Platypus: A Generalized Specialist Model for Reading Text in Various Forms

Peng Wang⋆, Zhaohai Li⋆, Jun Tang⋆, Humen Zhong, Fei Huang,
Zhibo Yang†, and Cong Yao†

Alibaba Group, Beijing, China
{wdp0072012,tjbestehen,yangzhibo450,yaocong2010}@gmail.com
zhaohai.li@foxmail.com,zhonghumen@smail.nju.edu.cn,f.huang@alibaba-inc.com

Abstract. Reading text from images (either natural scenes or documents) has been a long-standing research topic for decades, due to the high technical challenge and wide application range. Previously, individual specialist models are developed to tackle the sub-tasks of text reading (e.g., scene text recognition, handwritten text recognition and mathematical expression recognition). However, such specialist models usually cannot effectively generalize across different sub-tasks. Recently, generalist models (such as GPT-4V), trained on tremendous data in a unified way, have shown enormous potential in reading text in various scenarios, but with the drawbacks of limited accuracy and low efficiency. In this work, we propose Platypus, a generalized specialist model for text reading. Specifically, Platypus combines the best of both worlds: being able recognize text of various forms with a single unified architecture, while achieving excellent accuracy and high efficiency. To better exploit the advantage of Platypus, we also construct a text reading dataset (called Worms), the images of which are curated from previous datasets and partially re-labeled. Experiments on standard benchmarks demonstrate the effectiveness and superiority of the proposed Platypus model. Model and data will be made publicly available at AdvancedLiterateMachinery.

Keywords: Text Reading · Generalized Specialist Model · OCR · Multimodal Large Language Model

1 Introduction

The task of reading text from images, whether in natural scenes or documents, has evolved substantially to support a broad spectrum of applications, from archiving historical documents to real-time translation services [49, 91]. While traditionally categorized within the domain of Optical Character Recognition (OCR), the ambition of this task extends beyond the conventional bounds of OCR to encompass a wider notion of text reading.

⋆ Equal contribution. † Corresponding author.

GT : ["T.F.", "GREEN", "CRASH", "FIRE", "RESCUE", "R.", "L", "1949", "CRASH", "RESCUE"]	GT : ["The", "Kitchen", "is", "the", "Heart", "of", "the", "HOME"]	GT : ["NEW", "YEAR", "2013", "Happy", "New", "Year!", "New", "BEGINNINGS"]
PaddleOCR : ["GREEN", "T.F.", "CRASH", "RESCUE", "CRASH", "1949"] (6/10)	PaddleOCR : ["Hearl", "is", "the", "of", "the", "HOME"] (5/8)	PaddleOCR : ["NEW", "YEAR", "2012", "New.", "BEGINNINGS", "COG"] (4/8)
EasyOCR : ["CRASH", "[", "1949", "GREEN", "TF", "1", "8"] (3/10)	EasyOCR : ["Ekitchen", "is", "leart", "the", "of", "the", "HOME"] (5/8)	EasyOCR : ["NEW", "YEAR", "901Q", "let", "NEW-", "BEGINNINCS", "Happy.", "Yearl"] (4/8)
GPT-4V : ["T.F.", "GREEN", "CRASH", "FIRE", "RESCUE", "1949"] (6/10)	GPT-4V : ["The", "Kitchen", "is", "the", "Heart", "of", "the", "HOME"] (8/8)	GPT-4V : ["NEW", "YEAR", "2013", "Happy", "New", "Year!", "NEW", "BEGINNINGS"] (8/8)
Qwen-VL-Plus : ["T.F.", "GREEN", "CRASH", "RESCUE", "1949"] (5/10)	Qwen-VL-Plus : ["The", "Kitchen", "is", "Heart", "of", "HOME"] (6/8)	Qwen-VL-Plus : ["Happy", "New", "Year!", "New", "Beginnings!"] (5/8)
Our : ["GREEN", "T.F.", "CRASH", "RESCUE", "FIRE", "1949", "RESCUE", "CRASH"] (8/10)	Our :["The", "Kitchen", "is", "Heart", "the", "the", "of", "HOME"] (8/8)	Our :["NEW", "YEAR", "2013", "Year!", "Happy", "New", "BEGINNINGS"] (8/8)

Fig. 1: Comparative cases of Platypus against MLLMs (GPT-4V [1], Qwen-VL-Plus [6]) and OCR tools (PaddleOCR, EasyOCR) on CAT Benchmark, highlighting word accuracy ratio (red brackets) and Platypus's RAT performance.

The field's progression has seen the advent of models specialized for distinct text reading sub-domains, such as scene text recognition [5,63,64], handwritten note conversion [2,40,56,79], and formula decoding [86]. Despite their proficiency within specific tasks, these specialized models struggle with the unpredictability of text in broader scenarios [5,18].

Traditional text reading systems typically consist of two distinct stages: text detection [21,51,65,68,74,75,90] followed by text recognition [7,15,63,64]. The overall performance is highly reliant on the accuracy of the initial text detection phase, creating a cascading effect when errors occur. Moreover, the emergence of Multi-modal Large Language Models(MLLMs) has contributed a new dimension to the field by offering models with broad text reading capabilities [6,36,66,83], albeit with the trade-off of computational efficiency and specialized precision [45].

In response to these challenges, we present Platypus, a novel generalized specialist model for text reading that offers a comprehensive solution for diverse text reading scenarios. Platypus embodies the amalgamation of high-precision specialized text reading models with the wide-ranging adaptability of multimodal approaches. The name Platypus is inspired by the animal known for its unique combination of traits; similarly, our Platypus model harmonizes the specificity of specialized models with the versatility of generalist approaches, making it adept at tackling the heterogeneity of text reading tasks.

Fig. 1 shows Platypus's qualitative comparisons with MLLMs and open-source OCR pipelines on the Curated Artistic Text (CAT) Benchmark. It highlights Platypus's superior accuracy in identifying words from artistically rendered text images, a challenging scenario for many contemporary OCR systems.

Furthermore, Fig. 2 contrasts the integrated approach of Platypus with traditional OCR systems, underscoring its comprehensive and adaptable nature for handling a multitude of text reading tasks.

Our contributions are manifold:

– We introduce Platypus, a single unified architecture adept at various text reading tasks, serving as a versatile and unified text reading solution.
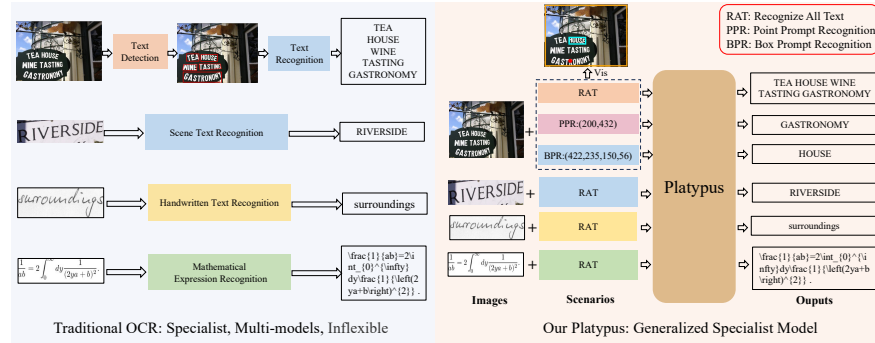
Fig. 2: Comparison of our Platypus with previous OCR systems.

- Platypus features a high degree of interactivity, providing users with the ability to specify areas for text recognition and select output granularity, enhancing usability and precision.
- A comprehensive text reading dataset, Worms, is curated and presented, supporting the extensive training and evaluation of Platypus.
- Platypus surpasses specialized text reading models and MLLMs in multiple text reading scenarios, establishing new state-of-the-art performances.

2 Related Work

Scene Text Spotting Methods The goal of scene text spotting is to simultaneously detect and recognize all the text in an image in natural scene. Early end-to-end text spotting methods [17,37,42,43,46,60,78] concatenate module of text detection and recognition with modified ROI module. Recently, transformer-based methods [24,29,47,57,84,87] achieve impressive achievement with their simple and efficient structures. DeepSolo [84] proposes learnable point queries to model text semantics and positions explicitly. The SPTS [47,57] series represent a text with a sequence combining center point and text.

Scene Text Recognition Methods Scene text recognition has been extensively researched [10,38,81,82], initially focusing on rigid OCR systems for document analysis. The emergence of convolutional neural networks (CNNs) has shifted the focus towards models capable of handling complex variations in natural scenes. Researchers have incorporated recurrent neural networks (RNNs) [22, 63] and attention mechanisms [11,34,64] to enhance performance. More recently, transformer-based models have been introduced, leveraging self-attention to capture long-range dependencies within text sequences [7,9,13–15,35,72,80,85].

Handwritten Text Recognition Methods The distinctiveness of handwritten text compared to printed text has necessitated specialized HTR models,

usually requiring intricate preprocessing to normalize the wide range of handwriting styles and strokes [8,52,73]. In the deep learning era, approaches such as Long Short-Term Memory (LSTM) networks have become prevalent, providing the ability to learn complex patterns in handwriting [2,40,79].

Mathematical Expression Recognition Methods MER poses unique challenges due to the spatial arrangements of symbols and the need to understand their semantic relationships. Approaches have extended beyond traditional symbol-based parsing [3] to employ neural network architectures that can directly translate images of formulas into LaTeX representations [33,76,86].

Multi-modal Large Language Models MLLMs represent a significant advancement in text processing by incorporating additional modalities such as vision and language [6, 16, 36, 41, 83]. These models have exhibited versatility across various applications. However, they often lack the fine-grained precision of specialized models [45, 66] and require substantial computational resources, limiting their practical deployment.

In summary, while text reading progress is notable, a holistic model for consistent performance across scenarios is needed. Platypus aims to provide a singular adaptive solution without the limitations of specialized and multimodal models.

3   Methodology

3.1   Task Unification

In addressing the diverse challenges of text reading from images, Platypus introduces a unified framework that encapsulates the various scenarios of text interpretation beyond traditional OCR boundaries. We categorize text reading tasks into four broad classes based on the source and presentation of the images.

Category We define our categories as natural scene full images, document full images, cropped text, and cropped formulas. Natural scene and document full images often contain a heterogeneous mixture of text styles and layouts, necessitating a more versatile approach to recognition. For these images, we introduce three distinct recognition scenarios:

- *RAT (Recognize All Text)*: This scenario involves recognizing all text within the entire image without any specific localization prompts.
- *PPR (Point Prompt Recognition)*: Text is recognized within a specified area surrounding a point prompt, allowing for targeted recognition of text.
- *BPR (Box Prompt Recognition)*: Text within a user-defined box area is recognized, which can be used to extract text from particular regions of interest.

For cropped text and formulas, the tasks involve direct interpretation of the text or mathematical expressions present in the images without the need for locational prompts, thus simplifying the recognition process.
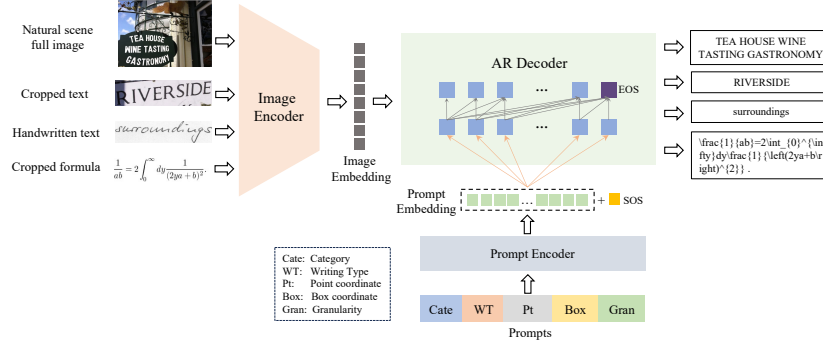
Fig. 3: Overall architecture of our proposed Platypus model.

Writing Type Our model is trained to differentiate between two primary writing types: printed and handwritten. These inherently distinct types of text require unique recognition strategies, which Platypus accommodates seamlessly.

Granularity We introduce granularity in recognition output to cater to the level of detail required by different text reading applications. This granularity is classified into two levels:

- *Word-level*: At this level, the model focuses on recognizing individual words as the atomic units of text.
- *Line-level*: Here, the model recognizes entire lines of text, which may contain multiple words or even sentences.

For natural scene and document full image tasks, granularity is a critical parameter that determines the output's specificity, while for cropped text and formula tasks, the granularity is implicitly determined by the cropping, thus requiring no additional specification.

Through this structured categorization, Platypus demonstrates a versatile capability to interpret text from a variety of sources and formats, bridging the gap between specialized text reading models and the demands of real-world text recognition challenges.

3.2   Model Architecture

Our Platypus architecture is an encoder-decoder framework benefiting from the principles of the SAM [30] model and incorporates a Prompt Encoder to encode various prompts. The overall architecture, illustrated in Fig. 3, combines these components into a cohesive system designed to robustly interpret text from various image forms.

Image Encoder The Image Encoder uses the pre-trained Swin-B Transformer [48] on the ImageNet 22k dataset to extract multi-scale visual features. The extracted features are further enhanced by a Feature Pyramid Network (FPN) [39], which provides a rich representation of the text at different scales and resolutions.

Prompt Encoder The Prompt Encoder is designed to handle the encoding of various prompts, which informs the model about the task category, writing type, output granularity, and the location information for the text to be recognized. We define the embedding for each category using a combination of position encoding and learned embeddings. For instance, the embedding for the task category is computed as follows:

$$\text{Embedding}_{\text{category}} = \text{PE} + E_{\text{category}} \tag{1}$$

where $E_{\text{category}} \in \mathbb{R}^{5 \times d}$ represents the learned embeddings for four explicit task categories—natural scene full images, document full images, cropped text, and cropped formulas plus one additional embedding to handle cases where the image category is not specified. This setup enhances the model's flexibility, allowing it to perform robustly even when the category is unknown during inference. Similarly, writing type and granularity prompts include an unspecified category to maximize the model's adaptability across various text recognition scenarios. Follwing [30], the prompt embeddings for point and box are generated in a similar way, but for the box representation, we adopt a quadrilateral form with four points (top-left, top-right, bottom-right, bottom-left) to provide more precise positioning.

Recognition Decoder Inspired by the Transformer [69] architecture, our Recognition Decoder is an autoregressive module generating the output text sequence. It integrates the visual features from the Image Encoder and the prompt embeddings from the Prompt Encoder to produce the final text recognition result. It consists of 6 transformer layers with 8 heads each, both initialized randomly.

3.3   Training Process

To effectively address the distinct challenges posed by different text reading scenarios, our training data is subdivided into two principal subsets. The first subset caters to full-image tasks such as natural scene full images and document full images, while the second subset is tailored for cropped-image tasks like cropped text and cropped formulas. The subsets not only differ in their image sizes but also utilize separate data loaders optimized for their specific needs.

   The loss function for our model is an amalgamation of four individual loss components, each corresponding to separate recognition tasks. The full-image subset incurs three types of losses: $\mathcal{L}_{RAT-f}$ for the RAT task, $\mathcal{L}_{PPR-f}$ for the PPR task, and $\mathcal{L}_{BPR-f}$ for the BPR task. In contrast, the cropped-image subset is associated with a single loss, $\mathcal{L}_{RAT-c}$, related solely to the RAT task. Each of these losses is weighted equally and set to 1 for simplicity.

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{RAT-f} + \lambda_2 \mathcal{L}_{PPR-f} + \lambda_3 \mathcal{L}_{BPR-f} + \lambda_4 \mathcal{L}_{RAT-c}, \qquad (2)$$

Each of these losses is computed using standard cross-entropy loss, which measures the predicted probabilities against the true labels of the text across our diverse dataset.

3.4   Inference Process

During inference, Platypus leverages a streamlined approach where the model generates predictions based on the input scene category and specific prompts provided. This enables the model to accurately interpret and recognize text from a wide range of input images without the direct need for detection or segmentation models.

4   Experiments

In this section, we conduct both qualitative and quantitative experiments on standard benchmarks of Scene Text Spotting (STS), Scene Text Recognition (STR), Handwritten Text Recognition (HTR), and Mathematical Expression Recognition (MER), to verify the effectiveness and advantages of Platypus.

4.1   Implementation Details

To train our Platypus for robust text reading across diverse scenarios, we meticulously prepared our dataset and conducted a two-phase training process.

Data Preparation As discussed in Sec. 3.3, we partition our training data into two subsets to accommodate the distinct dimensions of full-image and cropped-image tasks. Full-size images, encompassing natural scenes and documents, are resized to 1024 pixels on their longer edge. For cropped images, including text snippets and formulas, we resize the longer size to 768 pixels.

Pre-training The initial phase focused on pre-training with full image data, which presents complex, large-scale recognition challenges. For the RAT scenario, we set the point prompt to [0,0] and defined the box prompt as the full image size. For PPR, points are uniformly sampled within the text bounding boxes, while for BPR, ground-truth boundary boxes are perturbed with noise (10% of the box size, capped at 20 pixels). Besides, during each forward pass, multiple points and boxes are selected from a single image, capped at a maximum of eight prompts to allow the model to thoroughly learn the contextual relationships of text entities in the image. And granularity is selected randomly when both word and line annotations are available, guiding the corresponding ground truth selection. We employ a batch size of 2, augmented with instance-aware random cropping, rotations between -90°and 90°, random scaling, and color jittering. AdamW optimizer is utilized in the pre-training stage, with an initial learning rate of $5e^{-4}$, for a total of 1,000k steps, implementing a warm-up schedule for the first 5k steps and then linearly decaying the learning rate to zero.

Joint Training After the pre-training stage, we introduce cropped-image data into training. Separate data loaders are maintained for each subset, and batches alternated between full-image and cropped-image data to ensure a diversified learning experience. The data preparation and training techniques for the full-image subset remain consistent with the pre-training phase. For the cropped-image subset, recognizing the granularity of text is not necessary, as these images are pre-cropped to contain individual words or lines of text. Only the RAT scenario is applicable, with point prompts set to [0,0] and box prompts designed to cover the entire image size. A larger batch size of 16 is utilized for cropped images, and we apply common text image augmentation methods such as perspective and affine distortions, blurring, noise, and rotation to simulate various environmental conditions. The joint training continued with the AdamW optimizer, for 500k steps, using an initial learning rate of $3e^{-4}$ and the same learning rate schedule as established in the pre-training phase. All training is conducted on eight NVIDIA A100 GPUs.

Inference For full-image tasks, such as Scene Text Spotting (STS), we resize the longer size of image to 1024 and keep the aspect ratio. In the RAT recognition scenario, we adopt box prompt encompassing the entire image for evaluation. For the PPR scenario, the center point of each polygon annotation is adopted, and for the BPR scenario, the quadrilateral bounding box of groundtruth is used. For cropped-image tasks, such as Scene Text Recognition (STR), Handwritten Text Recognition (HTR), and Mathematical Expression Recognition (MER), which only entail the RAT scenario, we resize the longer size of image to 768, while keeping its aspect ratio. All evaluations are performed on one NVIDIA V100 GPU.

4.2   Comprehensive Text Reading Datasets (Worms)

The proposed model is trained in a unified way, thus, training data for STR, HTR, MER, and STS are combined for training. We collected and organized data from a multitude of public resources to build comprehensive text reading datasets, termed **Worms**. These datasets serve as the primary training set, akin to the favored diet of the Platypus in our model's training framework. Comprehensive details about the Worms dataset, including the specifics of each subset used, can be found in the Appendix.

4.3   Benchmarks and Evaluation Protocols

Scene Text Spotting (STS) To assess the ability scene text spotting of different methods, we evaluate on four popular scene text datasets, Total-Text [12], ICDAR 2013 (IC13) [28], ICDAR 2015 (IC15) [27], and CTW1500 [44].
Scene Text Recognition (STR) Experiments are conducted on six standard Latin scene text benchmarks, including 3 regular text datasets (IC13 [28], SVT [71], IIIT [54]) and 3 irregular ones (IC15 [27], SVTP [58], CUTE [61]).
Handwritten Text Recognition (HTR) To validate the performance on HTR, we also conduct experiments on handwritten text recognition datasets

Table 1: Comparison between Platypus and SOTA models on different text reading tasks, including MLLMs. Specialist models are limited to particular tasks, while Platypus, similar to MLLMs, is capable of accommodating multiple tasks.

SOTA Methods		STS	STR	HTR	MER	Params (M)
STS Models	SwinTextSpotter [23]	✓	✗	✗	✗	-
	SPTS [57]	✓	✗	✗	✗	36.5
	DeepSolo [84]	✓	✗	✗	✗	42.5
STR Models	MGP-STR [72]	✗	✓	✓	✗	148.0
	ABINet [15]	✗	✓	✓	✗	36.7
	PARSeq [7]	✗	✓	✓	✗	23.8
HTR Models	SeqCLR [2]	✗	✓	✓	✗	-
	PerSec-ViT [40]	✗	✓	✓	✗	-
	DiG-ViT-Base [79]	✗	✓	✓	✗	52
MER Models	LaTeX-OCR	✗	✗	✗	✓	25.5
MLLMs	mPLUG-Owl2 [83]	✓	✓	✓	✓	8200
	LLaVA1.5-7B [41]	✓	✓	✓	✓	7000
	GPT-4V [1]	✓	✓	✓	✓	-
Ours		✓	✓	✓	✓	118.8

with English (IAM [53] and CVL [31]) and French (RIMES [19]). In accordance with the methodologies employed in DiG [79], test sets of CVL and IAM include 12,012 and 13,752 cropped images respectively. For the RIMES dataset, 7,776 cropped images are used for testing.

Mathematical Expression Recognition (MER) For task of Mathematical Expression Recognition (MER), we use the benchmark of Latex-OCR[1] with cropped images of print formula. For convenience, we randomly select 1000 images of the test set for evaluation on MER.

Evaluation Metrics For benchmarks of STS, end-to-end text spotting methods are evaluated with H-mean of recall and precision on recognition of given lexicons. For MLLMs, the predictions and ground truths (GT) are split with spaces and then evaluated with precision and recall according to evaluation metric as defined in [66]. Precision stands for the percentages of correctly identified words to those generated by MLLMs, while recall is the ratio of correctly identified words to the total number of GT words. Thus H-mean is calculated through recall and precision.

For evaluation of the proposed Platypus on STS benchmarks, we calculated H-mean as the same as MLLMs, given the recognition scenarios such as RAT, PPR and BPR. While for benchmarks of STR and HTR, we evaluate performance with word accuracy ignoring case and symbols (WAICS) [26] as metric. For MER, we evaluate the performance using the Word Error Rate (WER) metric and Character Error Rate (CER).

---

[1] https://github.com/lukas-blecher/LaTeX-OCR

Table 2: Performance comparison between our method and SOTA text spotting models. For GPT-4V, we filter images with no results to evaluate the rest images marked with *valid*, and evaluation on all images is marked with *all*.

SOTA Methods	Scenarios	IC13		IC15		TotalText		CTW1500	
Lexicon	-	G	None	G	None	None	None	None	None
Granularity	-	word	line	word	line	word	line	word	line
ABCNet v2 [46]	Spotting	-	-	73.0	-	70.4	-	-	57.5
MANGO [59]	Spotting	88.7	-	73.9	-	72.9	-	-	58.9
SwinTextSpotter [23]	Spotting	-	-	70.5	-	74.3	-	-	51.8
TESTR [87]	Spotting	-	-	73.6	-	73.3	-	-	56.0
SPTS [57]	Spotting	88.5	-	65.8	-	74.2	-	-	63.6
ESTextSpotter [24]	Spotting	-	-	78.1	-	80.8	-	-	66.0
DeepSolo [84]	Spotting	90.1	-	79.1	-	82.5	-	-	64.2
GPT-4V [1](*all*)	RAT	88.3	52.1	34.4	20.6	70.7	45.2	75.7	51.7
GPT-4V [1](*valid*)	RAT	88.8	52.1	38.8	20.1	71.1	45.5	75.9	51.9
	RAT	92.7	83.8	68.3	63.1	78.9	58.0	80.9	73.5
Ours	PPR	94.6	85.9	85.7	78.7	88.3	76.9	86.6	75.6
	BPR	**96.9**	**86.0**	**88.2**	**77.9**	**93.0**	**77.4**	**91.9**	**76.1**

4.4   Comparison with State-of-the-Art Specialists and MLLMs

Overall Capability Comparisons on Extensive Text Reading Benchmarks We firstly make capability comparisons between our models and state-of-the-art (SOTA) models on different text reading tasks, including MLLMs. As can be concluded in Tab. 1, previous specialist SOTA models can only prevail on specific text reading tasks. While the MLLMs and our Platypus are able to compete on various text reading tasks as generalist models.

Comparison on Scene Text Spotting (STS) To further demonstrate the performance on natural scene full images, we perform evaluation on Scene Text Spotting benchmarks, in comparison with MLLMs and previous SOTA scene text spotters. All the evaluations of MLLMs and the proposed Platypus model are under metric defined in Sec. 4.3 and the comparison results are displayed in Tab. 2. For comparison with different prompts on Platypus, box prompt (i.e. BPR) and point prompt (i.e. PPR) with clear indication where text may be located, perform better than fullbox prompt (i.e. RAT). Compared with specialist models on STS, which are usually fine-tuned on the specific dataset and support only one granularity, Platypus can read text with word-level and line-level granularity. Besides, the performance of Platypus in PPR and BPR settings is better than that of SOTA on four benchmarks, and Platypus with RAT scenario is also comparable with specialist models on STS. When in comparison with MLLM GPT-4V, the proposed Platypus surpasses it with a large gap even in the RAT setting.

Table 3: Performance comparison between our method and SOTA scene text recognition models. Note that the results of MLLMs with † are obtained from [45].

Methods		IIIT5k	SVT	IC13	IC15	SVTP	CUTE80	Average
Specialist Models	ViTSTR [4]	88.4	87.7	93.2	78.5	81.8	81.3	85.6
	SRN [85]	94.8	91.5	95.5	82.7	85.1	87.8	90.4
	ABINet [15]	96.2	93.5	97.4	86.0	89.3	89.2	92.6
	TrOCR [35]	91.0	93.2	98.3	84.0	91.0	89.6	90.2
	MGP-STR [72]	96.4	97.3	94.7	87.2	91.0	90.3	93.3
	PARSeq [7]	99.1	97.9	98.3	90.7	**95.7**	98.3	96.4
Generalist Models	mPLUG-Owl2†	80.9	69.6	79.8	53.9	53.5	74.8	70.3
	LLaVA1.5-7B†	84.2	85.7	86.4	71.9	79.8	82.7	81.0
	UniDoc†	91.9	89.2	90.9	78.0	80.3	88.2	86.9
	Monkey†	83.7	75.1	85.4	53.4	58.4	73.9	72.9
	GPT-4V [1](*all*)	37.3	66.3	66.1	39.0	52.4	68.3	46.2
	GPT-4V [1](*valid*)	69.4	91.0	90.9	70.8	86.9	93.4	76.7
Ours		**99.1**	**98.1**	**98.5**	**90.7**	95.1	**98.9**	**96.4**

Comparison on Scene Text Recognition (STR) We also compare our proposed Platypus with SOTA models and MLLMs on STR, and the results on 6 standard benchmarks are summarized in Tab. 3. As can be seen from the results, though Platypus is a unified model for 4 tasks, Platypus can also achieve SOTA results surpassing not only generalist models such as GPT-4V and Monkey but also STR specialist models (i.e., ABINet [15], MGP-STR [72] and PARSeq [7]) on fine-grained task as STR.

Comparison on Handwritten Text Recognition (HTR) To validate the adaptability of Platypus on various types such as handwritten text, we conduct evaluation on HTR benchmarks, comparing with previous SOTA methods and MLLMs (i.e., GPT-4V). As depicted in Tab. 4, GPT-4V suffers from poor accuracy on 3 Latin benchmarks of HTR, indicating poor adaptability to handwritten text. The proposed Platypus instead sets a new SOTA on three benchmarks, and the performance gains are 8.2%, 0.1% and 1.2%, compared with the previous SOTA on IAM, CVL and RIMES respectively.

Comparison on Mathematical Expression Recognition (MER) We also compare Platypus with MLLMs and open source formula recognition method on MER benchmark LaTeX-OCR. As seen in Tab. 5, on structured text as formula, GPT-4V has poor adaptation performance. While the performance of Platypus can also surpass LaTeX-OCR for 1.5% and 1.8% in metric CER and WER respectively.

Comparisons of efficiency (FPS) The comparisons of inference speed (evaluated on a single V100 GPU) are depicted in Tab. 9. As can be observed, Platypus exhibits superior inference speed, surpassing MLLMs (*e.g.* mPLUG-Owl2) and

Table 4: Performance comparison between our method and SOTA handwritten text recognition models.

Methods	IAM	CVL	RIMES
SeqCLR [2]	79.9	77.8	92.4
PerSec-ViT [40]	83.7	82.9	-
DiG-ViT-Base [79]	87.0	91.3	-
GPT-4V [1](*all*)	40.4	26.9	15.9
GPT-4V [1](*valid*)	50.7	38.3	26.0
Ours	**96.4**	**91.4**	**93.6**

Table 5: Performance comparison between Platypus and SOTA mathematical expression recognition models.

Methods	LaTeX-OCR	
	CER($\downarrow$)	WER($\downarrow$)
LaTeX-OCR	8.7	9.6
GPT-4V [1](*all*)	47.2	50.1
GPT-4V [1](*valid*)	47.8	50.7
Ours	**7.2**	**7.8**

achieving higher efficiency than specialist models on STS, HTR and MER tasks. In particular, Platypus even runs ×3 faster than SPTS, mainly attributed to its efficient single-pass decoding.

4.5   Analyses

Exploring the Effectiveness of Prompts The efficacy of prompts in guiding text recognition was extensively examined. In the STS benchmark, point, box, and granularity prompts were invaluable for full-image tasks. For cropped-image tasks, we assessed category and writing type prompts across four scenarios: no prompts, only category, only writing type, and both combined (see Tab. 6). The combined prompts led to the best recognition accuracy, especially in HTR and MER tasks, with MER being more sensitive to category prompts. STR showed reasonable accuracy even without prompts, indicating robustness. These findings highlight the adaptive utility of prompts in various recognition contexts.

Evaluating Different Pre-training Approaches We explored three pre-training strategies, as shown in Tab. 7. The methods included (1) joint training without pre-training, (2) pre-training on cropped-image data followed by joint training, and (3) pre-training on full-image data followed by joint training with cropped-image data. Each method underwent a 200k-step training regimen. Evaluation on selected subsets from STS, STR, HTR, and MER benchmarks revealed significant accuracy gains when full-image pre-training was applied, especially notable in word-level IC15 and line-level CTW1500 for STS. Models pre-trained on full images showed marked improvements in STR, HTR, and MER tasks, affirming the benefit of this approach. Notably, the lack of pre-training led to substantial performance deficits, underscoring its importance.

Impact of Image Size To ascertain the influence of image resolution on text recognition, we conducted experiments with two distinct image size settings. For full-image tasks, we compared the performance of models trained with a maximum edge length of 768 pixels against those trained with 1024 pixels, while

Table 6: The effectiveness of category prompt and writing type prompt.

Prompt		STR(ACC ↑)			HTR(ACC ↑)	MER(WER ↓)
Category	Writing type	IIIT5k	IC15	CUTE80	IAM	LaTeX-OCR
✗	✗	98.8	90.5	98.5	88.6	8.5
✓	✗	98.8	90.5	98.6	94.1	7.9
✗	✓	98.8	90.7	98.3	96.2	8.5
✓	✓	99.1	90.7	98.9	96.4	7.8

Table 7: Performance comparison between different pre-training strategies.

Strategy	STS(H-mean ↑)		STR(ACC ↑)			HTR(ACC ↑)	MER(WER ↓)
	IC15	CTW1500	IIIT5k	IC15	CUTE80	IAM	LaTeX-OCR
w/o pre-training	0.1	0.1	2.3	2.6	1.1	4.1	98.7
cropped image pre-training	61.2	63.0	98.5	88.1	96.5	92.3	8.7
full image pre-training	86.7	76.1	98.7	90.2	97.5	95.5	8.3

Table 8: Performance comparison between different training image sizes.

full-img size	cropped-img size	STS(H-mean ↑)		STR(ACC ↑)			HTR(ACC ↑)	MER(WER ↓)
		IC15	CTW1500	IIIT5k	IC15	CUTE80	IAM	LaTeX-OCR
1024	768	88.2	76.1	99.1	90.7	98.9	96.4	7.8
768	512	82.4	75.4	98.9	90.5	98.9	96.3	8.3

for cropped-image tasks, we contrasted the sizes of 512 and 768 pixels. Following a pre-training phase of 1,000k steps, both models underwent 500k steps of joint training. The outcome, presented in Tab. 8, indicates an improvement in accuracy across all tasks when larger image sizes are used, with full-image tasks benefiting the most. This confirms our hypothesis that higher resolutions significantly aid in recognizing text within larger images, with a less pronounced but still considerable effect observed for cropped-image tasks.

Comparisons on Curated Artistic Text (CAT) Benchmark 100 images from LAION-5B [62] were curated and annotated to evaluate text reading pipelines on multi-orientation, occluded, overlapped, and artistic text. We make comparisons among proposed Platypus, MLLMs (GPT-4V and Qwen-VL-Plus), and open source OCR pipelines (PaddleOCR[2] and EasyOCR[3]) on this benchmark. Note that, the results of MLLMs are obtained from the APIs, and the results of open-source OCR pipelines are produced from the officially released codes and models. As displayed in Tab. 10, Platypus outperforms PaddleOCR and EasyOCR by 26.7% and 35.2% H-mean, respectively. Compared to GPT-4V on RAT, Platypus achieves comparable performance and better results on PPR and BPR scenarios. The qualitative comparisons are shown in Fig. 1. We can conclude through the results that the proposed Platypus has better gener-

---

[2] https://github.com/PaddlePaddle/PaddleOCR
[3] https://github.com/JaidedAI/EasyOCR

Table 9: Comparison of FPS.

SOTA Methods		STS	STR	HTR	MER
		TotalText	IC13	IAM	Latex-OCR
STS	SPTS	0.3	-	-	-
STR	PARSeq	-	25.9	-	-
HTR	DiG-ViT-Base	-	-	4.6	-
MER	LaTeX-OCR	-	-	-	1.3
MLLMs	mPLUG-Owl2	0.4	2.8	2.2	0.4
	GPT-4V	0.3	0.7	0.8	0.4
Ours		**1.0**	11.9	**13.0**	**2.1**

Table 10: Performance comparison on CAT.

Methods	Scenarios	Recall	Precision	H-mean
PaddleOCR	Spotting	43.7	62.6	51.4
EasyOCR	Spotting	38.5	48.5	42.9
GPT-4V [1]	RAT	80.1	85.1	82.5
Qwen-VL-Plus [6]	RAT	69.2	81.9	75.0
Ours	RAT	72.7	84.4	78.1
	PPR	84.3	84.3	84.3
	BPR	**89.8**	**89.8**	**89.8**
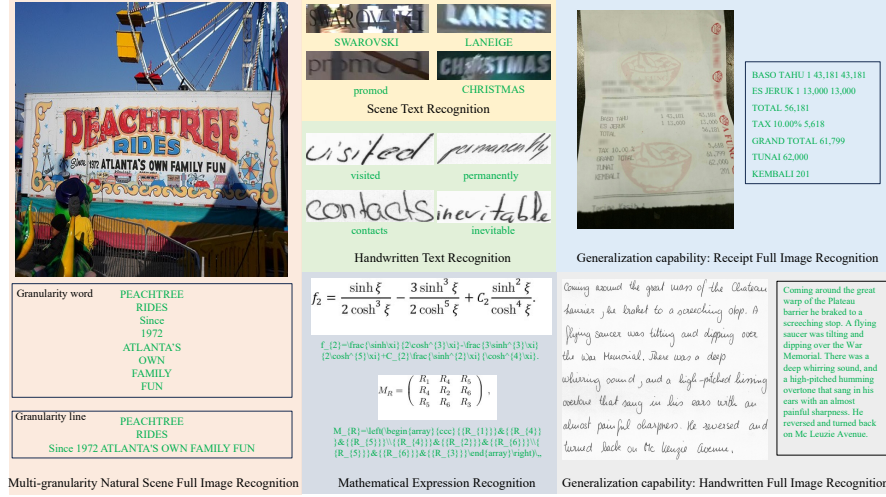
Fig. 4: Qualitative results of Platypus. All images are derived from public datasets.

alization ability on CAT benchmark with multi-orientation, artistic shape and overlapped issues, compared to OCR pipelines and even MLLMs.

Qualitative Results We display some qualitative results of Platypus in Fig. 4. The qualitative results further demonstrate the ability of Platypus on multi-granularity natural scene full image recognition, scene text and handwritten text recognition, mathematical expression recognition, and generalized capability on unseen images such as receipt full images and handwritten full images.

5 Conclusion

In summary, Platypus unifies text reading across various forms and complexities, achieving SOTA performance and outperforming specialized and multimodal models. The interactive prompt mechanism enhances precision, paving the way for future improvements. Although our evaluations focused on English text, there is potential for multi-language expansion. Future efforts will target linguistic diversity and refine the model's real-world text interpretation capabilities.

References

1. Gpt-4v(ision) system card (2023)
2. Aberdam, A., Litman, R., Tsiper, S., Anschel, O., Slossberg, R., Mazor, S., Manmatha, R., Perona, P.: Sequence-to-sequence contrastive learning for text recognition. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 15297–15307 (2020)
3. Alvaro, F., Sánchez, J.A., Benedí, J.M.: Recognition of on-line handwritten mathematical expressions using 2d stochastic context-free grammars and hidden markov models. Pattern Recognit. Lett. **35**, 58–67 (2014)
4. Atienza, R.: Vision transformer for fast and efficient scene text recognition. In: IEEE International Conference on Document Analysis and Recognition (2021)
5. Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S.J., Lee, H.: What is wrong with scene text recognition model comparisons? dataset and model analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
6. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023)
7. Bautista, D., Atienza, R.: Scene text recognition with permuted autoregressive sequence models. In: European Conference on Computer Vision (2022)
8. Bhunia, A.K., Das, A., Bhunia, A.K., Kishore, P.S.R., Roy, P.P.: Handwriting recognition in low-resource scripts using adversarial learning. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4762–4771 (2018)
9. Cheng, C., Wang, P., Da, C., Zheng, Q., Yao, C.: Lister: Neighbor decoding for length-insensitive scene text recognition. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 19484–19494 (2023)
10. Cheng, C., Wang, P., Da, C., Zheng, Q., Yao, C.: Lister: Neighbor decoding for length-insensitive scene text recognition. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
11. Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., Zhou, S.: Focusing attention: Towards accurate text recognition in natural images. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 5086–5094 (2017)
12. Ch'ng, C.K., Chan, C.S.: Total-text: A comprehensive dataset for scene text detection and recognition. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 935–942 (2017). `https://doi.org/10.1109/ICDAR.2017.157`
13. Da, C., Wang, P., Yao, C.: Levenshtein ocr. In: European Conference on Computer Vision (2022)
14. Da, C., Wang, P., Yao, C.: Multi-granularity prediction with learnable fusion for scene text recognition. CoRR **abs/2307.13244** (2023), `https://doi.org/10.48550/arXiv.2307.13244`
15. Fang, S., Xie, H., Wang, Y., Mao, Z., Zhang, Y.: Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 7094–7103 (2021)
16. Feng, H., Wang, Z., Tang, J., Lu, J., gang Zhou, W., Li, H., Huang, C.: Unidoc: A universal large multimodal model for simultaneous text detection, recognition, spotting and understanding. ArXiv **abs/2308.11592** (2023)

17. Feng, W., He, W., Yin, F., Zhang, X.Y., Liu, C.L.: Textdragon: An end-to-end framework for arbitrary shaped text spotting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)

18. Garcia-Bordils, S., Mafla, A., Biten, A.F., Nuriel, O., Aberdam, A., Mazor, S., Litman, R., Karatzas, D.: Out-of-vocabulary challenge report. In: ECCV Workshops (2022)

19. Grosicki, E., Carré, M., Brodin, J., Geoffrois, E.: Results of the RIMES evaluation campaign for handwritten mail processing. In: 10th International Conference on Document Analysis and Recognition, ICDAR 2009, Barcelona, Spain, 26-29 July 2009. pp. 941–945. IEEE Computer Society (2009)

20. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2315–2324 (2016)

21. He, M., Liao, M., Yang, Z., Zhong, H., Tang, J., Cheng, W., Yao, C., Wang, Y., Bai, X.: Most: A multi-oriented scene text detector with localization refinement. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8809–8818 (2021)

22. Hu, W., Cai, X., Hou, J., Yi, S., Lin, Z.: Gtc: Guided training of ctc towards efficient and accurate scene text recognition. ArXiv **abs/2002.01276** (2020)

23. Huang, M., Liu, Y., Peng, Z., Liu, C., Lin, D., Zhu, S., Yuan, N.J., Ding, K., Jin, L.: Swintextspotter: Scene text spotting via better synergy between text detection and text recognition. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4583–4593 (2022)

24. Huang, M., Zhang, J., Peng, D., Lu, H., Huang, C., Liu, Y., Bai, X., Jin, L.: Estextspotter: Towards better scene text spotting with explicit synergy in transformer. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 19438–19448 (2023)

25. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and artificial neural networks for natural scene text recognition. NIPS Deep Learning Workshop (2014)

26. Jiang, Q., Wang, J., Peng, D., Liu, C., Jin, L.: Revisiting scene text recognition: A data perspective. In: Proceedings of the IEEE/CVF international conference on computer vision (2023)

27. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S.K., Bagdanov, A.D., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., Shafait, F., Uchida, S., Valveny, E.: ICDAR 2015 competition on robust reading. In: ICDAR. pp. 1156–1160 (2015)

28. Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L.G., Mestre, S.R., Mas, J., Mota, D.F., Almazán, J., de las Heras, L.: ICDAR 2013 robust reading competition. In: ICDAR. pp. 1484–1493 (2013)

29. Kil, T.H., Kim, S., Seo, S., Kim, Y., Kim, D.: Towards unified scene text spotting based on sequence generation. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 15223–15232 (2023)

30. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.B.: Segment anything. pp. 3992–4003 (2023)

31. Kleber, F., Fiel, S., Diem, M., Sablatnig, R.: Cvl-database: An off-line database for writer retrieval, writer identification and word spotting. In: ICDAR. pp. 560–564 (2013)

32. Krylov, I., Nosov, S., Sovrasov, V.: Open images V5 text annotation and yet another mask text spotter. In: Balasubramanian, V.N., Tsang, I.W. (eds.) Asian Conference on Machine Learning. vol. 157, pp. 379–389 (2021)
33. Le, A.D.: Recognizing handwritten mathematical expressions via paired dual loss attention network and printed mathematical expressions. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 2413–2418 (2020)
34. Lee, C.Y., Osindero, S.: Recursive recurrent nets with attention modeling for ocr in the wild. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2231–2239 (2016)
35. Li, M., Lv, T., Cui, L., Lu, Y., Florêncio, D.A.F., Zhang, C., Li, Z., Wei, F.: Trocr: Transformer-based optical character recognition with pre-trained models. In: AAAI Conference on Artificial Intelligence (2021)
36. Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., Bai, X.: Monkey: Image resolution and text label are important things for large multi-modal models. ArXiv **abs/2311.06607** (2023)
37. Liao, M., Lyu, P., He, M., Yao, C., Wu, W., Bai, X.: Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. IEEE transactions on pattern analysis and machine intelligence **43**(2), 532—548 (February 2021). `https://doi.org/10.1109/tpami.2019.2937086`
38. 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 **45**(1), 919–931 (2022)
39. Lin, T.Y., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 936–944 (2016)
40. Liu, H., Wang, B., Bao, Z., Xue, M., Kang, S., Jiang, D., Liu, Y., Ren, B.: Perceiving stroke-semantic context: Hierarchical contrastive learning for robust scene text recognition. In: AAAI Conference on Artificial Intelligence (2022)
41. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. ArXiv **abs/2310.03744** (2023)
42. Liu, X., Liang, D., Yan, S., Chen, D., Qiao, Y., Yan, J.: Fots: Fast oriented text spotting with a unified network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
43. Liu, Y., Chen, H., Shen, C., He, T., Jin, L., Wang, L.: Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
44. Liu, Y., Jin, L., Zhang, S., Zhang, S.: Detecting curve text in the wild: New dataset and new solution. ArXiv **abs/1712.02170** (2017)
45. Liu, Y., Li, Z., Li, H., Yu, W., Huang, M., Peng, D., Liu, M., Chen, M., Li, C., Jin, L., Bai, X.: On the hidden mystery of ocr in large multimodal models. ArXiv **abs/2305.07895** (2023)
46. Liu, Y., Shen, C., Jin, L., He, T., Chen, P., Liu, C., Chen, H.: Abcnet v2: Adaptive bezier-curve network for real-time end-to-end text spotting. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**, 8048–8064 (2021)
47. Liu, Y., Zhang, J., Peng, D., Huang, M., Wang, X., Tang, J., Huang, C., Lin, D., Shen, C., Bai, X., Jin, L.: Spts v2: Single-point scene text spotting. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**, 15665–15679 (2023)
48. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 9992–10002 (2021)

49. Long, S., He, X., Yao, C.: Scene text detection and recognition: The deep learning era. International Journal of Computer Vision **129**, 161 – 184 (2018)

50. Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., Raptis, M.: Towards end-to-end unified scene text detection and layout analysis. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1039–1049 (2022)

51. Long, S., Ruan, J., Zhang, W., He, X., Wu, W., Yao, C.: Textsnake: A flexible representation for detecting text of arbitrary shapes. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)

52. Luo, C., Zhu, Y., Jin, L., Wang, Y.: Learn to augment: Joint data augmentation and network optimization for text recognition. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 13743–13752 (2020)

53. Marti, U., Bunke, H.: The iam-database: an english sentence database for offline handwriting recognition. Int. J. Document Anal. Recognit. **5**(1), 39–46 (2002)

54. Mishra, A., Alahari, K., Jawahar, C.V.: Scene text recognition using higher order language priors. In: BMVC. pp. 1–11 (2012)

55. Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., Luo, Z., Pal, U., Rigaud, C., Chazalon, J., Khlif, W., Luqman, M.M., Burie, J., Liu, C., Ogier, J.: 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. 01, pp. 1454–1459 (2017). `https://doi.org/10.1109/ICDAR.2017.237`

56. Nuriel, O., Fogel, S., Litman, R.: Textadain: Paying attention to shortcut learning in text recognizers. In: European Conference on Computer Vision (2021)

57. Peng, D., Wang, X., Liu, Y., Zhang, J., Huang, M., Lai, S., Zhu, S., Li, J., Lin, D., Shen, C., Jin, L.: Spts: Single-point text spotting. Proceedings of the 30th ACM International Conference on Multimedia (2021)

58. Phan, T.Q., Shivakumara, P., Tian, S., Tan, C.L.: Recognizing text with perspective distortion in natural scenes. In: ICCV. pp. 569–576 (2013)

59. Qiao, L., Chen, Y., Cheng, Z., Xu, Y., Niu, Y., Pu, S., Wu, F.: Mango: A mask attention guided one-stage scene text spotter. In: AAAI Conference on Artificial Intelligence (2020)

60. Qin, S., Bissacco, A., Raptis, M., Fujii, Y., Xiao, Y.: Towards unconstrained end-to-end text spotting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)

61. Risnumawan, A., Shivakumara, P., Chan, C.S., Tan, C.L.: A robust arbitrary text detection system for natural scene images. Expert Syst. Appl. **41**(18), 8027–8048 (2014)

62. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models. ArXiv **abs/2210.08402** (2022)

63. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(11), 2298–2304 (2017). `https://doi.org/10.1109/TPAMI.2016.2646371`

64. Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: Aster: An attentional scene text recognizer with flexible rectification. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(9), 2035–2048 (2019). `https://doi.org/10.1109/TPAMI.2018.2848939`

65. Shi, B., Bai, X., Belongie, S.J.: Detecting oriented text in natural images by linking segments. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3482–3490 (2017)
66. Shi, Y., Peng, D., Liao, W., Lin, Z., Chen, X., Liu, C., Zhang, Y., Jin, L.: Exploring ocr capabilities of gpt-4v(ision) : A quantitative and in-depth evaluation. ArXiv **abs/2310.16809** (2023)
67. 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: CVPR. pp. 8802–8812 (2021)
68. Tang, J., Yang, Z., Wang, Y., Zheng, Q., Xu, Y., Bai, X.: Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping. Pattern Recognit. **96** (2019)
69. Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Neural Information Processing Systems (2017)
70. Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S.J.: Coco-text: Dataset and benchmark for text detection and recognition in natural images. CoRR **abs/1601.07140** (2016)
71. Wang, K., Babenko, B., Belongie, S.J.: End-to-end scene text recognition. In: ICCV. pp. 1457–1464 (2011)
72. Wang, P., Da, C., Yao, C.: Multi-granularity prediction for scene text recognition. In: European Conference on Computer Vision (2022)
73. Wang, T., Zhu, Y., Jin, L., Luo, C., Chen, X., Wu, Y., Wang, Q., Cai, M.: Decoupled attention network for text recognition. In: AAAI Conference on Artificial Intelligence (2019)
74. Wang, W., Xie, E., Li, X., Hou, W., Lu, T., Yu, G., Shao, S.: Shape robust text detection with progressive scale expansion network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
75. Wang, W., Xie, E., Song, X., Zang, Y., Wang, W., Lu, T., Yu, G., Shen, C.: Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8440–8449 (2019)
76. Wu, J.W., Yin, F., Zhang, Y., Zhang, X.Y., Liu, C.L.: Handwritten mathematical expression recognition via paired adversarial learning. International Journal of Computer Vision **128**, 2386 – 2401 (2020)
77. Xie, Y., Mouchère, H., Liwicki, F.S., Rakesh, S., Saini, R., Nakagawa, M., Nguyen, C.T., Truong, T.N.: Icdar 2023 crohme: Competition on recognition of handwritten mathematical expressions. In: IEEE International Conference on Document Analysis and Recognition (2023)
78. Xing, L., Tian, Z., Huang, W., Scott, M.R.: Convolutional character networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
79. Yang, M., Liao, M., Lu, P., Wang, J., Zhu, S., Luo, H., Tian, Q., Bai, X.: Reading and writing: Discriminative and generative modeling for self-supervised text recognition. Proceedings of the 30th ACM International Conference on Multimedia (2022)
80. Yang, M., Yang, B., Liao, M., Zhu, Y., Bai, X.: Class-aware mask-guided feature refinement for scene text recognition. Pattern Recognit. **149**, 110244 (2024)

81. Yao, C., Bai, X., Shi, B., Liu, W.: Strokelets: A learned multi-scale representation for scene text recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4042–4049 (2014)
82. Yao, C., Zhang, X., Bai, X., Liu, W., Ma, Y., Tu, Z.: Detecting texts of arbitrary orientations in natural images (2012)
83. Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Dan, Y., Zhao, C., Xu, G., Li, C., Tian, J., Qi, Q., Zhang, J., Huang, F.: mplug-docowl: Modularized multimodal large language model for document understanding. ArXiv **abs/2307.02499** (2023)
84. 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. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 19348–19357 (2022)
85. Yu, D., Li, X., Zhang, C., Liu, T., Han, J., Liu, J., Ding, E.: Towards accurate scene text recognition with semantic reasoning networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
86. Yuan, Y., Liu, X., Dikubab, W., Liu, H., Ji, Z., Wu, Z., Bai, X.: Syntax-aware network for handwritten mathematical expression recognition. arXiv preprint arXiv:2203.01601 (2022)
87. Zhang, X., Su, Y., Tripathi, S., Tu, Z.: Text spotting transformers. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 9509–9518 (2022)
88. Zhang, Y., Gueguen, L., Zharkov, I., Zhang, P., Seifert, K., Kadlec, B.: Uber-text: A large-scale dataset for optical character recognition from street-level imagery. In: SUNw: Scene Understanding Workshop-CVPR. vol. 2017, pp. 1–5 (2017)
89. Zhong, X., Tang, J., Jimeno-Yepes, A.: Publaynet: Largest dataset ever for document layout analysis. 2019 International Conference on Document Analysis and Recognition (ICDAR) pp. 1015–1022 (2019)
90. Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., Liang, J.: East: An efficient and accurate scene text detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
91. Zhu, Y., Yao, C., Bai, X.: Scene text detection and recognition: recent advances and future trends. Frontiers of Computer Science **10**, 19 – 36 (2015)