W., Liu, M., Chen, M., Lu, N., Wen, Y., Liu, Y., Karatzas, D., Bai, X.: ICDAR 2023 competition on reading the seal title. arXiv preprint arXiv:2304.11966 (2023)
48. Yujian, L., Bo, L.: A normalized levenshtein distance metric. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(6), 1091–1095 (2007). https://doi.org/10.1109/TPAMI.2007.1078
49. Zhang, Q., Xu, Y., Zhang, J., Tao, D.: ViTAEv2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. International Journal of Computer Vision **131**(5), 1141–1162 (2023)
50. Zhang, X., Su, Y., Tripathi, S., Tu, Z.: Text spotting transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9519–9528 (June 2022)
51. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)

Supplementary Material

This supplementary appendix gives additional details relevant to the competition data, evaluation, and results. Section A describes the data set construction and annotation processes as well as a synopsis of the published versions. Section B details the evaluation protocol used in the competition, which differs in a few subtle but important ways from several related prior competitions. Section C graphically displays the quantitative competition results, while Section D provides a variety of qualitative competition result visualizations for all submissions, tasks, and data sets.

A Data Set Details

A.1 Rumsey Data Set Acquisition

Our procedure for selecting the most representative maps involved training a vision foundation model (e.g., ResNet) in a self-supervised manner with a contrastive learning objective. We further decomposed the cropped map images into patches and trained the model to embed image features such that patches cropped from the same image are close and patches from two distinct images are far away from each other. Subsequently, we clustered the maps into 1,000 groups. This clustering approach enables us to identify the most distinctive maps within the collection. Specifically, maps closest to the center of each cluster can be considered as representing the unique styles encapsulated by that cluster. We then randomly cropped map tiles of size 2K $\times$ 2K pixels and manually removed the tiles with low quality, little to no text, or extensive non-Latin characters.

A.2 Annotation Protocol

All the map annotations are performed by human annotators. For one map image, the annotators were instructed to perform text detection, recognition and linking annotations concurrently. Both data sets share the same ground truth and submission format.

Words Each word instance requires a bounding polygon and the transcription, with all annotations applied at the word level, regardless of inter-character spacing. For instance, if characters such as "L", "A", "K", and "E" are widely spaced, they are still annotated as "LAKE", even when the spacing is several times larger than the height of the character. If a word is cut off at the map crop boundary, we treat it as an *ignore* case in the evaluation metrics, similar to the *difficult* case in previous ICDAR competitions for blurry and tiny text labels. Illegible words are also ignored in the evaluations.

Groups Words belonging to one location phrase or other structural label group are gathered together into a reading-order sequence.

Character Set Because of the worldwide nature of the maps involved, the files are provided in UTF-8 encoding. However, the Latin alphabet (with diacritics), numbers, and punctuation are primarily to be expected.

A.3 Version History

The initial version (1.0) of the training and validation data for both evaluations (Rumsey [30] and IGN French Land Registers [8]) was released on February 2, 2024. As one might expect from a human annotation process with several stages and a wide scope, some issues were discovered by both competition participants and organizers along the way. As a result, the competition data sets underwent a few minor revisions during and shortly after the contest period. Here we summarize the various versions for clarity and transparency.

Rumsey Data Set It was reported that the images in the Rumsey data had been published with reversed color channel order (i.e., BGR format) and revision 1.1 was quickly published on February 19 to correct it [31]. It was subsequently discovered that: i) some images had missing groups (for Task 2 and 4) due to inconsistencies in the raw annotations; ii) several words were incorrectly linked into groups; and iii) some keypoints in the text bounding polygons were not following the correct order. Once again, we quickly addressed these deficiencies and published revision 1.2 on April 23 [32]. Both updates were announced on the contest web site for participants.

After the competition window closed, organizers discovered that some transcription markup (e.g., superscript and subscript tags) had inadvertently remained in the published ground truth. The training data had only one such instance, while twenty-five words were affected in the validation data. A post-competition revision 1.3 has since been published on June 7, which can and should be used for future benchmarking [29]. With only a single training instance affected (among 34.5K words), the impact on competitors and comparisons should be negligible. Importantly, the same corrections were made with the held out ground truth test data used for this report and all evaluations (past and ongoing) on the RRC server. (The difference in test performance typically amounts to a hundreth of a percent on Task 4 character accuracy.)

The public test data contains only images; its single version (1.0) was published March 4 [33].

IGN Data Set It was reported that coordinates of the ground truth polygons sometimes exceeded the dimensions of the accompanying image tile; this behavior was caused by a rounding in the transformation from the original full map image domain to the cropped version. Revision 1.1 was published on April 17 and announced the competition web site [7].

As with the Rumsey data, the public test data contains only images; its single version (1.0) was published March 4 [6].

B Evaluation Protocol

Evaluation follows a traditional one-to-one matching strategy between ground truth words and detected words. That is, the fundamental precursor to calculating the competition metrics is to place ground truth elements (words or groups of words forming a phrase) in correspondence with detected elements. However, this competition protocol differs from most prior RRC protocols in a few key ways:

1. Rather than using a greedy search to pair ground truths and detections, we use a metric-specific full optimization framework for matching.
2. Rather than discount detections that overlap (many-to-one) with "don't care" regions before matching, such detections participate in the match process with only one-to-one matches allowed.
3. Rather than sequentially apply geometric and textual constraints in the end-to-end task, the constraints are applied jointly in the match optimization.

These differences were proposed by Weinman et al. [44], which analyzes their benefits. The remainder of this section lays out the details of the evaluation protocol, which precedes calculation of the competition metrics.

The entire official evaluation code is publicly available at https://github.com/icdar-maptext/evaluation.

B.1 Optimization Framework

Because ground truth polygons may overlap significantly with one another, conventional means for determining the correspondence between predicted detections and ground truth (e.g., greedy [14,10] and self-consistent [35]) can be insufficient [44]. Maintaining the one-to-one matching restriction, we therefore use a full weighted bipartite matching algorithm to determine the maximal set of true positives TP from among valid candidates

$$\mathrm{TP}_{\mathrm{Det}} \subset \{(g,d) \in G \times D \mid \mathrm{IoU}\,(g,d) > 0.5\}\,, \tag{S1}$$

where G is the set of ground truth regions and D is the set of detected regions. The algorithm ensures that when there are multiple correspondence candidates, the true positive assignments maximize the total IoU score [44].

Formally, given a scoring function $\psi : G \times D \to \mathbb{R}$, bipartite linear sum assignment finds the $\mathbf{X} \in \mathbb{Z}_2^{|G|\times|D|}$ with entries x_{gd} maximizing the sum

$$\sum_{(g,d)\in G\times D} \psi\,(g,d)\,x_{gd} \tag{S2}$$

with constraints $\sum_{g \in G} x_{gd} \leq 1$ and $\sum_{d \in D} x_{gd} \leq 1$ for all $d \in D$ and $g \in G$, respectively [44]. Each $x_{gd} = 1$ in the matrix represents a correspondence.

Some ground truth regions are marked as a "don't care" (e.g., because it is illegible or truncated as part of the map tile cropping). Because such regions are annotated at the level of individual words (rather than covering multiple words), these "don't care" words can participate in the optimizing matching process described above. Their participation likewise ensures only one-to-one matches between detections and "don't care" regions. Any such matches are discounted from evaluation; they do not count as true positives. Conversely, unmatched "don't care" ground truth words do not count as false negatives. Any other unmatched detections count as false positives, and unmatched ground truth words count as false negatives. The standard alternative of most previous RRCs allows many-to-one matches with "don't care" regions, which can artificially inflate precision.

B.2 Task-Specific Evaluations

This section details specifics of the evaluation protocol for each task. While all tasks utilize the same basic optimization framework denoted by Eq. (S2), the underlying objects (G and D) and match weights $\psi(g, d)$ may differ across tasks.

Task 1: Word Detection For this task, sets G and D represent individual words. Since PDQ $\triangleq F \times T$ is the competition metric, we want to select correspondences that maximize not only the number of true positives for F, but also their tightness (IoU) for T. To that end, we use the match score function

$$\psi(g, d) = \begin{cases} \text{IoU}(g, d) & \text{if IoU}(g, d) > 0.5 \wedge \neg I(g) \\ \epsilon & \text{if IoU}(g, d) > 0.5 \wedge I(g) \\ -1 & \text{otherwise,} \end{cases} \tag{S3}$$

where predicate $I(g)$ represents whether g is a "don't care" word to be ignored. The very small value of $\epsilon > 0$ allows detections to be matched with "don't care" ground truth items (and later discounted) while still preferring matches with valid ground truths. For scoring with the competition metrics,

$$\text{TP} = \{(g, d) \in G \times D \mid x_{gd} = 1 \wedge \neg I(g)\}. \tag{S4}$$

Using this ψ to determine the correspondences that establish TP has a beneficial effect on final PDQ scores [44].

Task 3: Word Detection and Recognition As in prior RRCs, true positives must have matching transcriptions in addition to meeting the IoU threshold:

$$\text{TP}_{\text{Rec}} \subset \{(g, d) \in G \times D \mid \text{IoU}(g, d) > 0.5 \wedge g_{\text{text}} = d_{\text{text}}\}. \tag{S5}$$

However, unlike prior RRCs we use the optimization framework (described for Task 1) to jointly enforce the two constraints; the traditional sequential processing that prioritizes geometry can mismatch textual correspondences, particularly when significant numbers of detections and ground truth elements overlap [44].

The match score function is thus

$$\psi(g,d) = \begin{cases} \text{IoU}(g,d) & \text{if IoU}(g,d) > 0.5 \wedge g_{\text{text}} = d_{\text{text}} \wedge \neg I(g) \\ \epsilon & \text{if IoU}(g,d) > 0.5 \wedge I(g) \\ -1 & \text{otherwise.} \end{cases} \tag{S6}$$

The resulting set of true positives for metric calculation is also calculated using Eq. S4.

Task 2: Phrase Detection (Word Grouping) For this task, submissions come in the form of individual words grouped into sets corresponding to phrases. Thus, rather than finding correspondences between individual words, the evaluation finds correspondences between these unordered groups.

At a high level, the unions of the word polygons among the words forming predicted and ground truth phrase groups are used to calculate the IoU, and the maximizing correspondences are then found as for Task 1.

More formally, let G_i and D_i represent a set of words in the ground truth and detections, respectively, with $g_{ij} \in G_i$ and $d_{ij} \in D_i$ representing the individual words. We define the geometric regions

$$g_i \triangleq \bigcup_j g_{ij} \quad \forall i \tag{S7}$$

$$d_i \triangleq \bigcup_j d_{ij} \quad \forall i \tag{S8}$$

so that $G = \{g_i\}$ and $D = \{d_i\}$ become the sets over which correspondences are found.

If any word in the ground truth group is marked as a "don't care," the entire group is treated as a "don't care" and handled as described in Task 1:

$$I(g_i) = \begin{cases} 1 & \text{if } \exists j \; I(g_{ij}) \\ 0 & \text{otherwise.} \end{cases} \tag{S9}$$

Task 3 uses the same optimization process (S2) and match score function (S3) for Task 1, but applied to these groups G and D, which are comprised of the combined regions g_i and d_i.

Task 4: Phrase Detection and Recognition As with Task 3, both geometry and strings are used to determine the correspondences between lists of detected words and lists of ground truth words. For geometry, we use the same strategy as Task 2, where unions of the polygons in each list of detected and ground truth

words become input to the IoU calculation. However, because the competition metric PCQ does not require exact string matches, we also use the NED to promote string match quality in determining the correspondences. In the bipartite graph, the edge weight $\psi(g, d)$ between a ground truth group g and detected group d is given by

$$\psi(g,d) = \begin{cases} \mathrm{IoU}(g,d)\,(1 - \mathrm{NED}(g,d)) & \text{if } \mathrm{IoU}(g,d) > 0.5 \wedge \neg I(g) \\ \epsilon & \text{if } \mathrm{IoU}(g,d) > 0.5 \wedge I(g) \\ -1 & \text{otherwise,} \end{cases} \tag{S10}$$

where predicate $I(g)$ represents whether there is a "don't care" word in g to be ignored, as defined by Eq. (S9). Using this ψ to determine the correspondences that establish TP has a beneficial effect on final PCQ scores [44].

C Graphical Submission Ranking

This section features ranking summaries as visual bar plots for fast comparison:

- results for **task 1** are shown in Figure S1,
- results for **task 2** are shown in Figure S2,
- results for **task 3** are shown in Figure S3,
- results for **task 4** are shown in Figure S4.

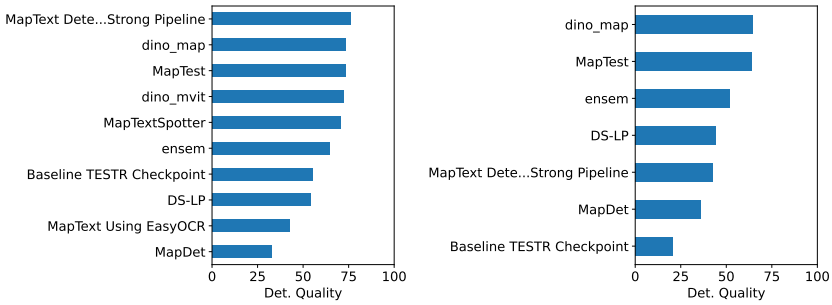


Fig. S1. Final ranking overview for task 1 on the Rumsey *(left)* and French Land Register *(right)* data sets. Methods are sorted by descending Detection Quality (%).

D Example Results

This section provides illustrations of select example predictions from each system on every task and data set. Additional examples are permanently archived at https://zenodo.org/communities/icdar-maptext.

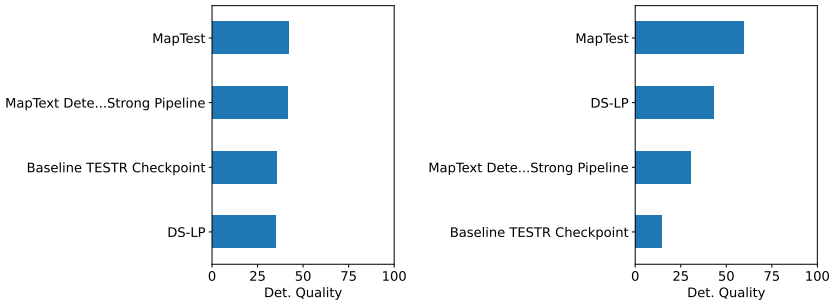


Fig. S2. Final ranking overview for task 2 on the Rumsey *(left)* and French Land Register *(right)* data sets. Methods are sorted by descending Detection Quality (%). Compared to task 1, words are grouped and must be detected as a whole.

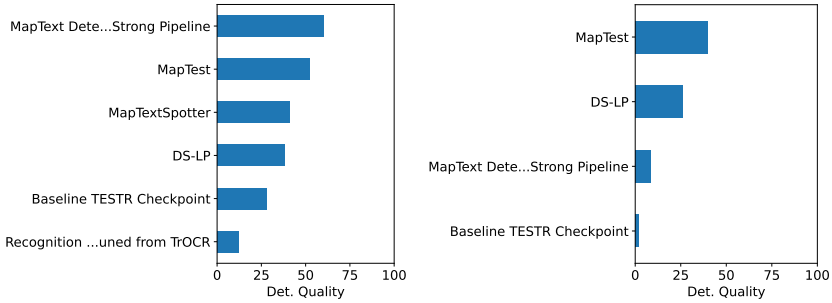


Fig. S3. Final ranking overview for task 3 on the Rumsey *(left)* and French Land Register *(right)* data sets. Methods are sorted by descending Detection Quality (%). Compared task 1, words have to be recognized perfectly to be considered as detected, and only for matching words the detection quality is considered.

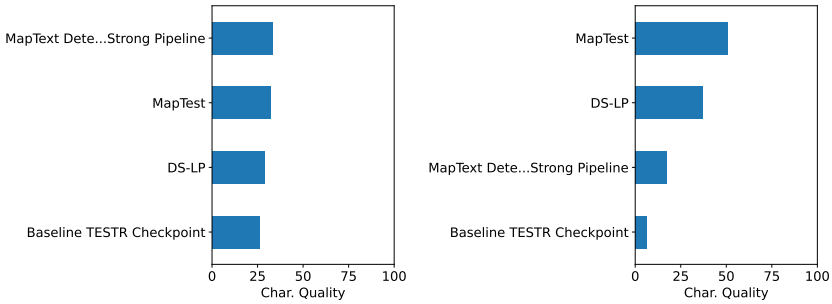


Fig. S4. Final ranking overview for task 4 on the Rumsey *(left)* and French Land Register *(right)* data sets. Methods are sorted by descending Character Quality (%), the product of detection quality and transcription quality (character accuracy). Compared to other tasks, words are grouped and must be detected and recognized as a whole.

Every image contains a comparison of the raw predictions of each submission with the ground truth. Each figure provides examples of map tiles that are easiest, hardest, and randomly selected. For each task, the selection of easy and hard images is based on the submissions' mean performance on the task's main evaluation metric:

Task 1: Panoptic Detection Quality for isolated words (Figures S5 and S6);
Task 2: Panoptic Detection Quality for word groups (Figures S7 and S8);
Task 3: Panoptic Recognition Quality for isolated words, which constrains matches between the ground truth and predictions to have exactly the same transcription (Figures S9 and S10); and
Task 4: Panoptic Character Quality for word groups, which combines both tightness and character-level accuracy with detection quality (Figures S11 and S12).

Refer the main paper for formal definitions of the metrics.

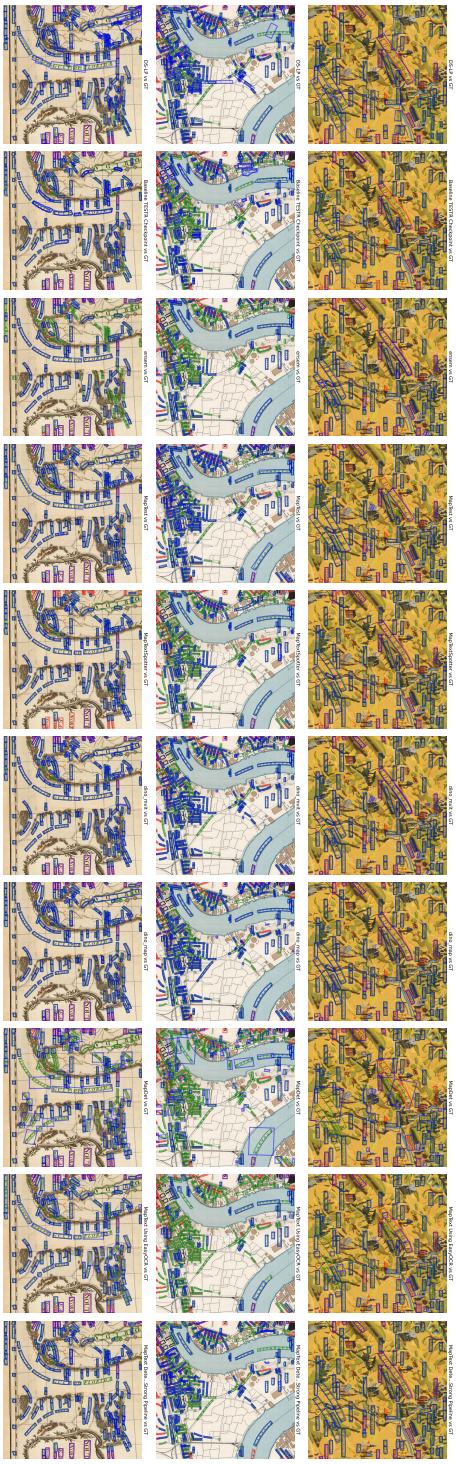


Fig. S5. Example results for Task 1 (word detection) on the Rumsey data set. Blue regions are predictions for the entitled submission. Green indicates valid ground truth words, while red indicates cropped or ignored words. Top–Bottom: Easiest, hardest, random.

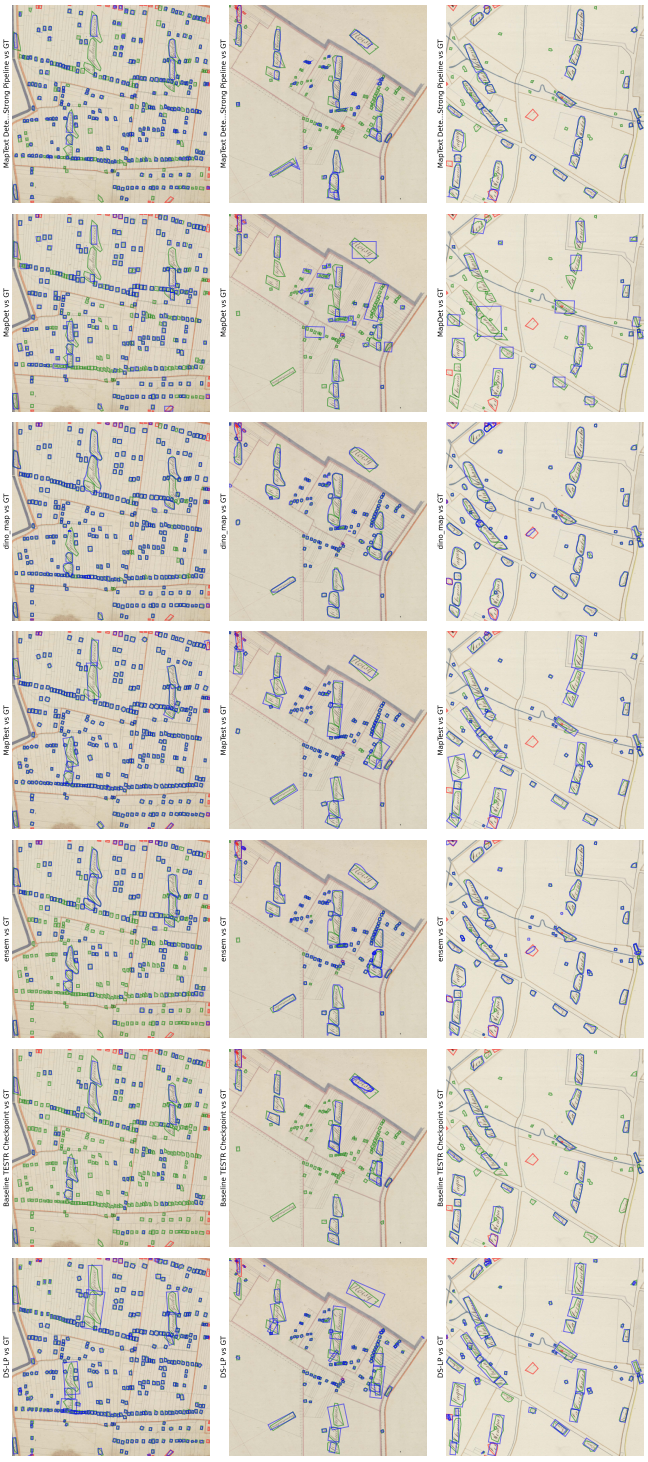


Fig. S6. Example results for Task 1 (word detection) on the French Land Register data set. Blue regions are predictions for the entitled submission. Green indicates valid ground truth words, while red indicates cropped or ignored words. Top–Bottom: Easiest, hardest, random.

Fig. S7. Example results for Task 2 (phrase detection/word grouping) on the Rumsey data set. Phrase groups have the same color with links drawn between successive group members. Top–Bottom: Easiest, hardest, random.

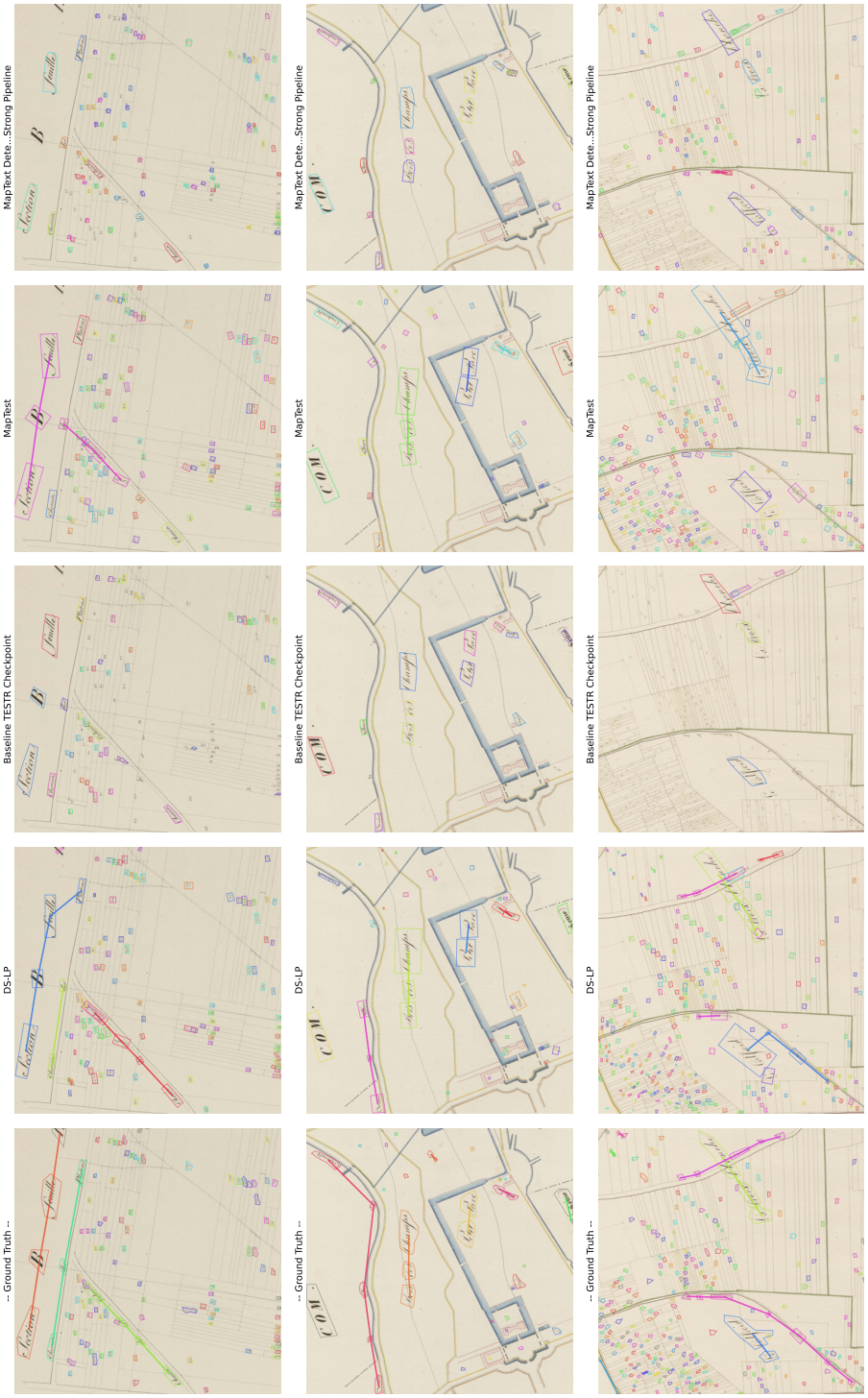


Fig. S8. Example results for Task 2 (phrase detection/word grouping) on the French Land Register data set. Phrase groups have the same color with links drawn between successive group members. Top–Bottom: Easiest, hardest, random.

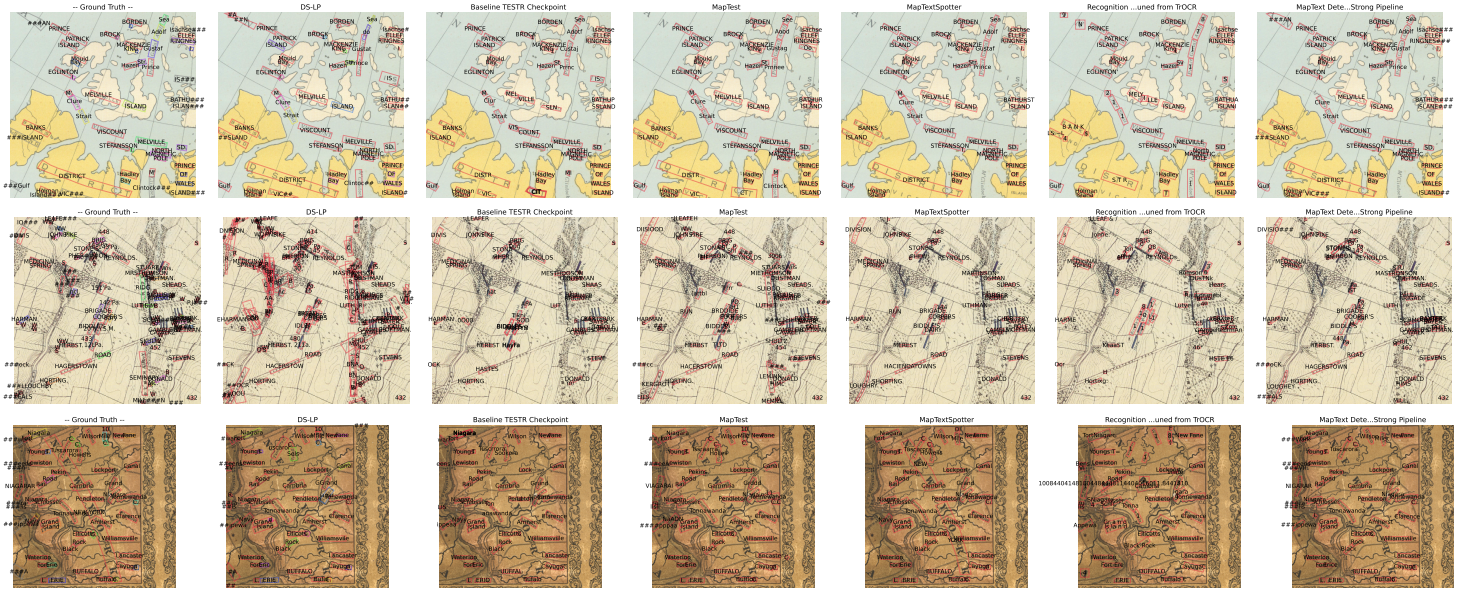


Fig. S9. Example results for Task 3 (word detection and recognition) on the Rumsey data set. Phrase groups have the same color with overlaid transcription. (Ground truth transcriptions including "###" indicate an ignored word.) Top–Bottom: Easiest, hardest, random.

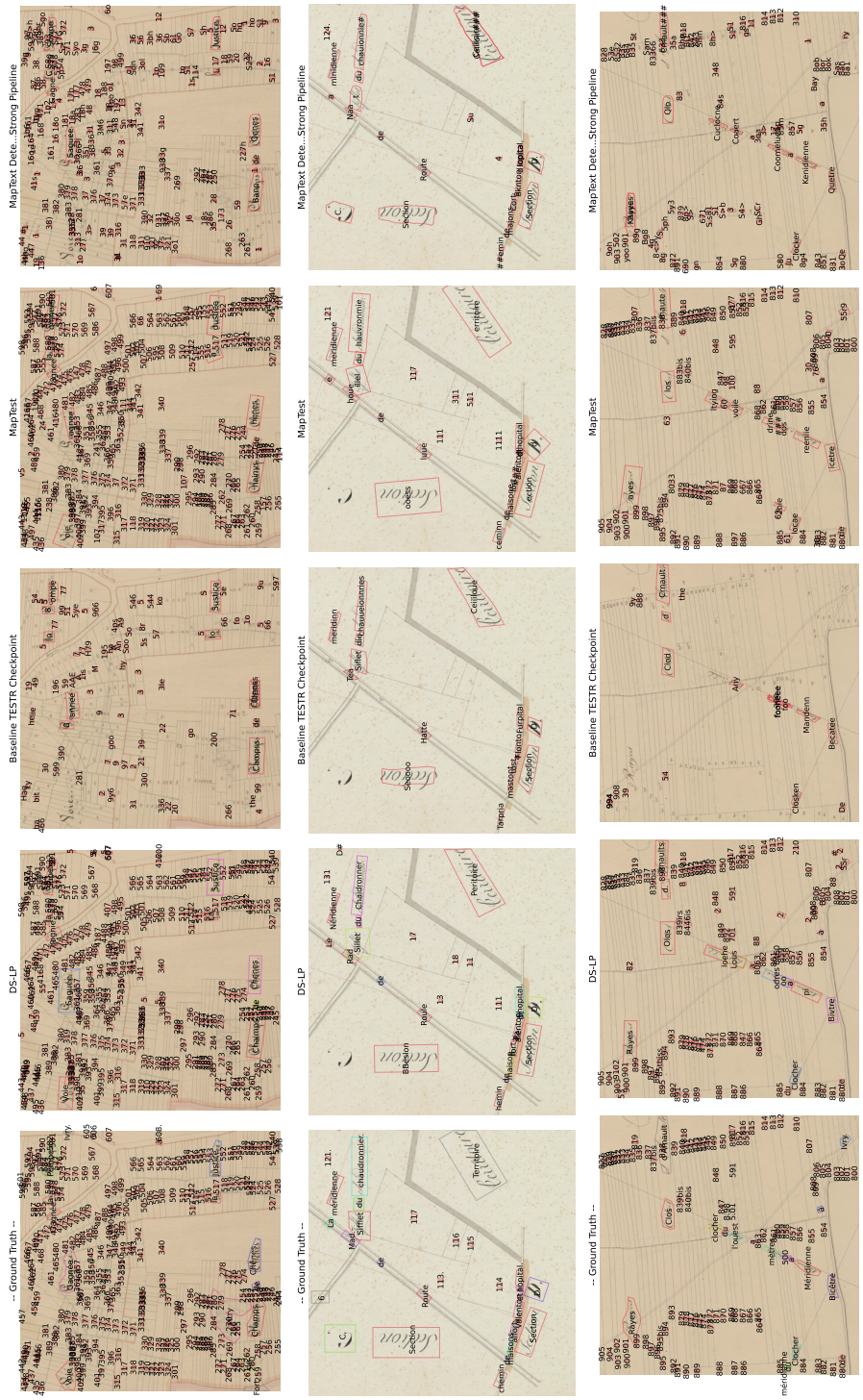


Fig. S10. Example results for Task 3 (word detection and recognition) on the French land register data set. Phrase groups have the same color with overlaid transcription. Top–Bottom: Easiest, hardest, random.

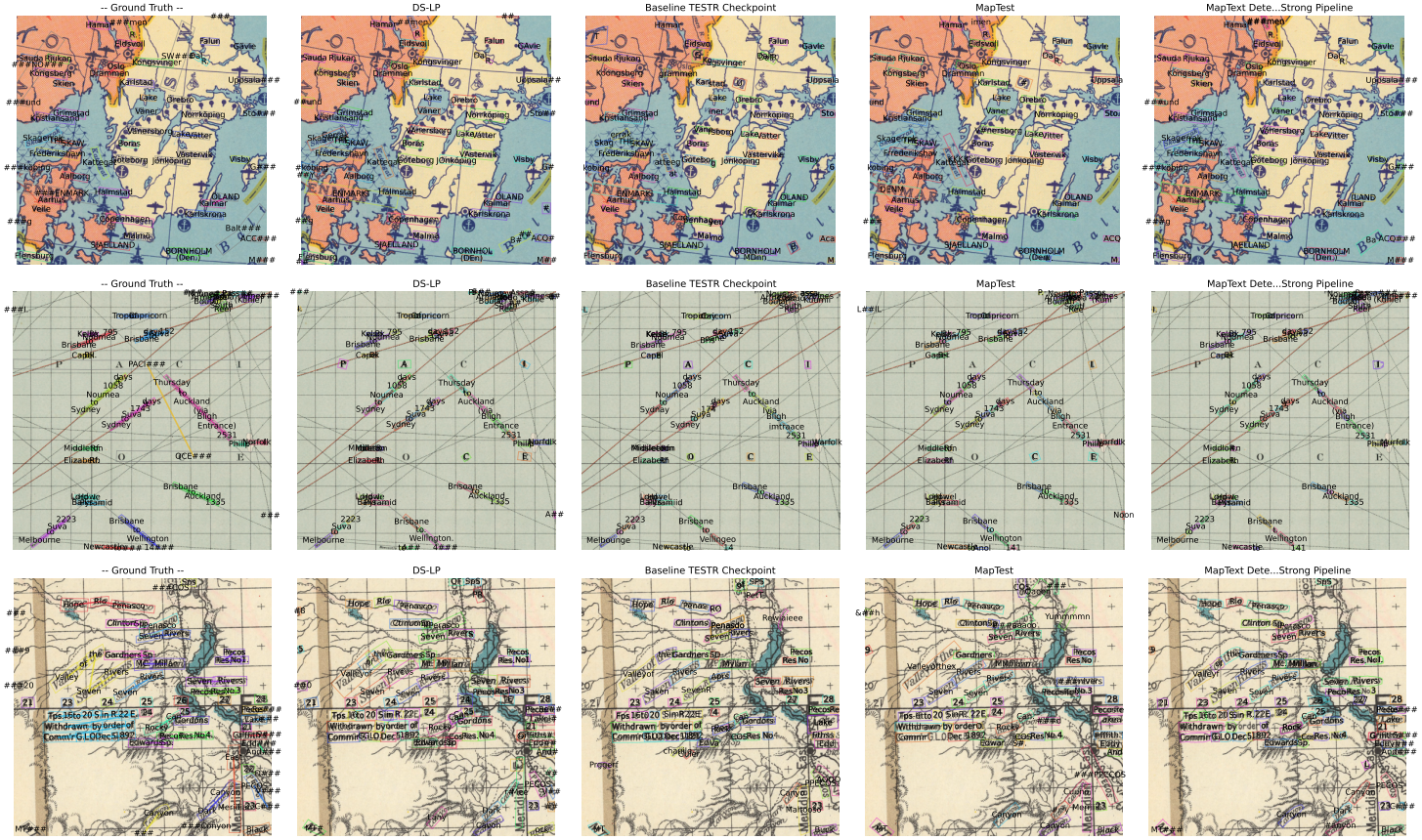


Fig. S11. Example results for Task 4 (phrase detection and recognition) on the Rumsey data set. Phrase groups have the same color with overlaid transcription. (Ground truth transcriptions including "###" indicate an ignored word.) Top–Bottom: Easiest, hardest, random.

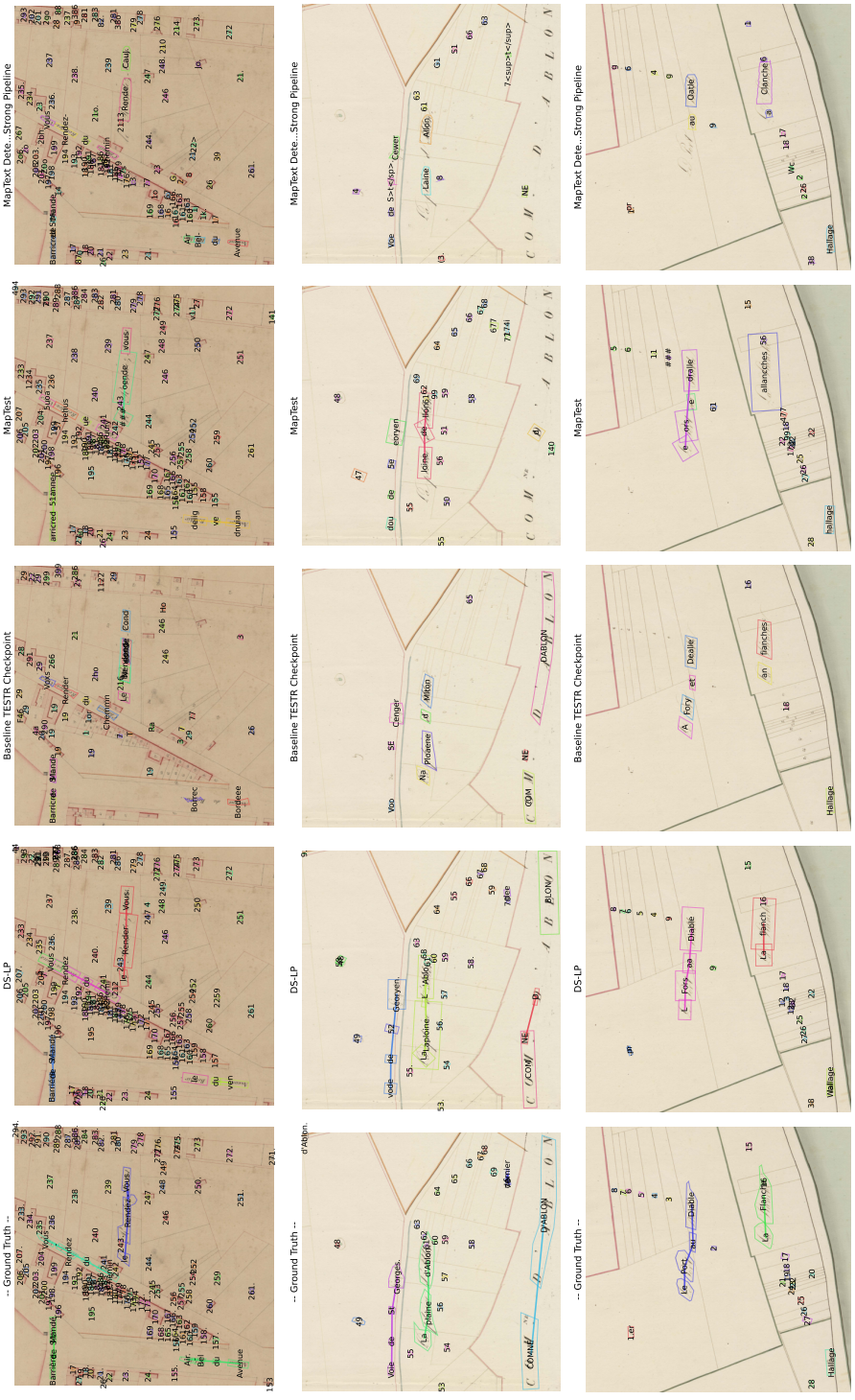


Fig. S12. Example results for Task 4 (phrase detection and recognition) on the French land register data set. Phrase groups have the same color with overlaid transcription. Top–Bottom: Easiest, hardest, random.