

MIST: Multilingual Incidental Dataset for Scene Text Detection

Anonymous WACV Algorithms Track submission

Paper ID 1327

Abstract

Scene text detection has progressed rapidly, largely driven by curated datasets and benchmarks. However, many of these have reached evaluation saturation and are heavily biased toward focused scenes, limiting their effectiveness in real-world environments where detection is hindered by environmental factors. To address this, we introduce **MIST** – a Multilingual Incidental Scene Text dataset featuring diverse text instances across 11 languages. MIST provides language, transcription, legibility, and fine-grained polygon-shaped annotations across 12K scene images and 600K word-level text instances. Images are captured along roads using a GoPro mounted on a moving car to capture real-world complexities, ensuring the scenes are **incidental** rather than deliberately framed. MIST establishes a new challenging benchmark to enable robust evaluation of scene text detection methods in real-world scenarios. The datasets and code will be made publicly available.

1. Introduction

Text detection and recognition in natural scene images is crucial for various real-world applications, including autonomous systems like cars and drones, accessibility tools, augmented reality, scene understanding, and real-time translation [19, 23, 46, 49, 58]. The availability of large datasets and computing resources has driven advancements in state-of-the-art methods. However, challenges persist due to scene text variations in language, color, font, size, orientation, and shape at the **micro level**. At the same time, complex backgrounds, occlusions, and poor imaging conditions cause low resolution, distortion, noise, corruption, or blur at the **macro level** [19, 23, 58]. Together, these factors make accurate text localization and recognition difficult.

Micro-level challenges are primarily captured through **focused** scene text datasets. Such scene text is deliberately captured with the text in focus, maximizing resolution and legibility (refer to Fig. 3). Focused scene text datasets like ICDAR-03 [20], 11 [31], and 13 [9] spurred initial

Dataset Comparison: M1 vs M3 vs Total Instances

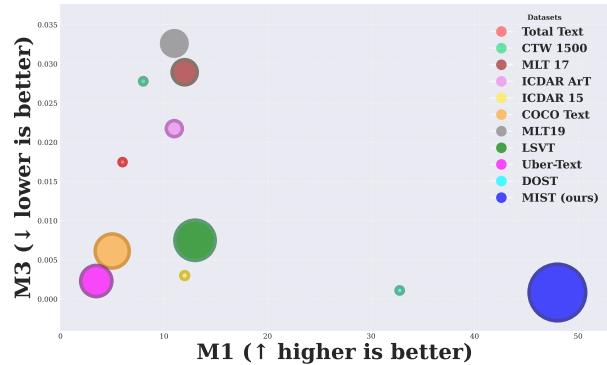


Figure 1. Displaying clusters of scene text detection datasets, including MIST, based on the average number of text instances per image (M1) and average area of instance about the image (M3). Bubble size represents the number of text instances in the whole dataset. Metrics are defined in Sec. 3.4

research but were limited by horizontal text assumptions. Succeeding datasets [22, 44, 47, 48] overcame this limitation by introducing multi-oriented rectangular bounding boxes, whereas Total-Text [2], CTW1500 [18] and ICDAR-ArT [3] addressed the challenge of curved text. Currently, state-of-the-art methods achieve approximately 90% F-Measure on Total-Text [2], CTW1500 [18], TD500 [44], and around 80% on ICDAR-ArT [3] and MLT17 [24].

These focused scene text datasets have driven much of the progress in scene text detection. However, their narrow capture angles and targeted objectives (for example, arbitrary shaped or multilingual text) reduce the prevalence of natural **macro level** complexities. In contrast, **incidental** scene text (refer to Fig. 2), which is captured over wider fields of view without any deliberate attempt to capture the text, naturally increases the likelihood of encountering higher frequency of **macro level** complexities. These challenges are highlighted in PSENet [38], which notes failures in densely crowded text, and TextBPN++ [53], which reports degradation under object occlusion.

Within *incidental* scene-text datasets (DOST [8],



Figure 2. An example from our MIST with detailed annotations, including transcription, script, legibility, and polygonal regions. The red polygons correspond to annotated text instances in the image.

058 COCO-Text [37], RoadText-1k [28], Uber-Text [54]), IC-
059 DAR15 [10] has long served as the de facto *primary* bench-
060 mark. With recent methods such as CPN [41] achieving
061 near-human performance (approaching 90% F-measure),
062 ICDAR15’s capacity to drive further progress is now lim-
063 ited. This motivates the need for a new, large-scale inciden-
064 tal dataset that captures diverse, complex text in real-world
065 conditions to drive further progress.



Figure 3. Shows examples from existing focused scene text detection datasets - Total-Text, CTW1500, and MLT17.

066 To address these concerns, we introduce **MIST** – a
067 Multilingual Incidental Scene Text dataset to advance re-
068 search in incorporating incidental scene text in addition to
069 focused scene text. Our main contributions are as follows:
070

- We introduce **MIST**, a **large-scale multilingual incidental** scene text dataset, comprising 12K images and 600K text instances across **11 languages** with transcription, legibility, language, and polygon-shaped annotations.
- We benchmark MIST using TextBPN++ [53], DPTText-DETR [45], MixNet [51], and DBNet++ [15]. The results highlight the need for further research in incidental scene text detection, as existing models show significant room for improvement despite dedicated training.
- We establish MIST as a general dataset on account of MIST trained models achieving **impressive zero-shot**

performance on existing datasets. Extending its generalisability, we display its **impressive transfer learning properties** and explore its **ability to mitigate existing issues in scene text detection**.

2. Related Works

2.1. Scene Text Detection Datasets

Early **focused** datasets [9, 20, 22, 31, 44, 47, 48] led the development up to multi-oriented, rectangular text. Total-Text [2], CTW1500 [18], and ICDAR-ArT [3] introduced arbitrary-shaped text with tight polygonal annotations and became primary benchmarks driving shape-generalised detectors. MLT17 [24] and MLT19 [25] are large-scale multilingual benchmarks spanning scripts such as Latin, Arabic, and Devanagari. However, a more challenging frontier lies in incidental scene text, as ICDAR15 [10] has saturated. Other incidental datasets face utility issues. For example, RoadText-1k [28] is a video-based dataset without static-image splits, making reproducible research difficult. DOST [8], which is closest to MIST in terms of incidental nature, includes a static-image split. However, it only consists of 338 images and the frames are sampled at every tenth interval, leading to redundancy. COCO-Text [37] remains under-utilized due to its sparse text density. To address this critical need for a more challenging incidental benchmark, we introduce MIST, a new dataset specifically designed to capture diverse macro-level complexities with high frequency. As we validate in Sec. 3.4 and Sec. 3.5, MIST demonstrates a superior incidental nature compared to ICDAR15 and existing datasets.

081
082
083
084

085

087
088
089
090
091

092
093
094

095
096
097
098
099
100
101
102
103
104
105
106
107
108
109

110 **2.2. Scene Text Detection**

111 **Regression-based Methods:** predict bounding boxes directly. TextBoxes [11] adapted SSD [16] by modifying convolution scales and anchors. TextBoxes++ [12] and 112 DMPNet [17] introduced quadrilateral regression for multi-oriented text, while SSTD [6] used attention to highlight 113 text regions. RRD [13] separated classification and regression for better multi-oriented text detection. EAST [57] and 114 DeepReg [7] used anchor-free, pixel-level regression, and 115 DeRPN [42] addressed scale variations. Though efficient 116 with simple post-processing (e.g., non-maximum suppression), 117 these methods struggle with irregular text shapes like 118 curved text. **Segmentation Based Methods:** predict text at 119 the pixel level, refining detections through post-processing. 120 Zhang *et al.* [55] combined segmentation with MSER for 121 multi-oriented text, while Xue *et al.* [43] used text borders 122 for instance separation. Mask TextSpotter [21] leveraged 123 Mask R-CNN [5] for arbitrary-shaped text. PSENet [38] 124 introduced progressive scale expansion, and Tian *et al.* [36] 125 used pixel embeddings for clustering. Though PSENet and 126 SAE [36] improved post-processing, they slowed inference. 127 DBNet [14] and DBNet++ [15] optimized segmentation by 128 integrating binarization without sacrificing speed.

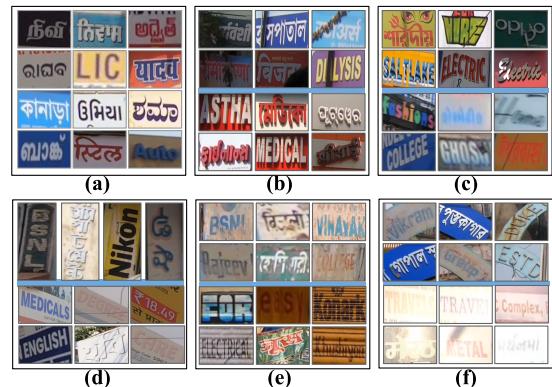
129 **Connected Components Methods:** decompose text 130 into segments and link them for detection. SegLink [32] 131 predicted segment boxes and connections for long text, 132 while SegLink++ [34] improved instance separation for 133 arbitrary shapes. Though effective for long text, these 134 methods depend on complex hyperparameters, making them 135 difficult to tune. **Boundary Based Methods:** detect text by 136 predicting key points along text boundaries [1, 35, 39, 56]. 137 CTD [50] used RNN decoding for 14 boundary points, 138 while [39] employed BLSTM for adaptive boundary prediction. 139 BPN [53], BPN++[52], and DPText-DETR[45] follow 140 a two-stage approach, generating coarse boundary proposals 141 followed by refinement via iterative deformation. Despite 142 achieving high accuracy, complex iterative corrections 143 remain computationally intensive.

144 **3. MIST Dataset**145 **3.1. Scene Image Collection**

146 Occurrence of small scale text in an environment with abundant 147 complexities such as occlusions, perspective distortions, motion blur, varying fonts, styles, colors, shapes, scripts, complex backgrounds, and cluttered arrangements is essential for curating an incidental scene text dataset. 148 Roadside scenes provide a rich and diverse setting for collecting such scene text images, making them an invaluable resource for real-world text detection datasets.

149 We follow a straightforward setup by mounting a GoPro 150 camera on a moving vehicle, facilitating the seamless acquisition 151 of large-scale datasets and eliminating the need 152 for extensive manual intervention. Collecting video sequences 153 across diverse in India also captures **multilingual text**, 154 adding multilinguality as a secondary attribute of the dataset, 155 alongside the incidental nature. The dynamic nature 156 of this collection process ensures the captured scenes 157 are incidental, as the vehicle captures text naturally occurring 158 in the environment rather than being deliberately framed. 159 The incidental nature of these images makes them highly 160 valuable for training robust scene text detection models 161 capable of handling unconstrained real-world scenarios.

162 **for extensive manual intervention. Collecting video sequences across diverse in India also captures multilingual text, adding multilinguality as a secondary attribute of the dataset, alongside the incidental nature. The dynamic nature of this collection process ensures the captured scenes are incidental, as the vehicle captures text naturally occurring in the environment rather than being deliberately framed. The incidental nature of these images makes them highly valuable for training robust scene text detection models capable of handling unconstrained real-world scenarios.**



163 Figure 4. Showcases the diverse and complex text instances in 164 MIST, including (a) multilingual text, (b) partially occluded and 165 3D text, (c) artistic and motion-blurred text, (d) vertical and 166 perspective text, (e) text blending into backgrounds, and (f) curved 167 and high-illumination text.

Dataset	$M_1 \uparrow$	$M_2 \downarrow$	$M_3 \downarrow$	#images	#text instances
Total-Text [2]	6	0.01093	0.01747	1555	9330
CTW1500 [18]	8	0.01796	0.02778	1500	12000
MLT17 [24]	12	0.01017	0.02893	18000	216000
ICDAR-ArT [3]	11	0.01230	0.02174	10166	111826
ICDAR15 [10]	12	0.00204	0.00300	1500	18000
COCO-Text [37]	5	0.00344	0.00613	63686	318430
MLT19 [26]	11	0.01125	0.03261	20000	220000
LSVT [33]	13	0.01270	0.00750	30000	390000
Uber-Text [54]	3.46	0.00277	0.00230	82572	285699
DOST [8]	32.77	0.00090	0.00110	338	11076
MIST (ours)	48	0.00067	0.00084	12000	576000

168 Table 1. Statistical comparison of scene text datasets. Best values 169 are **bolded**; second-best are underlined.

170 **3.2. Annotation**

171 MIST provides word-level polygon shaped annotations, 172 similar to Total-Text [2]. Each word is annotated with 173 a polygonal region, its script, legibility, and transcription. 174 The dataset covers 11 languages – English, Bengali, Gu- 175 jarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Pun- 176jabi, Tamil, and Telugu – while other languages, digital wa- 177 termarks, logos, and illegible text are marked as *do not care*. 178 Any detected *do not care* text is excluded from the evalua- 179

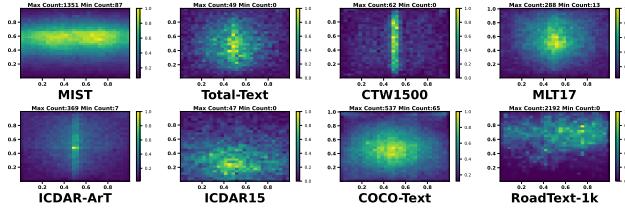


Figure 5. Presents a comparison of the spatial distribution of text instances between MIST and existing datasets. MIST, along with Total-Text, MLT17, ICDAR15, and COCO-Text, exhibits a broader spread of text instances compared to other datasets.

180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
tion. Fig. 2 shows a sample annotation and summarizes the
annotation collected for each text region.
We employ a two-step semi-automatic approach to
reduce manual effort and improve efficiency. First, selected
images are processed using Google Cloud OCR¹ to acquire
initial estimates for bounding box, script, and transcription
for each text instance. Next, our skilled and experienced in-
house annotators verify and edit the acquired estimates and
further annotate the text instances undetected by Google
Cloud OCR. They also label text as legible or illegible. Fi-
nal annotations include polygon boundaries, transcription,
script and legibility tag. The entire process took approx.
three months and approx. 2000 annotator hours. *More de-
tails provided in Appendix A of supplementary material.*

3.3. Dataset Statistics

From 150 hours of video, we extract diverse, non-overlapping frames to construct MIST. It comprises 12K scene images containing 576K text instances across 11 scripts, with each image at 1920×1080. To ensure temporal and regional diversity while avoiding duplicates images, we enforce per-region and per-sequence quotas and sample uniformly over time via equal temporal bins (see Appendix B of Supplementary material). We split the dataset into training, validation, and testing (benchmark) in a 4:1:1 ratio, with the benchmark comprising 2000 scene images.

The incidental nature of MIST images results in a more diverse text distribution, including a higher occurrence of smaller, naturally embedded, and legible text instances, enhancing the dataset’s real-world applicability. Table 1 and Fig. 1 provide a comprehensive distinction between MIST’s distribution and that of the existing dataset. Fig. 4 illustrates the diverse and complex text instances present in MIST. This variety makes MIST a highly challenging dataset for text detection and recognition. Fig. 5 compares the spatial distribution of text across various datasets, including MIST.

3.4. Inherent Characteristics

In this subsection, we first define the notion of an incidental dataset using relevant metrics. On this basis, we empiri-

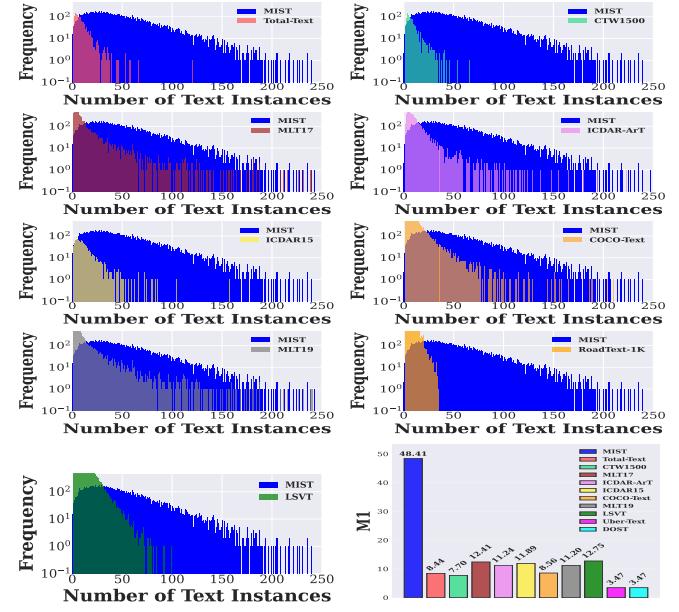


Figure 6. Compares the distribution of text instances in scene images (M_1) of against existing datasets: Total-Text, CTW1500, MLT17, ICDAR-ArT, ICDAR15, COCO-Text, and RoadText-1K.

Image-level metric	Dataset-level metric
$m_1 = N$	$M_1 = \frac{1}{ D } \sum_{i \in D} m_1^{(i)}$
$m_2 = \frac{A_t}{A_I}, \forall t \in T$	$M_2 = \frac{1}{\sum_{i \in D} T_i } \sum_{i \in D} \sum_{t \in T_i} m_2^{(i,t)}$
$m_3 = \frac{\sum_t A_t}{A_I N}$	$M_3 = \frac{1}{ D } \sum_{i \in D} m_3^{(i)}$

Table 2. Image- vs. dataset-level metrics.

cally validate the limitations of models trained on existing datasets, and thus motivate the need for an incidental scene text dataset and benchmark. MIST proves to be the most incidental in nature and hence promises to fulfill the requirement for an incidental dataset.

Metrics. We define two image-level metrics (m_1, m_3) and one text-instance-level metric (m_2), along with their dataset-level counterparts (M_1, M_2, M_3), to quantify incidental characteristics. For a single image I , T denotes its set of text instances ($N = |T|$), A_t the area of an instance $t \in T$, and A_I the image area. For a dataset $D = \{I_i\}$, T_i and $N_i = |T_i|$ denote the instances and count for image I_i . Table 2 summarizes the metric definitions. For the following discussion, let H denote the set of incidental scene text images and F the set of focused scene text images.

m_1 / M_1 . m_1 counts text instances in a single image, whereas M_1 averages this count over the dataset. Typically, $m_1^H > m_1^F$ and $M_1^H > M_1^F$.

¹<https://cloud.google.com/use-cases/ocr>

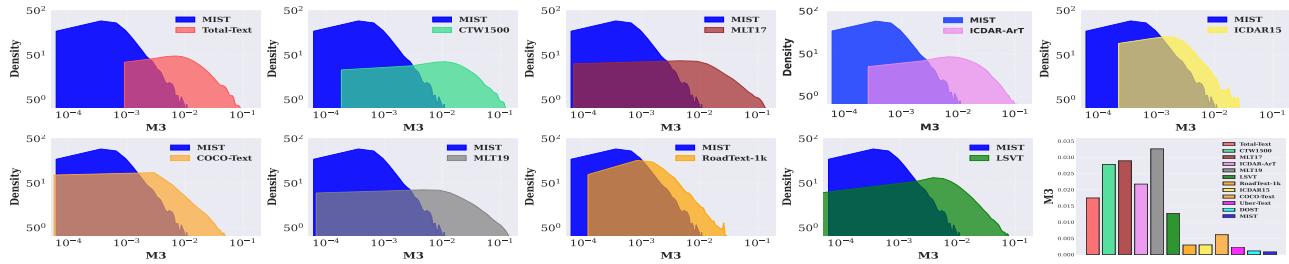


Figure 7. Comparison of the **average area of a text instance relative to the scene image (M_3)** in MIST against other benchmarks: Total-Text, CTW1500, MLT17, ICDAR-ArT, ICDAR15, COCO-Text, MLT19, RoadText-1K, and LSVT. The analysis employs kernel density estimation to create continuous distribution curves, with both axes shown on logarithmic scales (base 50 for y-axis and base 2 for x-axis).

236 m_2 / M_2 . m_2 is the per-instance area proportion,
237 whereas M_2 is the instance-weighted mean over the dataset.
238 Typically, $m_2^H < m_2^F$ and $M_2^H < M_2^F$.

239 m_3 / M_3 . m_3 is the average instance area proportion
240 within an image, whereas M_3 is the mean of m_3 across im-
241 ages. Typically, $m_3^H < m_3^F$ and $M_3^H < M_3^F$; as $m_3 \rightarrow 0$
242 or $M_3 \rightarrow 0$, scene text is more incidental.

243 Real-world text detection encounters challenges such as
244 occlusion, perspective distortion, motion blur, poor illu-
245 mination, and smaller legible texts. Collecting such arti-
246 facts for training and benchmarking is essential to drive
247 the progress of robust and generalizable scene text detec-
248 tion models. Thus, to maximize the likelihood of capturing
249 such complexities, we collect our data by capturing
250 text-dense high resolution scene images (detailed in Sec-
251 tion 3.1). Fig. 6 illustrates the probability mass function of
252 text instances per image. MIST displays a **well-balanced**
253 and **dense** text distribution compared to existing datasets,
254 which tend to be predominantly right-skewed. Compared to
255 its incidental counterparts, MIST boasts approximately **four**
256 times the **M1** of ICDAR15 and **six** times that of COCO-
257 Text, making it a richer and more comprehensive dataset.

258 If two datasets have similar M_1 , the scale of the scene
259 text is a differentiating cue between a focused and inci-
260 dental scene text. Since incidental scene text is captured
261 without deliberately focusing on the text, they appear rela-
262 tively small. To quantify this, we use the metric M_3 , which
263 measures the relative area occupied by text instances within
264 an image. Fig. 7 illustrates the distribution gap of M_3 be-
265 tween the MIST and existing datasets. A lower M_3 indi-
266 cates higher degree of *incidentalness*, positioning MIST as
267 a highly incidental dataset. Notably, MIST has huge den-
268 sities of small-scale text (emphasized by the **logarithmic**
269 **scale**) and its average M_3 is significantly smaller than other
270 datasets, averaging **15-20** times smaller than the existing fo-
271 cused datasets and **4** times smaller than its incidental coun-
272 terparts. Additionally, it extends coverage along the M_3
273 scale that was previously explored only to a limited extent
274 and scale, expanding the scope of text scales in scene text
275 detection literature.

3.5. Empirical Validation

In this section, we evaluate the limitations of current models on incidental scene text. Specifically, we use the TBPNN (Deformable-Resnet50 variant of TextBPN++[53]) models trained on MLT17, ICDAR-ArT, and Total-Text. For consistency, all models are tested at an input resolution of 640×1024 , using the official weights released in the TBPNN GitHub repository. The corresponding training configurations are described in the original paper[53]. Experiments are conducted on the MLT17 validation set, as well as the Total-Text and MIST test sets.

In our first experiment, we examine how the F-Measure varies with M_2 . For each image, we sweep a threshold τ over m_2 and label instances with $m_2 > \tau$ as *do-not-care regions*. The F-Measure at each τ is computed per image, averaged across the dataset, and plotted against τ . To isolate the effect of instance scale, we keep M_1 and M_3 approximately constant. **All results here are in-domain**, the models are trained and evaluated on the same dataset. Fig. 8(a) illustrates F-Measure as a function of M_2 : as the threshold decreases (retaining smaller instances), performance of the model trained on existing datasets declines, likely due to limited representation across scales or implicit scale biases. MIST aims to alleviate this by providing large numbers of text instances spanning a wide range of scales, especially those underrepresented in existing datasets.

We also investigate how a text detection model's performance correlates with an "incidental" scene. **All analyses here are in-domain**. Using the same datasets, we sort M_3 and divide it into batches of 5 for Total-Text and 20 for MLT17. Instead of aggregating the F-Measure across the entire dataset, we compute it for each batch. Plotting F-Measure against M_3 , shown in Fig. 8(b) and (c), reveals a performance decline as scenes transition from focused to incidental. Based on our assumption regarding macro-level complexities, this trend primarily reflects how scene text detection systems struggle as a scene shifts toward *incidental*. The region of poor performance for both datasets aligns with their left tail in the M_3 distribution, which no-

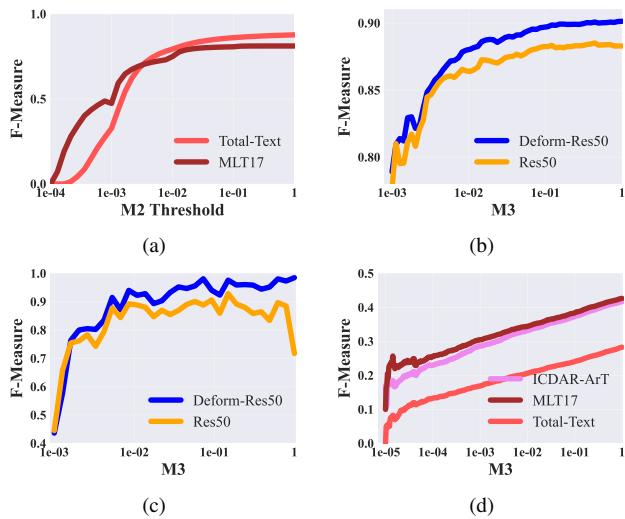


Figure 8. (a) F-Measure at varying M_2 thresholds for TBPN, trained and evaluated separately on Total-Text and MLT17. (b) F-Measure for each batch of 20 sorted M_3 elements on MLT17, comparing TBPN and ResNet-50 variant of TextBPN++ trained on MLT17. (c) F-Measure for each batch of 5 sorted M_3 elements on Total-Text, comparing the same two variants trained on Total-Text. (d) Zero-shot F-Measure across batches of 5 sorted M_3 elements on MIST, using TBPN trained on ICDAR-ArT, Total-Text, and MLT17.

315 tably overlaps with the right tail in MIST’s M_3 distribution
316 (Fig. 7). This underscores how even slight distribution
317 shifts can significantly impact scene text detection: models
318 trained on existing datasets perform poorly even under rela-
319 tively lenient real-world conditions (right tail of MIST’s M_3
320 distribution) due to skewed data distributions.

321 Zero-shot Evaluation: After analyzing the in-domain
322 performance of various models, we will now evaluate their
323 zero-shot(*out-of-domain*, OOD) performance on MIST to
324 better reflect real-world conditions. Assessing scene text
325 detection in these scenarios is crucial. Table 4 includes the
326 zero-shot results of Total-Text, MLT17, and ICDAR-ArT
327 on MIST, revealing a significant issue with extremely low
328 recall rates, indicating that many text instances are missed
329 or detected unsatisfactorily. The decline in performance
330 stems from domain shift, small text prevalence, and dataset-
331 specific challenges.

Similar to Fig. 8 (b), we plot F-Measure for each batch of 5 sorted elements from M_3 on MIST, using TBPN trained on ICDAR-ArT, Total-Text, and MLT17 in the zero-shot setting as Fig. 8 (d). The figure shows that the **F-Measure remains below 0.5** even for larger text, indicating that all text sizes are affected by macro-level complexities, with smaller text facing additional challenges due to their scale. This can be attributed to insufficient training samples in dif-

ficult conditions and the low variability of text scales in existing datasets. MIST aims to bridge this gap by aligning dataset benchmarks more closely with real-world scenarios.

To summarize, MIST stands out as a unique incidental dataset due to its high frequency of macro level complexities, owing to its dense text distribution and the small scale text instances. This combination enables it to mirror complex real-world scenarios better than existing datasets.

3.6. Quality Control

Since our road-captured scene images may naturally contain sensitive information, we employ a two-stage pipeline to conceal Personally Identifiable Information (PII). In Stage I, RetinaFace [29, 30] and EgoBlur [27] are used to automatically blur human faces and license plates. In Stage II, we manually review all images to blur any faces or license plates missed in the first stage, ensuring comprehensive PII protection. Due to the scale and complexity of our dataset, we performed thorough evaluation checks to reduce annotation errors. We performed inter-annotator verification over two iterations to ensure accurate and unambiguous annotation.

4. Experiments

4.1. Baselines and Implementation

We benchmark MIST with four detectors: TBPN (the Deformable ResNet-50 variant of TextBPN++ [53]), DP-DETR (DPText-DETR [45]), MixNet [51], and DB++ (DB-Net++ [15]). Each model is initialized from the authors' publicly released pretrained weights on their respective GitHub repositories. We follow the authors' training configurations for MixNet and DB++; for TBPN and DP-DETR, we keep all settings unchanged except the learning-rate schedule, setting the initial learning rate to 10^{-4} and applying a step decay of 0.9 every 10 epochs.

4.2. Evaluation

We report the Precision (**P**), Recall (**R**) and F-measure (**F**), calculated through the DetEval [40] protocol for evaluating scene text detection models, as used in Total-Text. This protocol supports – One-to-One, One-to-Many, and Many-to-One matching. We set $tr = 0.7$ and $tp = 0.6$ for a more interpretable benchmark evaluation. Incidental scenes feature smaller, clustered text, making $tp = 0.6$ ideal to prevent detections from merging with surrounding text or background. Meanwhile, macro level complexities can lead to fragmented detections, making $tr = 0.7$ a fairer choice to avoid excessive penalties. We also propose a **stratified evaluation benchmark** to better assess the challenges of incidental scenes. Standard evaluation often masks failure cases, as high overall scores may stem from easier instances. To prevent misrepresentation, we stratify MIST

389 into two sets based on M_3 thresholds of 0.0004 and 0.0002,
 390 ensuring a more informative evaluation. *More details can*
 391 *be found in Appendix B of supplementary material.*

Model	PT	P	R	F	F^α	F^β
DP-DETR	Syn	69.61	57.04	62.70	59.15	52.80
TBPN	MLT	70.87	47.75	<u>57.06</u>	52.09	44.68
MixNet	Syn	73.48	45.59	56.27	51.44	44.06
DB++	Syn	72.84	39.73	51.42	44.71	38.32

Table 3. Benchmarking results on MIST. The **PT** column reports the dataset used for pretraining: MLT represents MLT17[24] and Syn represents SynthText [4]. F^α and F^β denote stratified evaluation with M_3 thresholds of 0.0004 and 0.0002, respectively.

392 4.3. Benchmark Result

393 Table 3 presents the quantitative results, suggesting that
 394 the existing models have room for improvement to cater to
 395 complex real-world scenarios. The stratified evaluation fur-
 396 ther strengthens this claim, capturing the poor performance
 397 in much harder incidental scenes. *Refer to the supplemen-*
 398 *tary material for visual results.*

399 5. Insights & Takeaways

400 5.1. Error Type Analysis

401 We qualitatively analyze how TBPN, trained on MIST (fol-
 402 lowing Sec. 4.1), fares under different error types. Using
 403 samples images from MIST test-set, we quantified false
 404 negatives per category, as shown in Fig. 9(g,h) . The
 405 false negatives are dominated by low-resolution and low-
 406 illumination cases, both high in absolute frequency and in
 407 their within-category percentage. In contrast, object oc-
 408 clusion, text-background merging, and perspective distor-
 409 tion are relatively infrequent and exhibit lower percentage
 410 of within-category false negatives (all under 40%). These
 411 trends indicate that improving robustness to low resolution
 412 and poor illumination is likely to yield the greatest gains.

413 5.2. Generalization Capability of MIST

414 Scene text detection is shaped by intrinsic factors (style,
 415 shape, script, size) and extrinsic perturbations (occlusion,
 416 motion blur, perspective distortion). To assess generaliza-
 417 tion, performance of models on out of distribution data, un-
 418 der this variability, we use **TBPN** (Deformable-Resnet50
 419 version of TextBPN++) trained on MIST (same as in
 420 Sec. 4.1). For comparison, we use the authors' publicly
 421 released TBPN checkpoints trained on MLT17, TOTAL-
 422 TEXT, and ICDAR-ART (same as in Sec. 3.5). All
 423 checkpoints are evaluated zero-shot on the target datasets
 424 listed in Table 4 (i.e., MLT17, TOTAL-TEXT, ICDAR15,
 425 CTW1500, and MIST when applicable).

426 **Results.:** Let $TBPN_X$ denote the TBPN trained on X. Ta-
 427 ble 4 shows that $TBPN_{MIST}$ is competitive across all targets:

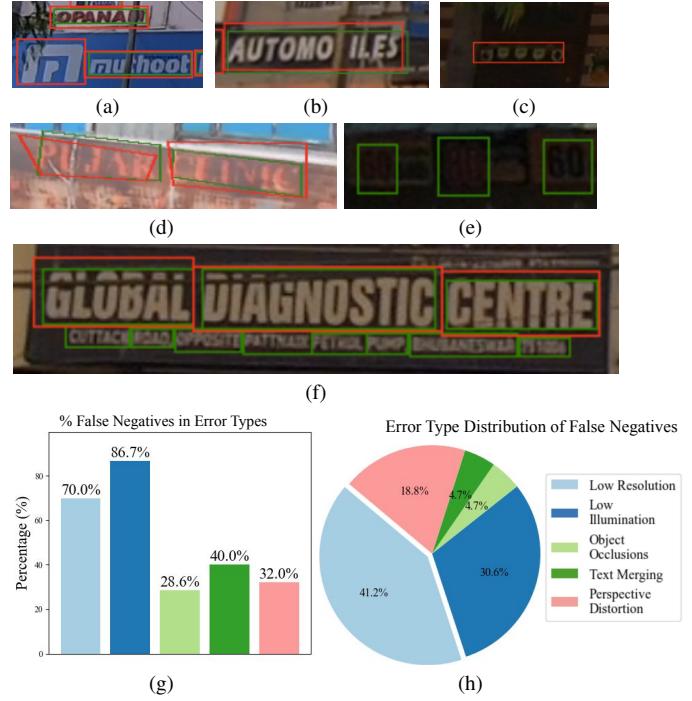


Figure 9. Green boxes denote ground truth and red boxes denote predictions. (a),(b) object occlusion on text, (c) false positive (door pattern as text), (d) text with merging backgrounds, (e) low-illumination text, (f) small-scale text. Error analysis of TBPN on MIST. (g) Within-category false-negative (FN) rate for each error type, defined as FN/#occurrences in that category. (h) Composition of all FNs across categories (normalized to 100%).

Train Set	Test Set	P	R	F
MIST	Total-Text	64.57	78.44	70.83
	ICDAR15	73.73	59.20	65.68
	CTW1500	80.62	86.00	<u>83.22</u>
	MLT17	80.00	63.00	<u>70.50</u>
	MIST	70.87	47.75	57.06
MLT17	Total-Text	69.82	70.05	<u>69.94</u>
	ICDAR15	79.00	47.50	59.40
	CTW1500	85.20	82.20	83.67
	MIST	88.88	28.06	42.65
	MLT17	83.74	72.10	77.48
Total-Text	ICDAR15	77.00	37.33	50.30
	CTW1500	86.30	78.69	82.00
	MLT17	83.67	48.84	61.68
	MIST	66.41	17.97	28.28
	Total-Text	92.44	87.93	90.13
ICDAR-ArT	ICDAR15	83.31	46.60	<u>59.77</u>
	MLT17	85.50	60.92	71.15
	MIST	86.04	28.01	<u>42.27</u>
	ICDAR-ArT	84.48	77.05	80.59

Table 4. Cross-dataset performance of TBPN [53] checkpoints. Inputs are 640×1024 and evaluation uses DetEval ($tr=0.7$, $tp=0.6$). Bold and underline indicate the best and second-best cross-domain results, respectively.

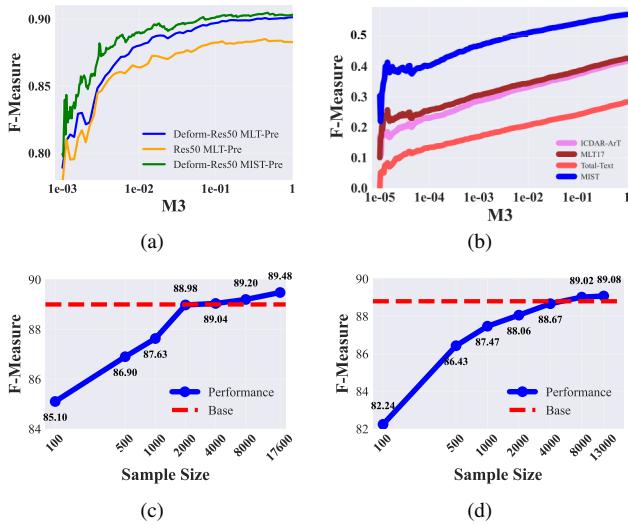


Figure 10. (a) Comparison of F-Measure vs. M_3 for TextBPN++ pre-trained on MLT17 vs. MIST, trained/evaluated on Total-Text. (b) F-Measure vs. M_3 for TextBPN++ trained on different datasets and evaluated on MIST. (c)–(d) Few-shot performance of DP-DETR pre-trained on MIST for Total-Text (c) and CTW1500 (d).

it outperforms TBPN_{TOTAL-TEXT} on MLT17 and CTW1500; on CTW1500 it closely matches TBPN_{MLT17}; and on IC-DAR15 it surpasses both TBPN_{MLT17} and TBPN_{TOTAL-TEXT}. These results indicate that training on MIST confers strong real-world generalization while remaining competitive on focused, out-of-domain benchmarks. Additional analyses are provided in the supplementary material.

Test Set	Model	P	R	F	F^α
Total-Text	H1	92.06	87.05	89.48	89.00
	H2	92.60	88.85	90.69	90.13
CTW1500	H1	91.58	86.72	89.08	88.80
	H2	88.56	86.83	87.69	86.49
MLT	H2	91.57	74.29	82.02	81.19

Table 5. Shows performance of DP-DETR[45] and TBPN[53] pre-trained on MIST and fine-tuned on Total-Text and CTW1500. H1 and H2 indicate DP-DETR and TBPN, respectively. F^α is the best reported F-Measure for Total-Text, CTW1500, and MLT17 as reported in [45, 53]. **Bold** and underline indicate the best and second-best value, respectively. Since DP-DETR is not evaluated on MLT17, we only use TBPN for MLT17.

5.3. Transfer Learning Capability

We assess the transfer learning capability of models trained on MIST through few-shot and whole dataset training, results reported in Table 5 (and Table 2 in supplementary). Fine-tuning on Total-Text, we surpass TextBPN++’s reported performance by 0.56% while surpassing DP-DETR

DET R’s reported performance by 0.45%. On CTW1500, we achieve an impressive 1.2% improvement over the base TextBPN++. Fig. 10c and Fig. 10d highlights that MIST achieves near baseline performance on specific datasets with a fraction of the training data - 2000 out of 17600 samples for TotalText and 4000 out of 13000 samples for CTW1500. In-depth details can be found in supplementary material.

5.4. Issue Mitigation Observation

We discussed how current models struggle as scene text gets more incidental. Here, we examine whether MIST improves performance on specific datasets beyond enhancing real-world generalization capability. We plot the F-Measure vs. M_3 (Fig. 10a) per image for TBPN, pre-trained on MIST and fine-tuned on Total-Text. The fine-tuned model consistently outperforms the base Total-Text model, pre-trained on MLT17, especially at lower and medium M_3 values, implying issue mitigation from macro complexities and small scale. We perform more analysis in supplementary material.

5.5. Limitations and Future Directions

Although the dataset is diverse in macro level complexities (Fig. 4), arbitrarily shaped or curved text is relatively underrepresented. It also exhibits global geographic bias and limited language diversity, as all images were captured in India (although across multiple subregions). Despite this bias, models trained on MIST exhibit strong out-of-domain (OOD) generalization to datasets from other regions. Our zero-shot performance on arbitrary-shaped text surpasses other models, but performance can degrade on small-scale instances. Finally, we aim to quantify (building on Sec 5.1) macro-level disturbances to improve issue traceability.

6. Conclusion

In this paper, we introduce MIST (Multilingual Incidental Scene Text), a large scale incidental scene text dataset, addressing key limitations in existing focused datasets. We demonstrate the necessity of incidental datasets by highlighting the limitations of models trained on existing datasets on incidental scene text. Our baseline experiments highlight the potential for improvement in scene text detection and MIST offers a challenging yet promising setting to drive such advancements. MIST also exhibits impressive transfer learning, few shot learning and issue mitigation capabilities, underscoring its utilities. We hope our work helps advance research towards building robust and generalizable real world text detection models.

References

- [1] Youngmin Baek, Bado Lee, Dongyo Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *CVPR*, pages 9365–9374, 2019. 3

- 490 [2] Chee Kheng Ch'ng and Chee Seng Chan. Total-Text: A
491 comprehensive dataset for scene text detection and recogni-
492 tion. In *ICDAR*, pages 935–942, 2017. 1, 2, 3
- 493 [3] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng,
494 Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaifao Zhang,
495 Junyu Han, Errui Ding, et al. ICDAR2019 robust reading
496 challenge on arbitrary-shaped text-RRC-ART. In *ICDAR*,
497 pages 1571–1576, 2019. 1, 2, 3
- 498 [4] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman.
499 Synthetic data for text localisation in natural images. In
500 *CVPR*, pages 2315–2324, 2016. 7
- 501 [5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Gir-
502 shick. Mask R-CNN. In *ICCV*, pages 2961–2969, 2017. 3
- 503 [6] Pan He, Weilin Huang, Tong He, Qile Zhu, Yu Qiao, and
504 Xiaolin Li. Single shot text detector with regional attention.
505 In *ICCV*, pages 3066–3074, 2017. 3
- 506 [7] Wenhao He, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu.
507 Deep direct regression for multi-oriented scene text detec-
508 tion. In *ICCV*, pages 745–753, 2017. 3
- 509 [8] Masakazu Iwamura, Takahiro Matsuda, Naoyuki Morimoto,
510 Hitomi Sato, Yuki Ikeda, and Koichi Kise. Downtown osaka
511 scene text dataset. In *ECCV*, pages 440–455, 2016. 1, 2, 3
- 512 [9] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida,
513 Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles
514 Mestre, Joan Mas, David Fernandez Mota, Jon Almazàn Al-
515 mazàn, and Lluís Pere de las Heras. ICDAR 2013 robust
516 reading competition. In *ICDAR*, pages 1484–1493, 2013. 1,
517 2
- 518 [10] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos
519 Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwa-
520 mura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chan-
521 drasekhar, Shijian Lu, et al. Icdar 2015 competition on robust
522 reading. In *ICDAR*, pages 1156–1160, 2015. 2, 3
- 523 [11] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang,
524 and Wenyu Liu. Textboxes: a fast text detector with a single
525 deep neural network. In *AAAI*, page 4161–4167, 2017. 3
- 526 [12] Minghui Liao, Baoguang Shi, and Xiang Bai. TextBoxes++:
527 A single-shot oriented scene text detector. *TIP*, 27:3676–
528 3690, 2018. 3
- 529 [13] Minghui Liao, Zhen Zhu, Baoguang Shi, Gui-song Xia, and
530 Xiang Bai. Rotation-sensitive regression for oriented scene
531 text detection. In *CVPR*, pages 5909–5918, 2018. 3
- 532 [14] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xi-
533 ang Bai. Real-time scene text detection with differentiable
534 binarization. In *AAAI*, pages 11474–11481, 2020. 3
- 535 [15] Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and
536 Xiang Bai. Real-time scene text detection with differentiable
537 binarization and adaptive scale fusion. *TPAMI*, 45(1):919–
538 931, 2022. 2, 3, 6
- 539 [16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian
540 Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C.
541 Berg. SSD: Single shot multibox detector. In *ECCV*, pages
542 21–37, 2016. 3
- 543 [17] Yuliang Liu and Lianwen Jin. Deep matching prior network:
544 Toward tighter multi-oriented text detection. In *Proceed-
545 ings of the IEEE Conference on Computer Vision and Pattern
546 Recognition (CVPR)*, 2017. 3
- 547 [18] Yuliang Liu, Lianwen Jin, Shuaifao Zhang, Canjie Luo, and
548 Sheng Zhang. Curved scene text detection via transverse and
549 longitudinal sequence connection. *PR*, 90:337–345, 2019. 1,
550 2, 3
- 551 [19] Shangbang Long, Xin He, and Cong Yao. Scene text detec-
552 tion and recognition: The deep learning era. *IJCV*, 129(1):
553 161–184, 2021. 1
- 554 [20] Simon M. Lucas, Alex Panaretos, Luis Sosa, Anthony Tang,
555 Shirley Wong, Robert Young, Kazuki Ashida, Hiroki Nagai,
556 Masayuki Okamoto, Hiroaki Yamamoto, Hidetoshi Miyao,
557 JunMin Zhu, WuWen Ou, Christian Wolf, Jean-Michel Jo-
558 lion, Leon Todoran, Marcel Worring, and Xiaofan Lin. Icdar
559 2003 robust reading competitions: entries, results, and future
560 directions. *International Journal of Document Analysis and
561 Recognition (IJDAR)*, 7(2):105–122, 2005. 1, 2
- 562 [21] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and
563 Xiang Bai. Mask textspotter: An end-to-end trainable neural
564 network for spotting text with arbitrary shapes. In *ECCV*,
565 pages 67–83, 2018. 3
- 566 [22] Robert Nagy, Anders Dicker, and Klaus Meyer-Wegener.
567 NEOCR: A configurable dataset for natural image text recog-
568 nition. In *CBDAR*, pages 150–163, 2011. 1, 2
- 569 [23] Fatemeh Naiemi, Vahid Ghods, and Hassan Khalesi. Scene
570 text detection and recognition: a survey. *Multimedia Tools
571 Appl.*, 81(14):20255–20290, 2022. 1
- 572 [24] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan
573 Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal,
574 Christophe Rigaud, Joseph Chazalon, et al. ICDAR2017 ro-
575 bust reading challenge on multi-lingual scene text detection
576 and script identification-RRC-MLT. In *ICDAR*, pages 1454–
577 1459, 2017. 1, 2, 3, 7
- 578 [25] Nibal Nayef, Yash Patel, Michal Busta, Pinaki Nath Chowd-
579 hury, Dimosthenis Karatzas, Wafa Khelif, Jiri Matas, Uma-
580 pada Pal, Jean-Christophe Burie, Cheng-lin Liu, and Jean-
581 Marc Ogier. Icdar2019 robust reading challenge on multi-
582 lingual scene text detection and recognition – rrc-mlt-2019,
583 2019. 2
- 584 [26] Nibal Nayef, Yash Patel, Michal Busta, Pinaki Nath Chowd-
585 hury, Dimosthenis Karatzas, Wafa Khelif, Jiri Matas, Uma-
586 pada Pal, Jean-Christophe Burie, Cheng-lin Liu, and Jean-
587 Marc Ogier. ICDAR2019 robust reading challenge on multi-
588 lingual scene text detection and recognition—RRC-MLT-
589 2019. In *ICDAR*, pages 1582–1587, 2019. 3
- 590 [27] Nikhil Raina, Guruprasad Somasundaram, Kang Zheng,
591 Sagar Miglani, Steve Saarinen, Jeff Meissner, Mark
592 Schwesinger, Luis Pesqueira, Ishita Prasad, Edward Miller,
593 Prince Gupta, Mingfei Yan, Richard Newcombe, Carl Ren,
594 and Omkar M Parkhi. Egoblur: Responsible innovation in
595 aria. *arXiv*, 2023. 6
- 596 [28] Sangeeth Reddy, Minesh Mathew, Lluis Gomez, Marçal
597 Rusinol, Dimosthenis Karatzas, and CV Jawahar. Roadtext-
598 1k: Text detection & recognition dataset for driving videos.
599 In *ICRA*, pages 11074–11080, 2020. 2
- 600 [29] Sefik Serengil and Alper Ozpinar. A benchmark of facial
601 recognition pipelines and co-usability performances of mod-
602 ules. *Journal of Information Technologies*, 17(2):95–107,
603 2024. 6

- 604 [30] Sefik Ilkin Serengil and Alper Ozpinar. LightFace: A hybrid
605 deep face recognition framework. In *Innovations in Intel-*
606 *ligent Systems and Applications Conference (ASYU)*, pages
607 23–27, 2020. 6
- 608 [31] Asif Shahab, Faisal Shafait, and Andreas Dengel. Icdar
609 2011 robust reading competition challenge 2: Reading text
610 in scene images. In *2011 International Conference on Doc-*
611 *ument Analysis and Recognition*, pages 1491–1496, 2011. 1, 2
- 612 [32] Baoguang Shi, Xiang Bai, and Serge Belongie. Detecting
613 oriented text in natural images by linking segments. In
614 *CVPR*, pages 2550–2558, 2017. 3
- 615 [33] Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Can-
616 jie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo
617 Liu, Dimosthenis Karatzas, et al. Icdar 2019 competition
618 on large-scale street view text with partial labeling-rrc-lsvt.
619 In *2019 International Conference on Document Analysis and*
620 *Recognition (ICDAR)*, pages 1557–1562. IEEE, 2019. 3
- 621 [34] Jun Tang, Zhibo Yang, Yongpan Wang, Qi Zheng, Yongchao
622 Xu, and Xiang Bai. Seglink++: Detecting dense and
623 arbitrary-shaped scene text by instance-aware component
624 grouping. *PR*, 96:106954–106964, 2019. 3
- 625 [35] Jingqun Tang, Wenqing Zhang, Hongye Liu, MingKun
626 Yang, Bo Jiang, Guanglong Hu, and Xiang Bai. Few could
627 be better than all: Feature sampling and grouping for scene
628 text detection. In *CVPR*, pages 4563–4572, 2022. 3
- 629 [36] Zhuotao Tian, Michelle Shu, Pengyuan Lyu, Ruiyu Li, Chao
630 Zhou, Xiaoyong Shen, and Jiaya Jia. Learning shape-aware
631 embedding for scene text detection. In *CVPR*, pages 4234–
632 4243, 2019. 3
- 633 [37] Andreas Veit, Tomas Matera, Lukás Neumann, Jiri Matas,
634 and Serge J. Belongie. Coco-text: Dataset and benchmark
635 for text detection and recognition in natural images. *CoRR*,
636 2016. 2, 3
- 637 [38] Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu,
638 Gang Yu, and Shuai Shao. Shape robust text detection with
639 progressive scale expansion network. In *CVPR*, pages 9336–
640 9345, 2019. 1, 3
- 641 [39] Xiaobing Wang, Yingying Jiang, Zhenbo Luo, Cheng-Lin
642 Liu, Hyunsoo Choi, and Sungjin Kim. Arbitrary shape scene
643 text detection with adaptive text region representation. In
644 *CVPR*, pages 6449–6458, 2019. 3
- 645 [40] Christian Wolf and Jean-Michel Jolian. Object count/area
646 graphs for the evaluation of object detection and segmen-
647 tation algorithms. *IJDAR*, 8(4):280–296, 2006. 6
- 648 [41] Longhuang Wu, Shangxuan Tian, Youxin Wang, and Pengfei
649 Xiong. CPN: complementary proposal network for uncon-
650 strained text detection. In *AAAI*, pages 6057–6065, 2024.
651 2
- 652 [42] Lele Xie, Yuliang Liu, Lianwen Jin, and Zecheng Xie.
653 Derpn: Taking a further step toward more general object de-
654 tection. In *AAAI*, pages 9046–9053, 2019. 3
- 655 [43] Chuhui Xue, Shijian Lu, and Fangneng Zhan. Accurate
656 scene text detection through border semantics awareness and
657 bootstrapping. In *ECCV*, pages 355–372, 2018. 3
- 658 [44] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu.
659 Detecting texts of arbitrary orientations in natural images. In
660 *CVPR*, pages 1083–1090, 2012. 1, 2
- 661 [45] Maoyuan Ye, Jing Zhang, Shanshan Zhao, Juhua Liu, Bo
662 Du, and Dacheng Tao. DPText-DETR: Towards better scene
663 text detection with dynamic points in transformer. In *AAAI*,
664 pages 3241–3249, 2023. 2, 3, 6, 8
- 665 [46] Qixiang Ye and David Doermann. Text detection and recog-
666 nition in imagery: A survey. *PAMI*, 37(7):1480–1500, 2015.
667 1
- 668 [47] Chucai Yi and YingLi Tian. Text string detection from natu-
669 ral scenes by structure-based partition and grouping. *TIP*, 20
670 (9):2594–2605, 2011. 1, 2
- 671 [48] Xu-Cheng Yin, Wei-Yi Pei, Jun Zhang, and Hong-Wei Hao.
672 Multi-orientation scene text detection with adaptive cluster-
673 ing. *PAMI*, 37(9):1930–1937, 2015. 1, 2
- 674 [49] Xu-Cheng Yin, Ze-Yu Zuo, Shu Tian, and Cheng-Lin Liu.
675 Text detection, tracking and recognition in video: A compre-
676 hensive survey. *TIP*, 25(6):2752–2773, 2016. 1
- 677 [50] Liu Yuliang, Jin Lianwen, Zhang Shuaitao, and Zhang
678 Sheng. Detecting curve text in the wild: New dataset and
679 new solution. *arXiv preprint arXiv:1712.02170*, 2017. 3
- 680 [51] Yu-Xiang Zeng, Jun-Wei Hsieh, Xin Li, and Ming-Ching
681 Chang. Mixnet: toward accurate detection of challenging
682 scene text in the wild. *arXiv preprint arXiv:2308.12817*,
683 2023. 2, 6
- 684 [52] Shi-Xue Zhang, Xiaobin Zhu, Chun Yang, Hongfa Wang,
685 and Xu-Cheng Yin. Adaptive boundary proposal network for
686 arbitrary shape text detection. In *ICCV*, pages 1305–1314,
687 2021. 3
- 688 [53] Shi-Xue Zhang, Chun Yang, Xiaobin Zhu, and Xu-Cheng
689 Yin. Arbitrary shape text detection via boundary transformer.
690 *TMM*, 26:1747–1760, 2023. 1, 2, 3, 5, 6, 7, 8
- 691 [54] Ying Zhang, Lionel Gueguen, Ilya Zharkov, Peter Zhang,
692 Keith Seifert, and Ben Kadlec. Uber-text: A large-scale
693 dataset for optical character recognition from street-level im-
694 agery. In *SUNw: Scene Understanding Workshop - CVPR*
695 2017, Hawaii, U.S.A., 2017. 2, 3
- 696 [55] Zheng Zhang, Chengquan Zhang, Wei Shen, Cong Yao,
697 Wenyu Liu, and Xiang Bai. Multi-oriented text detection
698 with fully convolutional networks. In *CVPR*, pages 4159–
699 4167, 2016. 3
- 700 [56] Jinzhi Zheng, Libo Zhang, Yanjun Wu, and Chen Zhao.
701 Bpdo: Boundary points dynamic optimization for arbitrary
702 shape scene text detection. In *ICASSP*, pages 5345–5349,
703 2024. 3
- 704 [57] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang
705 Zhou, Weiran He, and Jiajun Liang. EAST: An efficient and
706 accurate scene text detector. In *CVPR*, pages 2642–2651,
707 2017. 3
- 708 [58] Yingying Zhu, Cong Yao, and iang Bai. Scene text detection
709 and recognition: recent advances and future trends. *Frontiers*
710 *Comput. Sci.*, 10(1):19–36, 2016. 1
- 711