HAND-PUNCHED BRAILLE DOCUMENT RECOGNITION: A NEW BENCHMARK DATASET AND A ROBUST DETECTION TRANSFORMER

Anonymous authors
Paper under double-blind review

ABSTRACT

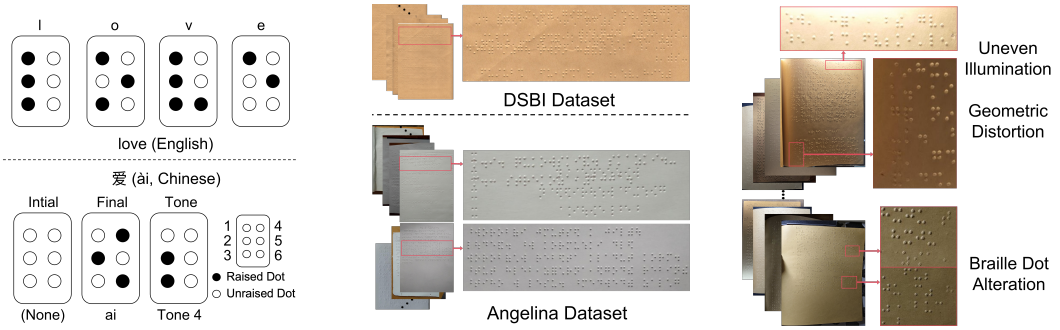
Optical Braille recognition plays an important role in digitizing Braille documents and supporting accessible learning for blind and visually impaired learners. However, existing page-level Braille benchmarks provide limited coverage of smartphone-captured hand-punched documents with Braille dot appearance variation, which normally exhibit illumination variation and geometric distortion. To address this data gap, we introduce BrailleStuDoc, a benchmark dataset of student-hand-punched Braille documents captured with smartphones. It contains 331 images from distinct physical pages and 149,720 cell-level annotations covering all 63 non-empty six-dot patterns. To jointly address degraded local appearance cues and geometric variations in smartphone-captured hand-punched Braille documents, we propose BRIDGE, an illumination- and geometry-aware Braille cell detection Transformer. BRIDGE introduces Micro-Surface Feature Enhancement (MSFE) to strengthen local appearance representation under illumination and dot appearance variations, and Geometry-Aware Scale Modeling (GASM) to model the local geometric regularity of neighboring Braille cells under geometric distortion. Extensive experiments demonstrate the complementary value of BrailleStuDoc to existing public datasets and the effectiveness of BRIDGE, which achieves the best overall recognition performance among the compared methods.

1 INTRODUCTION

Braille is an essential medium for blind and visually impaired learners to access written information, complete assignments, and communicate in educational settings Croake et al. (2025); Rose et al. (2025). Optical Braille recognition converts paper-based Braille into machine-readable content, supporting document digitization, assignment review, accessible learning services, and more equitable access to educational resources Hoskin et al. (2024). As illustrated in Fig. 1(a), six-dot Braille follows a 3×2 cell structure, with the interpretation of each dot pattern determined by the underlying Braille system Kana & Hagos (2025). Existing optical Braille recognition methods have achieved promising performance on scanned documents and images acquired under relatively controlled conditions Li et al. (2019; 2020); Ovodov (2021). With the widespread availability of smartphones, Braille documents can be conveniently captured in everyday learning environments. However, practical document capture is often affected by illumination variation, shadows, and geometric distortion Jiang et al. (2022); Li et al. (2023b;c); Zhang et al. (2024), while hand-punched Braille may further exhibit dot appearance variation caused by irregular punching, wear, or correction remnants. These factors alter local visual cues and apparent cell geometry, complicating Braille cell localization and six-dot pattern recognition.

Despite the practical relevance of recognizing student-produced Braille documents captured with smartphones, existing page-level benchmarks remain limited in representing such documents from authentic educational settings. More specifically, three major gaps remain. 1) Publicly available page-level Braille document datasets are extremely limited. To the best of our knowledge, only DSBI Li et al. (2018) and Angelina Ovodov (2021) fall into this category. DSBI contains 114 scanned double-sided Chinese Braille images Li et al. (2018), while Angelina provides 240 photographed Braille pages, only a subset of which consists of student-produced work Ovodov (2021). 2) Existing page-level benchmarks provide limited coverage of the joint occurrence of illumina-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Six-dot Braille encoding

(b) Existing Braille document datasets

(c) The proposed BrailleStuDoc dataset

Figure 1: (a)Illustration of Braille Encoding Rule. (b,c) Comparison between existing Braille Document datasets and our newly proposed BrailleStuDoc dataset.

Table 1: Comparison of representative page-level Braille document datasets.

Dataset	DSBI	Angelina	BrailleStuDoc
Language	Chinese	Russian	Chinese
Images	114	240	331
Image acquisition	Scanner	Camera /Smartphone	Smartphone
Hand-punched	No	Partial	Yes
Illumination variation	No	Not reported	Yes
Perspective variation	No	Yes	Yes
Cell-level annotations	Yes	Yes	Yes

tion variation and geometric distortion in handheld smartphone capture. 3) Braille dot appearance variations commonly observed in hand-punched student work, including irregular punching, wear, and correction remnants, remain underrepresented in existing benchmarks. Consequently, current datasets provide limited coverage of the combined illumination, geometric, and dot appearance variations encountered in smartphone-captured hand-punched Braille documents, as illustrated in Fig. 1(b)–(c).

To address these limitations, we construct BrailleStuDoc, a benchmark dataset of student-hand-punched Chinese Braille documents captured with smartphones. It comprises 331 images, each from a distinct physical page, and 149,720 cell-level annotations covering all 63 non-empty six-dot patterns. Captured using multiple smartphones under everyday indoor conditions, the images exhibit illumination variation, geometric distortion, and Braille dot appearance variation. As shown in Table 1, BrailleStuDoc contains the largest number of images among the compared datasets and is the only one composed entirely of hand-punched documents. Therefore, BrailleStuDoc can better support the evaluation of optical Braille recognition models under practical smartphone-capture conditions, particularly for hand-punched Braille documents.

Our analysis of BrailleStuDoc highlights two key modeling challenges for smartphone-captured, hand-punched Braille documents. First, illumination variation and variations in Braille-dot appearance can weaken the local visual cues of raised dots. Their appearance relies on subtle shading and edge cues, which may be degraded by spatially varying illumination, wear, and irregular punching. Second, geometric distortion changes the apparent scale and spatial arrangement of Braille cells across the page, while nearby cells typically remain approximately aligned along local rows and columns and exhibit smoothly varying scales. These characteristics require both robust local appearance representation and effective modeling of inter-cell geometry. DETR-style detectors provide a strong foundation through end-to-end set prediction and advances in query formulation and encoder design Carion et al. (2020); Liu et al. (2022); Zhang et al. (2023a); Li et al. (2023a). In particular, RT-DETRv2 Lv et al. (2024) provides an efficient end-to-end framework with selective multi-scale feature extraction, making it a suitable foundation for page-level Braille recognition. Nevertheless, as a general-purpose detector, RT-DETRv2 does not explicitly model illumination-sensitive Braille-dot appearance or the structured geometric relationships among neighboring Braille cells.

Accordingly, we propose BRIDGE, an illumination- and geometry-aware Braille cell detection framework built upon RT-DETRv2 Lv et al. (2024). BRIDGE augments RT-DETRv2 with two task-specific modules to jointly predict the bounding box and dot-pattern class of each Braille cell. **Micro-Surface Feature Enhancement (MSFE)** is incorporated into multi-scale feature aggregation to enhance subtle Braille structures by integrating illumination-aware responses, spatial reliability, and local high-frequency information. **Geometry-Aware Scale Modeling (GASM)** is introduced into the last three decoder layers to exploit the local row–column organization and scale consistency of neighboring Braille cells for refining query representations and reference-box scales. Together, MSFE and GASM enhance local appearance representation and model Braille-specific geometric variation while preserving the rapid end-to-end detection paradigm of RT-DETRv2. Extensive experiments on the proposed BrailleStuDoc dataset and representative existing Braille document benchmarks validate the effectiveness of BRIDGE.

Our main contributions are summarized as follows:

- We introduce a new benchmark dataset for optical Braille recognition, termed BrailleStuDoc, which contains 331 smartphone-captured images from distinct physical pages and 149,720 cell-level annotations covering all 63 non-empty six-dot patterns. The dataset consists of student-hand-punched Chinese Braille documents and captures illumination variation, geometric distortion, and Braille dot appearance variation encountered in practical smartphone capture.
- We propose a new illumination- and geometry-aware Braille cell detection framework, termed BRIDGE, based on RT-DETRv2. By incorporating Micro-Surface Feature Enhancement (MSFE) and Geometry-Aware Scale Modeling (GASM), BRIDGE enhances local Braille appearance representation and exploits the geometric regularity of neighboring Braille cells for robust cell localization and dot-pattern classification.
- Extensive experiments on BrailleStuDoc and two existing Braille document datasets demonstrate the effectiveness of BRIDGE and the value of the proposed dataset. BRIDGE achieves the best F_1 score of 99.50 under BrailleStuDoc training and the highest F_1 scores across all three datasets under joint training.

2 Related Work

Recent advances in optical Braille recognition have been driven by the development of deep learning-based methods and increasingly diverse Braille image datasets. In this section, we briefly review the datasets and recognition methods closely related to this task.

2.1 Optical Braille Image Datasets

Existing optical Braille image datasets can be broadly categorized by input granularity into character-, line-, page-, and scene-level groups. Since this work focuses on page-level Braille document recognition, we primarily review page-level datasets and briefly summarize representative resources at the other granularities.

Among page-level datasets, DSBI contains 114 color images of double-sided Braille documents acquired at 200 dpi with a flatbed scanner, with annotations for recto and verso dots as well as Braille cells Li et al. (2018). Angelina extends page-level recognition to photographed Braille documents, comprising 212 pages from double-sided Braille books and 28 pages of student work captured using cameras or mobile phones. The dataset includes curved pages and perspective distortion, with Braille cells annotated by bounding boxes and class labels corresponding to six-dot patterns Ovodov (2021). More recently, Yamashita et al. constructed a dataset from 120 pages of the *Tenji Mainichi* Braille newspaper and 126 pages of the Japanese Braille magazine *Shikaku Shogai*, with each page captured under 12 controlled lighting directions Yamashita et al. (2026). However, the dataset has not been publicly released.

Representative character-level datasets include the Handwritten Braille–Omniglot dataset, the Braille Character Dataset, and the Arabic Braille dataset, which consist of isolated Braille character images for classification Elaraby et al. (2024). At the line level, Am-Br-line-image contains 2,100 Amharic Braille line images with corresponding Amharic text annotations Asfaw et al. (2024).

Scene-level datasets focus on Braille embedded in surrounding environments. NSBD contains 212 images collected from the Internet or captured using mobile phones, covering scenes such as walls, cards, and elevator buttons, with bounding-box annotations for Braille characters Lu et al. (2022). NSBID further extends scene-level coverage with 554 original images collected from mobile phones and online sources across 12 complex background types for Braille segment detection Lu et al. (2023).

Despite the growing diversity of existing datasets, student-produced material remains scarce in page-level benchmarks, and the combined variations in illumination, page geometry, and Braille dot appearance encountered in practical smartphone capture are still underrepresented.

2.2 Optical Braille Recognition Methods

Existing optical Braille recognition methods can be broadly categorized into dot-based recognition, segmentation-based cell recognition, detection-based cell recognition, and line-level sequence recognition.

Dot-based methods localize individual Braille dots and reconstruct cells from their spatial arrangement. TS-OBR detects dots for grid construction and applies SVM classification at candidate locations Li et al. (2019), while BddNet uses Gaussian heatmaps for dot-center localization followed by spatial clustering and grid recovery Meng et al. (2024). Segmentation-based methods recover Braille cell regions from page images. BraUNet formulates page-level recognition as multi-class semantic segmentation with auxiliary foreground segmentation and morphological post-processing Li et al. (2020). Zemouri et al. combine projection-based cell segmentation and spatial partitioning with subsequent classification Zemouri et al. (2026).

Detection-based methods treat complete Braille cells as object instances and jointly predict their locations and classes from Braille document images without explicit cell segmentation. Ovodov adapted RetinaNet for Braille cell detection on photographed pages with curvature and perspective distortion Ovodov (2021), while Lu et al. introduced an anchor-free detector with edge features for Braille character detection in natural-scene images Lu et al. (2022). Yamashita et al. employed YOLOX for Braille cell detection and evaluated its robustness under systematically varied lighting directions Yamashita et al. (2026). Beyond existing Braille detectors, transformer-based detectors such as DETR Carion et al. (2020), DN-DETR Li et al. (2022) ,H-DETR Jia et al. (2023) , Lite DETR Li et al. (2023a) ,DDQ Zhang et al. (2023b) , DEIM Huang et al. (2025), RT-DETRZhao et al. (2024), and RT-DETRv2 Lv et al. (2024) provide efficient end-to-end detection via global context modeling, making them a promising option for detecting densely arranged and visually similar Braille cells.

At a higher recognition granularity, line-level methods directly map Braille text-line images to character sequences without explicit character segmentation. Asfaw et al. combine CNN-based visual feature extraction with BiLSTM sequence modeling and CTC decoding for Amharic Braille line recognition Asfaw et al. (2024). Despite recent progress, joint modeling of illumination variation, geometric distortion, and Braille dot appearance variations remains limited, particularly for smartphone-captured Braille document recognition.

3 Dataset

We introduce BrailleStuDoc, a benchmark for recognizing hand-punched Chinese Braille documents captured with smartphones. Unlike existing datasets dominated by scanned documents or images acquired under controlled conditions, BrailleStuDoc focuses on student work and preserves the imaging and writing variations encountered in real-world smartphone capture.

3.1 Dataset Collection

BrailleStuDoc was collected from real-world Chinese Braille assignments and practice materials produced by visually impaired students at Hunan Vocational School of Special Education. All documents are single-sided and hand-punched. Data collection was conducted with authorization from the school, and personally identifiable information was removed or masked before data release. Images were captured using six smartphones running iOS or Android, primarily in indoor environ-

ments under natural or artificial illumination. No flash was used during acquisition. The collection includes near-frontal and oblique views. The dataset contains 331 images originating from 331 distinct physical pages, with exactly one image collected from each page. Each page is assigned a unique identifier to support page-level data partitioning and prevent cross-split duplication.

3.2 Annotation Protocol

Each six-dot Braille cell is treated as an annotation instance and assigned a bounding box and a dot-pattern class. Each bounding box covers the full spatial region of the corresponding six-dot Braille cell, including both raised and unraised dot positions. Following the standard six-dot Braille numbering convention, positions 1–3 correspond to the left column from top to bottom, and positions 4–6 to the right column. Let $z_k \in \{0, 1\}$ indicate whether a raised dot is present at position k. We encode the dot-pattern class as

$$c = \sum_{k=1}^{6} z_k 2^{k-1}, \qquad c \in \{1, \ldots, 63\}, \tag{1}$$

yielding 63 possible non-empty six-dot patterns. Blank cells with no raised dots are not treated as detection instances. Annotations are stored in COCO format Lin et al. (2014). For cells whose dot patterns are obscured or altered by wear, irregular punching, stains, or correction remnants, the class label is determined from the dot configuration confirmed on the physical page. When the dot pattern cannot be reliably determined from the image alone, the original paper document is re-examined and confirmed through tactile inspection.

The annotation workflow combines manual labeling with model-assisted proposal generation. A subset of images was first manually annotated and used to train an initial detector. The detector was then applied to the remaining images to generate candidate bounding boxes and class labels. These predictions were used only as annotation proposals. Annotators manually reviewed the proposed annotations, corrected inaccurate bounding boxes and class labels, removed erroneous proposals, and added missed Braille cells. All annotations were reviewed by a second annotator. Disagreements and ambiguous cases were further reviewed by an expert with professional knowledge of Braille. Instances that remained unresolved after expert review were excluded from the annotations.

3.3 Dataset Statistics

BrailleStuDoc comprises 331 document images with 149,720 annotated Braille cells spanning all 63 non-empty six-dot patterns, averaging approximately 452 cells per image. We use a fixed page-level split into training, validation, and test sets at a ratio of 7:2:1. The class distribution is highly imbalanced. Class 1 is the most frequent, with 11,203 instances, whereas Class 40 contains only 11 instances. Such class imbalance poses an additional challenge for recognizing low-frequency dot patterns.

The major visual challenges in BrailleStuDoc can be broadly grouped into three categories including illumination variation, geometric distortion, and Braille dot appearance variation. Among the 266 images exhibiting illumination variation, 190 images exhibit uneven illumination, 99 contain shadows, 49 contain surface reflections, and 31 are captured under low-light conditions. For geometric distortion, 77 images exhibit noticeable perspective distortion. For Braille dot appearance variation, 66 images exhibit wear, irregular punching, stains, or correction remnants. These conditions are not mutually exclusive, and multiple conditions may occur in the same image.

4 Method

4.1 Overview

We propose BRIDGE, an illumination- and geometry-aware detection framework for recognizing Braille cells in smartphone-captured documents. As shown in Fig. 2, BRIDGE is built upon RT-DETRv2, while introducing two task-specific modules, i.e., Micro-Surface Feature Enhancement (MSFE) and Geometry-Aware Scale Modeling (GASM), to address the appearance degradation and geometric variation commonly observed in BrailleStuDoc. Specifically, an input image is first processed by the backbone to extract multi-scale visual features, which are subsequently fused in the

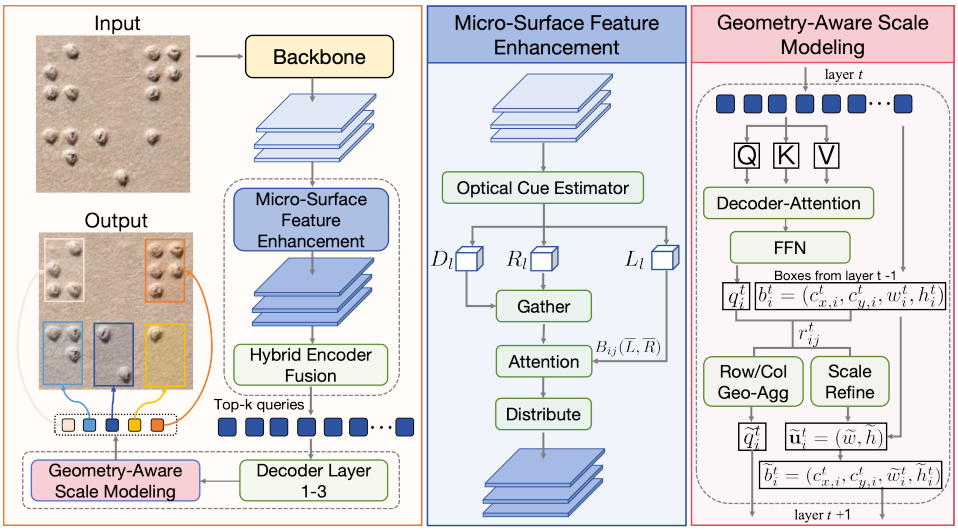


Figure 2: Overall framework of the proposed BRIDGE approach.

Hybrid Encoder. MSFE is embedded into the multi-scale feature aggregation process to enhance subtle Braille structures under spatially varying illumination by jointly exploiting illumination-aware responses, spatial reliability, and local high-frequency information. The enhanced features are then passed to the transformer decoder. In the last three decoder layers, GASM further exploits the local row-column arrangement and scale consistency of neighboring Braille cells to refine query representations and reference-box scales. Finally, the decoder predicts the bounding box and six-dot pattern class of each Braille cell. MSFE and GASM introduce no additional supervision and are jointly optimized with the detector using only Braille cell class and bounding-box annotations. The following sections describe the two modules in detail.

4.2 Micro-Surface Feature Enhancement

In RGB images, low-relief Braille dots are mainly characterized by subtle local edges, shading transitions, and surface-texture cues. These cues can be weakened or distorted by uneven illumination, specular reflections, low contrast, and variations in hand-punched embossing. Directly amplifying high-frequency responses may also enhance paper texture and imaging artifacts. To address this, we introduce MSFE, which preserves local structural details while learning spatial reliability to guide multi-scale feature aggregation.

For each scale of the Hybrid Encoder, MSFE derives three complementary feature cues including an illumination-aware response L_l, a spatial reliability map R_l, and a local high-frequency residual D_l. D_l is injected into the multi-scale gather process to preserve local structural responses, while R_l modulates the contribution of different spatial locations during aggregation. L_l and R_l from different scales are spatially aligned and averaged to obtain $\overline{L}$ and $\overline{R}$, which are then used to guide global self-attention through an additive bias,

$$B_{ij} = \gamma_a \,\mathrm{Clamp}\Big(-\frac{|\overline{L}_i - \overline{L}_j|}{\tau} + \lambda_R \log\big(\max(\overline{R}_j, \epsilon)\big), -4, 1\Big), \tag{2}$$

where i and j index the target and source spatial locations, τ and λ_R control the contributions of illumination compatibility and source reliability, respectively, ϵ ensures numerical stability, and γ_a is a learnable modulation coefficient.

The aggregated context is subsequently redistributed to each feature scale through reliability-dependent gating based on R_l. Locations with lower reliability receive stronger contextual enhancement, complementing the reliability-aware weighting used during feature aggregation. The newly introduced modulation coefficient are initialized to zero, allowing MSFE to start from the original multi-scale aggregation pathway and progressively learn task-specific enhancement during end-to-end training.

4.3 GEOMETRY-AWARE SCALE MODELING

Perspective projection and local page deformation cause spatially varying Braille cell geometry. To reduce the influence of unstable early-stage predictions on geometric modeling, we introduce GASM, which exploits the local row-column and scale regularities of neighboring cells to refine both detection queries and reference-box scales.

At decoder layer t, let q_i^t and $b_i^t = (c_{x,i}^t, c_{y,i}^t, w_i^t, h_i^t)$ denote the i-th query and its reference box. The relative geometry between queries i and j is encoded as

$$r_{ij}^t = \left[\frac{c_{x,j}^t - c_{x,i}^t}{w_i^t}, \frac{c_{y,j}^t - c_{y,i}^t}{h_i^t}, \log \frac{w_j^t}{w_i^t}, \log \frac{h_j^t}{h_i^t} \right]. \tag{3}$$

Based on the preliminary classification confidence scores and reference-box centers at layer t, GASM selects candidate neighbors and constructs row- and column-oriented neighborhoods $\mathcal{N}_{i,\mathrm{row}}^t$ and $\mathcal{N}_{i,\mathrm{col}}^t$. Neighboring query features are aggregated together with their relative geometry:

$$\begin{aligned} m_{i,d}^t &= \sum_{j \in \mathcal{N}_{i,d}^t} a_{ij,d}^t \phi_d\big([q_j^t; g(r_{ij}^t)]\big), \quad d \in \{\mathrm{row}, \mathrm{col}\}, \\ \widetilde{q}_i^t &= q_i^t + \gamma_{\mathrm{row}} m_{i,\mathrm{row}}^t + \gamma_{\mathrm{col}} m_{i,\mathrm{col}}^t, \end{aligned} \tag{4}$$

where $a_{ij,d}^t$ denotes a normalized weight determined by classification confidence and direction-specific spatial proximity, $g(\cdot)$ encodes relative geometry, and $\phi_d(\cdot)$ transforms the geometry-conditioned neighbor feature. The coefficients γ_{row} and γ_{col} control the residual contributions from the two directional neighborhoods.

Using the same local geometric context, GASM further refines the scale of the reference box. Let $\mathbf{u}_i^t = (w_i^t, h_i^t)$ denote the the scale of the i-th reference box including the width and height at decoder layer t, and $\mathcal{N}_{i,s}^t$ the local scale neighborhood constructed from valid neighboring candidates. A neighborhood scale estimate is first obtained and then used to adjust the current scale with a bounded residual update:

$$\overline{\mathbf{u}}_i^t = \sum_{j \in \mathcal{N}_{i,s}^t} \beta_{ij}^t \mathbf{u}_j^t, \quad \widetilde{\mathbf{u}}_i^t = \mathrm{Clamp}_{[\varepsilon,1]} \left[\mathbf{u}_i^t + \mathrm{Clamp}_{[-\delta,\delta]} \left(\boldsymbol{\eta} \odot (\overline{\mathbf{u}}_i^t - \mathbf{u}_i^t) \right) \right], \tag{5}$$

where β_{ij}^t is a normalized weight determined by local geometry and classification confidence, $\boldsymbol{\eta}$ is a learnable scale coefficient, δ bounds the element-wise update, and ε prevents degenerate scales. If no valid scale estimate is available, the current scale is retained. The refined scale $\widetilde{\mathbf{u}}_i^t = (\widetilde{w}_i^t, \widetilde{h}_i^t)$ is combined with the unchanged center coordinates to form the refined reference box $\widetilde{b}_i^t = \left(c_{x,i}^t, c_{y,i}^t, \widetilde{w}_i^t, \widetilde{h}_i^t \right)$.

The $\widetilde{q}_i^t$ and $\widetilde{b}_i^t$ are then used for the current decoder layer prediction. The preliminary classification confidence scores and reference boxes used for neighborhood construction are detached from gradient computation, and the current predicted box is detached before serving as the reference box for the next decoder layer.

5 EXPERIMENTS

5.1 DATASETS AND EVALUATION METRICS

We conduct a comprehensive evaluation of BRIDGE against four representative Braille recognition methods, including BraUNet Li et al. (2020), RetinaNet Ovodov (2021), BddNet Meng et al. (2024), and YOLOX Yamashita et al. (2026), as well as its base detector, RT-DETRv2 Lv et al. (2024). Unlike the other methods, BddNet performs dot-level detection rather than direct Braille cell detection. We therefore reconstruct Braille cell predictions from its detected dots following the procedure described in the original paper and evaluate them using the same cell-level protocol. Experiments are conducted on three page-level Braille document datasets, including two publicly available datasets, DSBI Li et al. (2018) and Angelina Ovodov (2021), and our newly constructed BrailleStuDoc.

Table 2: Comparison with public methods on our datasets.

Method	Overall				Illumination Variation			Geometric Distortion			Dot Appearance Variation		
	Precision	Recall	F_1	Latency (s)	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
BddNet	26.69	9.90	14.44	3.0300	19.60	7.02	10.34	25.79	8.81	13.13	33.49	14.27	20.01
BraUNet	97.03	98.98	98.00	0.1660	96.82	98.83	97.81	96.74	98.93	97.82	97.35	99.13	98.23
RetinaNet	97.93	98.48	98.21	0.1890	97.68	98.38	98.03	97.65	98.43	98.04	98.31	98.66	98.49
YOLOX	97.92	97.01	97.46	0.0840	97.64	96.99	97.31	97.87	96.84	97.35	98.04	97.18	97.61
RT-DETRv2	98.39	98.07	98.23	**0.0258**	98.37	98.16	98.26	98.29	97.73	98.01	98.47	97.99	98.23
BRIDGE	**99.41**	**99.59**	**99.50**	0.0323	**98.94**	**99.43**	**99.18**	**98.94**	**99.38**	**99.16**	**99.08**	**99.54**	**99.31**

Table 3: Comparison with public methods on public dataset.

Train Dataset	Method	DSBI			Angelina			BrailleStuDoc		
		Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
DSBI +Angelina	BddNet	77.05	79.71	78.36	18.48	17.49	17.97	13.41	16.33	14.73
	BraUNet	89.51	94.76	92.06	96.12	97.58	96.84	62.98	85.19	72.42
	RetinaNet	96.55	97.40	96.97	94.68	93.24	93.95	91.04	89.04	90.03
	YOLOX	98.47	97.33	97.90	96.53	96.34	96.43	89.10	77.00	82.61
	RT-DETRv2	98.54	97.94	98.24	97.96	97.39	97.67	95.42	92.04	93.70
	BRIDGE	99.72	99.62	99.67	99.11	99.19	99.15	94.80	93.61	94.20
DSBI + Angelina +BrailleStuDoc	BddNet	78.79	82.24	80.48	23.54	22.60	23.06	17.19	16.96	17.07
	BraUNet	90.91	94.22	92.54	98.68	99.49	99.09	98.08	99.15	98.61
	RetinaNet	99.24	99.08	99.16	97.04	97.49	97.27	97.75	97.69	97.72
	YOLOX	98.51	98.82	98.66	98.28	98.38	98.33	97.95	96.51	97.23
	RT-DETRv2	99.06	98.77	98.92	98.01	98.17	98.09	98.54	98.77	98.65
	BRIDGE	99.58	99.69	**99.63**	99.75	99.47	**99.61**	99.33	99.54	**99.43**

Performance is evaluated at the Braille cell level using Precision, Recall, and F_1-score. Detailed experimental protocols, evaluation criteria, and implementation settings are described in our supplementary materials.

5.2 Comparison on Public OBR document Benchmarks

• **Result on BrailleStuDoc Dataset.** As shown in Table 2, we compare BRIDGE with the other methods on BrailleStuDoc in terms of overall recognition, inference efficiency, and robustness across the three major challenging subsets summarized in Sec. 3.3. BRIDGE achieves the best overall Precision, Recall, and F_1 of 99.41, 99.59, and 99.50, respectively. Compared with RT-DETRv2, BRIDGE improves these three metrics by 1.02/1.52/1.27. It also improves F_1 over BraUNet and RetinaNet by 1.50 and 1.29. BRIDGE introduces a small inference overhead over RT-DETRv2 (0.0323 s vs. 0.0258 s), while remaining faster than BraUNet, RetinaNet, and YOLOX. For the three challenging subsets corresponding to illumination variation, geometric distortion, and Braille dot appearance variation, BRIDGE achieves F_1 scores of 99.18/99.16/99.31, outperforming RT-DETRv2 by 0.92/1.15/1.08. The corresponding Recall gains are 1.27/1.65/1.55, indicating that the improvements mainly come from reducing missed Braille cells. BRIDGE also improves F_1 over RetinaNet by 0.82–1.15 and over BraUNet by 1.08–1.37 across the three challenging subsets. Since the subsets are not mutually exclusive, these results characterize model robustness under different visual conditions rather than performance on independent test sets.

• **Result on Public Dataset.** As shown in Table 3, we evaluate the effect of incorporating BrailleStuDoc into training with the two existing public datasets. When trained only on DSBI and Angelina, all methods show clear performance degradation on BrailleStuDoc. RT-DETRv2 obtains F_1 scores of 98.24/97.67/93.70 on DSBI, Angelina, and BrailleStuDoc, respectively, while BRIDGE achieves 99.67/99.15/94.20, indicating a domain gap between existing public datasets and student-produced, hand-punched Braille documents. BddNet exhibits a particularly large degradation, with F_1 decreasing from 78.36 on the scanned DSBI dataset to 17.97/14.73 on the camera-captured Angelina/BrailleStuDoc datasets. Unlike the cell-level detection methods, BddNet performs dot-level detection followed by cell reconstruction, suggesting greater sensitivity to appearance and geometric variations introduced by camera capture. After incorporating BrailleStuDoc into training, all six methods improve on the BrailleStuDoc test set, with F_1 gains of 2.34–26.19. Under joint training,

Table 4: Ablation study of MSFE and GASM in BRIDGE on BrailleStuDoc.

#	MSFE	GASM	Precision	Recall	F_1
1	–	–	98.39	98.07	98.23
2	✓	–	98.95	99.10	99.03
3	–	✓	99.23	99.03	99.13
4	✓	✓	**99.41**	**99.59**	**99.50**

BRIDGE achieves F_1 scores of 99.63/99.61/99.43 on DSBI, Angelina, and BrailleStuDoc, respectively, obtaining the best F_1 among the compared methods on all three datasets. These results demonstrate the complementary value of BrailleStuDoc for improving recognition of hand-punched Braille documents.

5.3 Ablation Study

We conduct ablation experiments to analyze the individual and combined contributions of MSFE and GASM, as shown in Table 4. Adding MSFE improves F_1 from 98.23 to 99.03, while GASM increases it to 99.13, demonstrating that both modules independently improve Braille cell recognition. Combining MSFE and GASM achieves the best performance, with Precision, Recall, and F_1 of 99.41/99.59/99.50, respectively. The full model improves F_1 by 1.27 over the baseline and by 0.37 over the stronger single-module variant, showing that the two modules can be effectively integrated within BRIDGE.

6 Conclusion

This paper addresses the limited coverage of hand-punched Braille documents in existing page-level benchmarks by introducing BrailleStuDoc, a benchmark dataset of Chinese Braille documents captured with smartphones. It contains 331 images from distinct physical pages and 149,720 cell-level annotations covering all 63 non-empty six-dot patterns, while reflecting illumination variation, geometric distortion, and Braille dot appearance variation encountered in practical smartphone capture. To address these challenges, we further propose BRIDGE, an illumination- and geometry-aware Braille cell detection framework built upon RT-DETRv2. BRIDGE incorporates Micro-Surface Feature Enhancement (MSFE) for robust local appearance representation and Geometry-Aware Scale Modeling (GASM) for modeling the local geometric regularity of neighboring Braille cells. Experimental results on BrailleStuDoc and existing public benchmarks demonstrate the effectiveness of BRIDGE and the value of BrailleStuDoc for Braille document recognition. BrailleStuDoc and the source code will be publicly released upon acceptance to support further research on optical Braille recognition and equitable education.

In our future work, we plan to further expand the scale and diversity of BrailleStuDoc by incorporating more document types and capture conditions for more comprehensive experimental evaluation. Moreover, we will investigate lightweight and robust Braille recognition models to achieve a better balance between recognition accuracy, robustness, and inference efficiency.

References

N. A. Asfaw, B. H. Belay, and K. M. Alemu. A deep learning approach for line-level amharic braille image recognition. *Scientific Reports*, 14(1):1, 2024.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, volume 12346 of *Lecture Notes in Computer Science*, pp. 213–229. Springer, 2020.

Karen Croake, Frances Gentle, and Jill Duncan. Evidence-based pedagogy used to teach beginning readers braille in a mainstream setting: A scoping review. *British Journal of Visual Impairment*, 43(2):407–422, 2025.

Nagwa Elaraby, Sherif Barakat, and Amira Rezk. A generalized ensemble approach based on transfer learning for braille character recognition. *Information Processing & Management*, 61:103545, 2024.

Elizabeth R. Hoskin, Morag K. Coyne, Michael J. White, Stephan C. D. Dobri, T. Claire Davies, and Shane D. Pinder. Effectiveness of technology for braille literacy education for children: A systematic review. *Disability and Rehabilitation: Assistive Technology*, 19(1):120–130, 2024.

Shihua Huang, Zhichao Lu, Xiaodong Cun, Yongjun Yu, Xiao Zhou, and Xi Shen. Deim: Detr with improved matching for fast convergence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15162–15171, 2025.

Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19702–19712, 2023.

Xiangwei Jiang, Rujiao Long, Nan Xue, Zhibo Yang, Cong Yao, and Gui-Song Xia. Revisiting document image dewarping by grid regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4543–4552, 2022.

Fituma Yadasa Kana and Asmerom Tekle Hagos. Factors hindering the use of braille for instruction and assessment of students with visual impairments: A systematic review. *British Journal of Visual Impairment*, 43(2):396–406, 2025.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13619–13627, 2022.

Feng Li, Ailing Zeng, Shilong Liu, Hao Zhang, Hongyang Li, Lei Zhang, and Lionel M. Ni. Lite detr: An interleaved multi-scale encoder for efficient detr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18558–18567, 2023a.

Heng Li, Xiangping Wu, Qingcai Chen, and Qianjin Xiang. Foreground and text-lines aware document image rectification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19574–19583, 2023b.

Renqiang Li, Hong Liu, Xiangdong Wang, and Yueliang Qian. DSBI: Double-sided braille image dataset and algorithm evaluation for braille dots detection. In *Proceedings of the 2nd International Conference on Video and Image Processing*, pp. 65–69. Association for Computing Machinery, 2018.

Renqiang Li, Hong Liu, Xiangdong Wang, and Yueliang Qian. Effective optical braille recognition based on two-stage learning for double-sided braille image. In *PRICAI 2019: Trends in Artificial Intelligence*, volume 11672 of *Lecture Notes in Computer Science*, pp. 150–163. Springer, 2019.

Renqiang Li, Hong Liu, Xiangdong Wang, Jianxing Xu, and Yueliang Qian. Optical braille recognition based on semantic segmentation network with auxiliary learning strategy. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2362–2368, 2020.

Zinuo Li, Xuhang Chen, Chi-Man Pun, and Xiaodong Cun. High-resolution document shadow removal via a large-scale real-world dataset and a frequency-aware shadow erasing net. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12449–12458, 2023c.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pp. 740–755. Springer, 2014.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=oMI9PjOb9Jl.

Liqiong Lu, Dong Wu, Jianfang Xiong, Zhou Liang, and Faliang Huang. Anchor-free braille character detection based on edge feature in natural scene images. *Computational Intelligence and Neuroscience*, 2022:7201775, 2022.

Liqiong Lu, Changjiang Chen, Dong Wu, and Jianfang Xiong. Natural scene braille image dataset and braille segment detection method. *Computer Engineering*, 49(10):171–177, 2023. In Chinese.

Wenyu Lv, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. RT-DETRv2: Improved baseline with bag-of-freebies for real-time detection transformer. *arXiv preprint arXiv:2407.17140*, 2024.

Zezheng Meng, Zefeng Cai, Jie Feng, Hanjie Ma, Haixiang Zhang, and Shaohua Li. Braille character segmentation algorithm based on gaussian diffusion. *Computers, Materials & Continua*, 79 (1):1481–1496, 2024.

Ilya G. Ovodov. Optical braille recognition using object detection neural network. In *2021 IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1741–1748, 2021.

Aasha Rose, Melissa Fanshawe, Georgina Barton, and Melissa Cain. Quality of life for braille users: A scoping review. *Review of Education*, 13(2):e70083, 2025.

Akihiro Yamashita, Hiroki Furukawa, Taichi Shirakawa, and Katsushi Matsubayashi. Braille character recognition independent of lighting direction using object detection and segmentation models. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 30(1):222–231, 2026.

ET-Tahir Zemouri, Nabil Zerrouki, Fouzi Harrou, and Ying Sun. AI-driven braille character recognition using partitioned spatial modeling and sequential learning. *Multimedia Systems*, 32(3):219, 2026.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations*, 2023a.

Jiaxin Zhang, Dezhi Peng, Chongyu Liu, Peirong Zhang, and Lianwen Jin. Docres: A generalist model toward unifying document image restoration tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15654–15664, 2024.

Shilong Zhang, Xinjiang Wang, Jiaqi Wang, Jiangmiao Pang, Chengqi Lyu, Wenwei Zhang, Ping Luo, and Kai Chen. Dense distinct query for end-to-end object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7329–7338, 2023b.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. DETRs Beat YOLOs on Real-time Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16965–16974, 2024.