

Towards End-to-End Unified Scene Text Detection and Layout Analysis

Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, Michalis Raptis
Google Research

{longshangbang, qinb, dpantelev, bissacco, yasuhisaf, mraptis}@google.com

Abstract

Scene text detection and document layout analysis have long been treated as two separate tasks in different image domains. In this paper, we bring them together and introduce the task of unified scene text detection and layout analysis. The first hierarchical scene text dataset is introduced to enable this novel research task. We also propose a novel method that is able to simultaneously detect scene text and form text clusters in a unified way. Comprehensive experiments show that our unified model achieves better performance than multiple well-designed baseline methods. Additionally, this model achieves state-of-the-art results on multiple scene text detection datasets without the need of complex post-processing. Dataset and code: <https://github.com/google-research-datasets/hiertext>.

1. Introduction

The ability to read and understand text in natural scenes and digital documents plays an important role in anthropocentric applications of computer vision. While state-of-the-art text detection systems such as [44, 61] excel at localizing individual text entities, visual text understanding [2] requires comprehension of the semantic and geometric layout [5, 7] of the textual content. In the current literature, most works focus on the individual tasks of text entities detection [3, 18, 61] and layout analysis [26, 58] in a separate way, devoting all the power of deep learning models to task-specific performance. We argue that joint treatment of these two closely related problems can result not only in simpler and more efficient models, but also models that are more accurate across all tasks. Additionally, an all-in-one, unified text and layout detection architecture can become indispensable for text reasoning tasks such as text-based VQA [4, 47] and image captioning [57].

The division between text detection and geometric layout analysis tasks has led to parallel and separate research directions. Text detectors [14, 18, 40, 61] usually treat word-level annotations, i.e. sequence of characters not interrupted by



Figure 1. **Top:** We introduce the task of unified text detection and layout analysis, and collect a dataset called **HierText** with hierarchical annotations. Blue boxes are word level bounding boxes. Yellow boundaries mark the ground-truth clustering of text. Line-level annotations and transcriptions are not visualized to avoid overcrowding. **Bottom:** We propose an end-to-end model called **Unified Detector** which can simultaneously detect text as masks and further group them into clusters. The model produces masks for text detection and an affinity matrix to cluster text lines, both in an end-to-end fashion without complex post-processing. We visualize the layout analysis results by coloring text line masks according to their clusters.

space, as the only supervision signal. Conversely, geometric layout analysis algorithms [2, 26, 54, 58, 62] focus on digital documents and either assume word-level text information as given [2, 54, 58] or directly predict geometric structures without reasoning for their atomic elements [62]. We ask: *Can there be a reconciliation of text entity detection and geometric layout analysis? Can geometric layout analysis target both natural scenes and digital documents?* These questions are important given their relevance in real-world

applications, such as screen readers for visually impaired and image-based translation.

Our work aims to unify text detection and geometric layout analysis. We introduce a new image dataset called **HierText**. It is the first dataset featuring hierarchical annotations of text in natural scenes and documents (Fig. 1, top). The dataset contains high quality *word*, *line*, and *paragraph* level annotations. “Text lines” are defined as logically connected sequences of words that are aligned in spatial proximity. Text lines that belong to the same semantic topic and are geometrically coherent form “paragraphs”. Images in HierText average more than 100 words per image, twice denser than the current highest density scene text dataset [48]. Experimental results show our dataset is complementary to other public datasets [10, 11, 19, 22, 37, 38, 48, 49, 59, 60] for the standalone text detection task.

In addition to HierText, we present a novel model, **Unified Detector**, that simultaneously detects text entities and performs layout analysis by grouping text entities, as illustrated in the bottom of Fig. 1. Unified detector consolidates an end-to-end instance segmentation model, MaX-DeepLab [53], to detect arbitrarily shaped text and multi-head self-attention layers [51] to form text clusters. The proposed model enables end-to-end training and inference with a single-stage simplified pipeline. It eliminates complex label generation processes [3, 44] during training and sophisticated post-processing [33, 63] during inference. Unified Detector outperforms competitive baselines and even a commercial solution on the task of unified text detection and geometric layout analysis, demonstrating its effectiveness.

Along with the unified task, we also evaluate our model on the standalone scene text detection task using existing public datasets, including ICDAR 2017 MLT [38], TotalText [10], CTW1500 [60], MSRA-TD500 [59], and achieve state-of-the-art results. While fine-tuning is a common practice in recent works [44, 63], the proposed model is directly trained using a combination of datasets without fine-tuning on each individual target dataset. The unified detector is the first end-to-end model that achieves state-of-the-art performance on the text detection task and simultaneously recovers important text layout information.

In conclusion, our core contributions are as follows:

- We propose the task of unified text detection and layout analysis, bringing together two tasks that have been studied independently, yet are intrinsically connected.
- A new high quality dataset with hierarchical text annotations is introduced to facilitate research on this task.
- We propose an end-to-end unified model, which outperforms competitive multi-stage baselines that treat the two tasks separately.
- Our model, which is free of complex post-processing, achieves state-of-the-art results on multiple challenging public text detection benchmarks.

2. Related Works

2.1. Scene text and documents datasets

There have been a variety of scene text datasets and document datasets. Scene text datasets range from straight text [19] to curved text [10, 60], sparse text to dense text [48, 50], monolingual text [10, 19] to multilingual text [37, 38], word level to line level [59, 60], narrower image domain to broader image domain [22, 48], varying in characteristics. However, these datasets only focus on the retrieval of individual words or text lines. There are also datasets that provide additional higher-level annotations for text based VQA [48] and image captioning [57]. However, they focus on specific tasks and do not analyze the layout of text, which has universal usage in downstream tasks. Document datasets [1, 12, 13, 27, 62] only provide annotations for layout analysis without labeling the atomic entities i.e. words. Furthermore, these datasets only contain scanned or digital documents for a specific domain such as academic papers [62] and historical newspapers [13]. The text reading order dataset [27] only contains images that have well-defined reading order, such as product labels and instruction manuals, and thus is not general. The proposed dataset is the first dataset that allows joint detection and layout analysis for general natural images.

2.2. Scene text detection

Recently, scene text detection research [32] has mainly focused on the representation of irregularly shaped text and post-processing method that recovers the text contours from geometric attributes such as word or character center regions, pixel level orientation, and radius of the text [3, 14, 28, 33, 44, 55]. The custom representation for text complicates the label generation process and post-processing, such as semi-supervised and iterative generation of character center regions [3], boundary shrinking and recovery [28] with Vatti clipping [52], and polygonal non-maximum suppression [14]. Raisi et al. [41] introduce the end-to-end detector DETR [6] to detect text using rotated bounding boxes, but it does not handle curved text. Besides, these works only provide solutions to the task of text detection. Conversely, our research works on the unification of text detection and layout analysis with an end-to-end neural network that greatly simplifies the whole pipeline.

2.3. Layout analysis

Driven by the success of object detection [17, 42] and semantic segmentation [8, 31] in images, layout analysis in documents is also framed as detection and segmentation tasks in some works [26, 43, 62], where detector models are trained to detect semantically coherent text blocks as objects. These methods fail to produce word or line level detections and can only be used in company with standalone

Dataset	#Img			#Word			Ann Level
	Train	Val	Test	(avg/total)	Word	Line	
IC15 [19]	1,000	0	500	4.4/6.5K	✓		
Total-Text [10]	1,255	0	300	7.4/11K	✓		
CTW1500 [60]	1,000	0	500	6.7/10K		✓	
MSRA-TD500 [59]	300	0	200	6.9/3.5K		✓	
IC17 MLT* [38]	7,200	1,800	9,000	9.5/85K	✓		
IC19 MLT* [37]	10,000	0	10,000	8.9/89K	✓		
IC19 LSVT* [49]	30,000	0	20,000	8.1/243K		✓	
IC19 ArT* [11]	5,603	0	4,563	8.9/50K	✓		
TextOCR [48]	21,778	3,124	3,232	32.1/903K	✓		
Intel OCR [22]	191,059	16,731	0	10.0/2.1M	✓		
HierText	8,281	1,724	1,634	103.8/1.2M	✓	✓	✓

Table 1. HierText v.s. other datasets. Our dataset is characterized by high text density and hierarchical annotations. Datasets marked with * do not provide test set annotations. The train and validation sets were used to count the number of words.

text detectors, increasing the complexity of the pipeline. Another branch of work [54] takes a hierarchical view and apply graph-based models on the finest granularity, i.e. individual words, to analyze the layout. All of these prior arts have mainly focused on document datasets. Unlike these works, we introduce layout analysis into scene text domain and propose an end-to-end unified model.

3. Hierarchical Text Dataset (HierText)

3.1. Data collection

Images in HierText are collected from the Open Images v6 dataset [24]. We scan Open Images using a public commercial OCR engine, *Google Cloud Platform Text Detection API (GCP)*¹, to search for images with text. We filter out images: a) with few detected words, b) low recognition confidence and c) with non-English dominant text. Finally, we randomly sample a subset from the remaining images to construct our dataset. 11639 images are obtained and further splitted into *train*, *validation*, and *test* set. HierText images are of higher resolution with their long side constrained to 1600 pixels compared to previous datasets based on Open Images [22, 48] that are constrained to 1024 pixels, resulting in more legible text.

We annotate these images in a hierarchical way [16]. We first annotate word locations with polygons. Legible words are also transcribed regardless of their language. The top-left corner and the orientation of the polygon define the word’s reading direction. Then words are grouped to text lines. Paragraphs are firstly annotated using polygon and then text lines and words are associated with the corresponding polygon based on their binary mask intersection. As a result, we obtain a tree structure annotation hierarchy. Note that, the clustering of words into lines and lines into paragraphs are relatively low-cost, since precise pixel level annotation is not required.

Coverage check: We check the cross-dataset coverage between HierText and the other two text datasets from Open

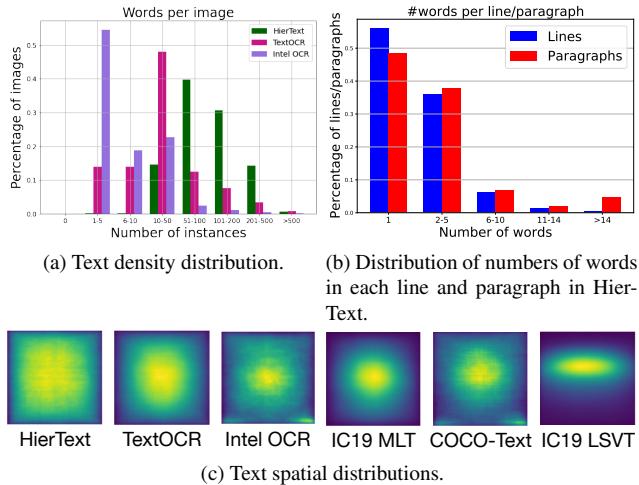


Figure 2. **Statistics of HierText dataset.** Compared to other datasets, our dataset has higher word density per image and more uniformly distributed text.

Images, i.e. TextOCR [48] and Intel OCR [22]. Only 1.5% of our images are in TextOCR, and 3.6% in Intel OCR. We also ensure that our training images are not in the validation or test set images of TextOCR and Intel OCR, or vice versa.

3.2. Dataset characteristics

Table 1 compares statistics between HierText and other datasets. HierText has 103.8 words per image on average; approximately 3 times the text density of the second densest dataset, i.e. TextOCR [48]. Even though HierText has fewer images than TextOCR, it contains more legible words. Finally, HierText is the only dataset that provides hierarchical annotations. Fig. 2a shows that HierText represents a different domain of images from existing public datasets. It has a large proportion of high text density images. While Intel OCR [22] has the largest number of images and some coverage of images with more than 100 words, HierText contains more of them in terms of absolute number: 5.3K v.s. 3.4K. Fig. 2c illustrates that the spatial distribution of text is also more uniform in HierText. In other datasets, text tend to be located in the center of the images. The distribution of the number of words in each line and paragraph is shown in Fig. 2b. A significant proportion of lines and paragraphs have more than one word making the layout analysis a challenging problem.

Overall, we demonstrate that the proposed HierText dataset has unique characteristics and captures an uncovered domain from other datasets. Additionally, it enables research into unified text detection and layout analysis.

3.3. Task and evaluation protocol

Tasks: There are two task categories for HierText dataset. The first category involves the instance segmentation of text

¹<https://cloud.google.com/vision/docs/ocr>

at word or line levels. Conceptually, word level and line level outputs are interchangeable since modern text recognition systems [9, 35, 45] are highly effective with both type of input image patches. For the second task of layout analysis, we also frame it as an instance segmentation task by treating each text cluster, i.e. “paragraph”, as one object instance, following previous works [62]. The ground-truths for text lines and paragraphs are defined as the union of pixel-level masks of the underlying word level polygons.

A candidate method for the unified detection and layout analysis task should produce text entity detection results at either word or line level, and also the grouping of these entities into paragraphs.

Evaluation: To evaluate these tasks as instance segmentation, we use the recently proposed Panoptic Quality (PQ) metric [21] as the main evaluation metric:

$$PQ = \frac{\sum_{(p,g) \in TP} IoU(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \quad (1)$$

where TP , FP , FN represent true positives, false positives, and false negatives respectively. Mathematically, PQ equals to the product of *ICDAR15* [19] style *F1 score* [15] and *average IoU of all TP pairs*. The motivation for a segmentation metric is that text entities are sensitive to missing or superfluous pixels which result in missing or unexpected characters in recognition. Although there have been recent works [25, 30, 46] investigating the scoring of text detection, they do not generalize to complex geometric entities like text lines and paragraphs. PQ metric treats word, line and paragraph segmentation tasks in a uniform way. Therefore, we adopt PQ metric for the evaluation of all tasks for its simplicity and ubiquity.

4. Methodology

4.1. Unified detector

We propose an end-to-end model to perform unified scene text detection and layout analysis. We term it *Unified Detector*. It is designed to produce (1) a set of text detection masks and (2) the clustering of these detections simultaneously without complex post-processing.

End-to-end text detection: Inspired by recent advances in end-to-end object detection and panoptic segmentation [6, 53], we represent the detection of text as producing a fixed number of N softly exclusive masks $\{\hat{m}_i\}_{i=1}^N$ and N binary classifications $\{\hat{y}_i\}_{i=1}^N$. The masks satisfy $\sum_{i=1}^N \hat{m}_i = \mathbb{1}^{H \times W}$. The binary classification \hat{y}_i denotes the probability of the i -th mask being a text object. This representation is suitable for text of arbitrarily shape and can accurately capture both word and line level detections.

Unified layout analysis: Unified detector analyzes the layout and performs text clustering by producing an affinity matrix: $\hat{A} \in [0, 1]^{N \times N}$. Entry $\hat{A}_{i,j}$ in this matrix represents

the probability of text represented by \hat{m}_i and \hat{m}_j belonging to the same semantic/paragraph group.

Inference: The inference of unified detector is straightforward. We first obtain text detection results by applying argmax on the masks to assign each pixel to one text object. Then, we remove low-confidence pixels. As a result, for the i -th object, the final mask is represented as:

$$m'_{i,x,y} = \mathbb{1}(i = \operatorname{argmax}_j [\hat{m}_{j,x,y}] \text{ and } \hat{m}_{i,x,y} > t_m) \quad (2)$$

where t_m is the threshold for pixel’s confidence. We further filter text instances by applying a threshold t_c on the binary classification score \hat{y}_i . For layout analysis inference, we cluster a pair of text instances if their affinity score $\hat{A}_{i,j}$ is above a threshold, denoted as t_A . A union-find algorithm is utilized to merge these connected nodes into clusters.

4.2. Model architecture

The architecture of the proposed unified detector is illustrated in Fig. 3. Our unified detector is based on the recent Max-DeepLab [53] end-to-end panoptic segmentation framework. In this framework, we augment the input pixels with a set of N learned object queries that are D -dimensional. Then we feed the pixels and object queries into a transformer-based encoder, the MaX-DeepLab backbone, in which the bidirectional communication between pixels and object queries allows the model to encode text instances in each of the object queries. With the encoded queries and pixel features, the *text detection branch* produces the text mask output, $\{\hat{m}_i\}_{i=1}^N$. The *layout branch* produces the affinity matrix $\hat{A} \in [0, 1]^{N \times N}$ for the relations between each pair of text instances. A third branch produces the binary classification scores $\{\hat{y}_i\}_{i=1}^N$.

Backbone: The MaX-DeepLab [53] backbone is composed of an alternating stack of hourglass [39] style CNNs and the proposed dual-path transformer. The Hourglass style [39] CNNs are applied to pixel features. They encode features from coarse to fine resolutions iteratively and thus can produce high resolution features. The dual-path transformer [53] allows bidirectional communication between pixel features and the learnable object queries. It enables attention within pixel space and interaction among object queries. This makes it possible to encode long-range information in pixel features, and allows object queries to locate and retrieve text objects exclusively from pixels. The MaX-DeepLab produces output at $\frac{1}{4}$ resolution of the input, i.e. $(H', W') = (\frac{H}{4}, \frac{W}{4})$. We urge readers to refer to the original paper [53] for full details.

Text detection branch: The text detection branch takes the outputs of the MaX-DeepLab backbone and produces the text mask outputs. Two fully-connected layers produce mask queries from the encoded queries, denoted as $f \in \mathbb{R}^{N \times D}$. Similarly, two convolutional layers produce

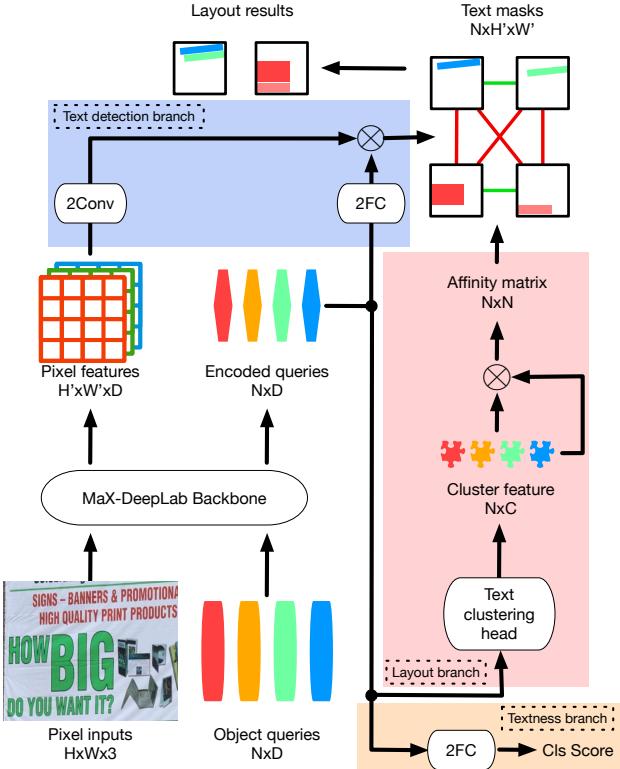


Figure 3. Illustration of our approach. Our method produces mask outputs for the text detection task and an affinity matrix for layout analysis task in a unified way. The text detection branch produces an $N \times H' \times W'$ tensor that represents the N softly exclusive masks. The layout analysis branch produces an $N \times N$ affinity matrix that models the pairwise relationship of the predicted masks. The green links in the top right suggest clustering of the text instances, while the red links indicate the opposite. The binary classification score produced by textness branch is used to filter out non-text objects from object queries.

normalized pixel features, denoted as $g \in \mathbb{R}^{D \times H' \times W'}$. The text mask prediction is the inner product of f and g :

$$\hat{m} = \text{softmax}_N(f \cdot g) \in \mathbb{R}^{N \times H' \times W'} \quad (3)$$

Layout branch: Layout branch takes the encoded queries from the backbone as the sole input. In order to separate layout features from text detection features, we apply an extra projection head for cluster embedding projection. For this projection head, we adopt a 3-layered multi-head self-attention layer [51] to obtain the normalized layout features, denoted as $h \in \mathbb{R}^{N \times C}$. We apply inner product of the layout features followed by a sigmoid function with temperature τ to get the affinity matrix:

$$\hat{A}[i, j] = \frac{1}{1 + e^{-\frac{h_{i,:} h_{j,:}^T}{\tau}}} \quad (4)$$

Textness branch: The textness branch applies another 2-layered fully connected layers and a sigmoid function to produce the binary classification scores $\{\hat{y}_i\}_{i=1}^N$.

4.3. Training targets

Unified detector enables end-to-end training for both the scene text detection task and the layout analysis task. The key ingredient is to perform a bipartite matching between prediction and groundtruth since our model produces an unordered set of outputs. We first describe the matching between prediction and groundtruth of the detection task and the metric we use. Then we show the joint optimization of our unified detector for both tasks.

Text matching: We adopt the PQ-style similarity score proposed in MaX-DeepLab [53]. For a pair of prediction (\hat{m}_i, \hat{y}_i) and groundtruth (m_j, y_j) , the score is defined as:

$$\text{sim}(i, j) = [\hat{y}_i y_j + (1 - \hat{y}_i)(1 - y_j)] \times \text{Dice}(\hat{m}_i, m_j) \quad (5)$$

where $\text{Dice}(\hat{m}_i, m_j)$ denotes the Dice coefficient [36] between the pair of masks. It measures mask similarity. This score considers both the classification score and mask score.

The goal of matching is to find a permutation of N elements $\sigma \in \mathfrak{S}_N$ to maximize the total similarity between predictions and groundtruths:

$$\hat{\sigma} = \underset{\sigma}{\operatorname{argmax}} \sum_{i=1}^N \text{sim}(i, \sigma(i)) \quad (6)$$

Following previous works [6, 53], we solve this optimal assignment problem with the Hungarian algorithm [23] on the fly during training.

Text detection loss: The training target for text detection is adopted from MaX-DeepLab [53]:

$$\begin{aligned} \mathcal{L}_{det} = & \frac{1}{N} \sum_{i=1}^N \{(1 - \alpha)(1 - y_{\sigma(i)})[-\log(1 - \hat{y}_i)] \\ & + \alpha y_{\sigma(i)}[-\ddot{\hat{y}}_i \text{Dice}(\hat{m}_i, m_{\sigma(i)}) - \text{Dice}(\hat{m}_i, m_{\sigma(i)}) \log(\hat{y}_i)]\} \end{aligned} \quad (7)$$

where dotted variables $\ddot{\hat{y}}_i$ and $\text{Dice}(\hat{m}_i, m_{\sigma(i)})$ denote constant weights and gradients do not pass through them. α is a balancing factor between positive and negative samples.

Layout analysis loss: We first define the ground-truths for output of layout analysis branch. Each text instance comes with a text cluster ID, denoted as $\{c_i\}_{i=1}^N$. This is part of the annotations of the proposed HierText dataset. The groundtruth affinity matrix can be intuitively defined as:

$$A[i, j] = \mathbb{1}(c_i == c_j) \quad (8)$$

Then, the layout analysis loss can be computed as:

$$\begin{aligned} \mathcal{L}_{lay} = & \sum_{i=1}^N \sum_{j=1}^N y_{\sigma(i)} y_{\sigma(j)} \{ \alpha_L w_p A_{\sigma(i), \sigma(j)} [-\log(\hat{A}_{i,j})] \\ & + (1 - \alpha_L) w_n (1 - A_{\sigma(i), \sigma(j)}) [-\log(1 - \hat{A}_{i,j})] \} \quad (9) \end{aligned}$$

where $w_p = [\sum_{i=1}^N \sum_{j=1}^N y_{\sigma(i)} y_{\sigma(j)} A_{\sigma(i), \sigma(j)}]^{-1}$, $w_n = [\sum_{i=1}^N \sum_{j=1}^N y_{\sigma(i)} y_{\sigma(j)} (1 - A_{\sigma(i), \sigma(j)})]^{-1}$, and α_L is a balancing factor.

The final training target is the weighted sum of the text detection loss \mathcal{L}_{det} , the layout analysis loss \mathcal{L}_{lay} . We also find it useful to incorporate the semantic segmentation loss \mathcal{L}_{seg} and instance discrimination loss \mathcal{L}_{ins} as defined in MaX-DeepLab [53]. As a result, the model is jointly optimized for the following loss function:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{det} + \lambda_2 \mathcal{L}_{lay} + \lambda_3 \mathcal{L}_{seg} + \lambda_4 \mathcal{L}_{ins} \quad (10)$$

where $\lambda_{1, \dots, 4}$ are weighting factors.

5. Experiments

In this section, we set up experiments to evaluate our proposed unified detector in a comprehensive way. First, we compare our method with competitive baselines. We show that the unified detector achieves better performance. We also perform thorough ablation studies to analyze the design selections of the proposed approach. Finally, we train and evaluate the unified detector on public datasets for the sole task of scene text detection, verifying the effectiveness of the text detection branch.

5.1. Baselines

The task of unified detection and layout analysis largely remains untouched in the academic literature, despite the incredible progress of scene text detection methods and increasing number of layout analysis algorithms. We therefore carefully select the following baselines representing non-end-to-end methods:

Commercial solution: The *GCP API*, as mentioned above, is a commercial solution that produces text detection and recognition results at word, line and paragraph level.

GCN Post-Processing: The GCN [20] based post-processing method (GCN-PP) [54] applies the GCN on text line bounding boxes to cluster lines into paragraphs.

Object detection baselines: PubLayNet [62] formulates the layout analysis as an instance segmentation task predicting text clusters as pixel masks. Following this work, we build a baseline using Mask R-CNN [17] as in [62] that produces text cluster masks. Each such mask represents one text cluster. The layout analysis is performed by assigning each detected text entity (word or line) to the text cluster



Figure 4. Outputs of the unified detector trained on HierText. Images are sampled from the val and test set of HierText, Total-Text, CTW1500, IC15, ICDAR17 MLT, and MSRA-TD500. In each pair of images bounded by dotted boxes, the image on the **top or left** visualizes **text detection results**. The image on the **bottom or right** visualizes results of **layout analysis**. These results demonstrate that our unified detector is able to detect arbitrarily shaped text and produce text clusters regardless of the variability in shapes, fonts, colors and backgrounds.

Method	Text detection branch	Layout analysis branch	Text line	Layout
			Detection PQ	Analysis PQ
GCP API	unknown	unknown	56.17	46.33
GCN-PP		GCN		50.10
Mask-RCNN-Cluster	Text detection branch of	Mask R-CNN [17]		51.67
MaX-DeepLab-Cluster	unified detector	MaX-DeepLab [53]	62.23	52.52
Unified Detector		Layout branch of unified detector		53.60

Table 2. Results of text detection and layout analysis on HierText test set. The last row represents our end-to-end unified scene text detection and layout analysis model.

with the maximum area of intersection. Since this model does not produce word or line level detections, it is used in combination with a text entity detection model as specified in Sec. 5.2. This two-stage baseline is dubbed *Mask-RCNN-Cluster*. Similarly, we build *MaX-DeepLab-Cluster* using MaX-DeepLab [53], which represents a more competitive method with state-of-the-art advance in instance object segmentation task.

5.2. Experimental settings

Unified Detector: We use the DeepLab2 [56] library for the implementation of the MaX-DeepLab part of our method.

We use the MaX-DeepLab-S backbone, with an input size of 1024×1024 . The number of object query is set to 384, due to the high density of text in our dataset. Query dimensions are $D = 256$ and $C = 128$ respectively. In our main experiments, we only use HierText as training data. The models are trained on 128 TPUv3 cores with a batch size of 256 for 100K steps, AdamW [34] optimizer with weight decay rate of 0.05, and cosine learning rate starting from 10^{-3} . The weights for PQ-loss, layout analysis loss, instance discrimination loss, and semantic segmentation loss are 3.0, 1.0, 1.0, 1.0 respectively. The balancing factors are set as $\alpha = 0.5$ and $\alpha_L = 0.5$. During inference, we filter out text masks that have less than 32 pixels or less than $t_c = 0.5$ in confidence. We also use $t_m = 0.4$ to filter out low confidence pixels. For text clustering, we use a threshold of $t_A = 0.5$ on the affinity matrix. In our main experiments, the text detection branch of unified detector is trained to detect text lines as opposed to words. Note that most of these hyper-parameters follow the original settings of MaX-DeepLab.

Baselines: For Mask-R-CNN-Cluster, we use the implementation from the public TF-Vision repository². Input size is set to 1024×1024 . For MaX-DeepLab-Cluster, we follow the same hyper-parameter and training settings of our unified detector for fair comparison. For GCN-PP, we follow the settings in [54] to train the line clustering model. As mentioned above, these methods can only perform layout analysis based on detected text entities. Therefore, we pair these three baselines with the text detection branch of our unified detector for fair comparison.

5.3. Main Results

We evaluate our method and compare with baselines as detailed above. Results are summarized in Tab. 2. Compared with other standalone text clustering methods including GCN-based and detection-based ones, our end-to-end unified approach achieves better layout analysis performance by a considerable margin of 1.08% in PQ score. Note that, these baseline methods are applied on the outputs of the text detection branch of unified detector. Therefore, the only difference is in the layout analysis method. This shows that the built-in end-to-end text clustering module of unified detector is more effective and better than standalone baseline modules. Note that the baselines are two stage approaches that require almost double the computational resources. For text detection, our system achieves higher performance than the GCP API (62.23 v.s. 56.17).

We also demonstrate results on images from various domains, as shown in Fig. 4. The proposed method is able to work on various layouts, including text clusters with curved text and with non-uniform fonts and colors.

²<https://github.com/tensorflow/models/tree/master/official/vision/beta>

#Obj query	Method	Layout analysis (PQ)
128	Unified-Detector-Word	34.38
	Unified-Detector-Line	51.48
256	Unified-Detector-Word	36.71
	Unified-Detector-Line	52.50
384	Unified-Detector-Word	39.11
	Unified-Detector-Line	53.60

Table 3. Comparison between word-based and line-based unified detector. Results demonstrate that line-based unified detector outperforms word-based unified detector consistently with different numbers of object queries.

Balancing method	Text Line Detection				Layout analysis	
	P	R	F	T	PQ	(PQ)
Vanilla	75.34	75.02	75.18	77.27	58.10	50.04
α -balanced loss ($\alpha = 0.25$)	76.32	75.20	75.76	77.57	58.76	51.48
focal loss	75.16	74.58	74.87	77.38	57.94	45.22

Table 4. The effect of balancing factor in text clustering loss on text and layout metrics.

5.4. Ablation studies

In this section, we do ablation studies to further explore the design details. Except the detection granularity experiments (i.e. word v.s. line), we use $N = 128$ object queries.

Word-based v.s. line-based: Our unified detector framework is able to perform end-to-end text entity detection on either word or line level, and then cluster these entities into the paragraph level as layout analysis results. Though word and line detections are largely interchangeable in terms of the subsequent recognition algorithms, we observe significant difference in layout analysis as shown in Tab. 3. While both word and line level models benefit from more object queries, line level models consistently outperform their word level peers. One potential cause may be that detecting at line level reduces the number of objects compared to word-level detections, making the optimization for the clustering head easier.

Text clustering loss: We compare the use of different ways to balance the clustering loss. Results are listed in Tab. 4. α -balancing is the default method described in Sec. 4.3. Vanilla means no balancing at all; it normalizes the loss term directly by $w = [\sum_{i=1}^N \sum_{j=1}^N y_{\sigma(i)} y_{\sigma(j)}]^{-1}$. Applying α -balancing factor achieves considerable improvements in both text detection and layout analysis. Balancing the loss with focal style factors [29] results in worse performance in both tasks.

Text clustering head: We compare our default setting, a 3-layered multi-head self-attention (MHSA) [51] head, with other viable choices, as shown in Tab. 6. We also list the performance of a MaX-DeepLab line detector without layout analysis branch. If we do not use any extra layer, the text detection performance is deteriorated compared to

Method	Venue	Training Data		Word Detection						Line Detection					
		Pub	HierText	ICDAR 17 MLT			Total-Text			CTW1500			MSRA-TD500		
				P	R	F	P	R	F	P	R	F	P	R	F
CRAFT [3]	CVPR19	✓		80.6	68.2	73.9	87.6	79.9	83.6	86.0	81.1	83.5	88.2	78.2	82.9
PSENet [55]	CVPR19	✓		75.3	69.2	72.2	84.0	78.0	80.9	84.8	79.7	82.2	-	-	-
FCE [64]	CVPR21	✓		-	-	-	89.3	82.5	85.8	87.6	83.4	85.5	-	-	-
MOST [18]	CVPR21	✓		82.0	72.0	76.7	-	-	-	-	-	-	90.4	82.7	86.4
CentripetalText [44]	NeurIPS21	✓		-	-	-	90.67	85.19	87.85	87.66	80.57	83.97	86.62	84.54	85.57
ABPNet [61]	ICCV21	✓		-	-	-	90.6	82.5	86.3	88.3	79.9	83.9	90.0	82.5	86.1
PCR [14]	CVPR21	✓					88.5	82.0	85.2	87.2	82.3	84.7	90.8	83.5	87.0
Ours (word)	-	✓		77.71	75.88	76.78	85.49	90.53	87.94	-	-	-	-	-	-
Ours (line)	-	✓	✓	78.05	76.44	77.24	84.96	91.06	87.90	-	-	-	-	-	-

Table 5. Results of word and text line detection on public scene text datasets. Both our word and line detectors are outperforming the latest methods, even though our models are not fine-tuned for any target datasets. The proposed new dataset also proves to be a helpful complement to existing scene text datasets.

Text clustering head	Text Line Detection					Layout (PQ)
	P	R	F	T	PQ	
Line detector only	76.21	75.11	75.66	77.38	58.55	-
no-extra layer	75.50	74.16	74.82	77.14	57.72	48.07
1 FC-ReLU-BN	75.71	74.85	75.28	77.41	58.28	47.79
MHSA x1	76.11	75.65	75.88	77.43	58.76	51.00
MHSA x3	76.32	75.20	75.76	77.57	58.76	51.48

Table 6. The impact of different text clustering head architecture on text and layout metrics.

line detector only, indicating that it is necessary to separate the features. However, using fully connected layer cannot fully recover the ability to detect text and worsens layout analysis. Using 1 layer of MHSA is better than only using fully-connected layer in both the detection and layout tasks. This is intuitive since Transformer’s [51] architecture block provides stronger modelling of interactions between text entities. Finally, additional transformer layers improve the performance.

5.5. Scene text detection on public datasets

In this section, we evaluate our models on the most widely used benchmarks for scene text detection. We adopt the same training and optimization settings in Sec. 5.2 except that the layout analysis branch is excluded since other public datasets do not have layout labels. We use $N = 384$ object queries. We do not initialize our models from any checkpoint. Nor do we pretrain on any synthetic datasets. We directly train on the union of public datasets without fine-tuning on any of them³. We directly evaluate the models with the checkpoint of last training iteration. We evaluate on the following 4 benchmarks: *MLT17* [38], *Total-Text*

³For word detection, we use TextOCR [48], MLT17 [38], Total-Text [10], HierText. For line detection, we use LSVT [49], CTW1500 [60], MSRA-TD500 [59], HierText.

[10], *CTW1500* [60], and *MSRA-TD500* [59]. Results and comparison with previous papers are summarized in Tab. 5. Overall, our detectors are characterized by higher recall and lower precision compared to state-of-the-art methods. Notably, even though curve text takes up a very small proportion in the training datasets, our model still excels at both curved text datasets, CTW1500 and Total-Text, showing case the adaptability of the proposed method.

For word detection, we achieve state-of-the-art result (77.24) on MLT 17. The performance is still very competitive (76.78) when trained on other public datasets only. On Total-text, we achieve state-of-the-art regardless of the use of HierText (87.94 and 87.90).

For line detection, we achieve very competitive results on CTW1500 and MSRA-TD500 without training on HierText. We observe considerable improvements when we add HierText in our training data (84.88 → 85.97 and 86.69 → 87.70). This demonstrates that HierText is a helpful complement to the collection of public line datasets.

6. Conclusion

In this paper, we motivate the task of unified scene text detection and layout analysis. To facilitate research into this direction, we collect a dataset with hierarchical text annotations. We further propose an end-to-end model for unified detection and layout analysis that outperforms previous methods while at the same time greatly simplifying the pipeline. With the new task, dataset, and model, we push the envelop of text extraction and understanding in images and enable better support for downstream tasks.

References

- [1] Apostolos Antonacopoulos, David Bridson, Christos Papadopoulos, and Stefan Pletschacher. A realistic dataset for

- performance evaluation of document layout analysis. In *2009 10th International Conference on Document Analysis and Recognition*, pages 296–300. IEEE, 2009. 2
- [2] Srikanth Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 993–1003, 2021. 1
- [3] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9365–9374, 2019. 1, 2, 8
- [4] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019. 1
- [5] Thomas M Breuel. Two geometric algorithms for layout analysis. In *International workshop on document analysis systems*, pages 188–199. Springer, 2002. 1
- [6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2, 4, 5
- [7] Roldano Cattoni, Tarcisio Coianiz, Stefano Messelodi, and Carla Maria Modena. Geometric layout analysis techniques for document image understanding: a review. *ITC-irst Technical Report*, 9703(09), 1998. 1
- [8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2
- [9] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 4
- [10] Chee Kheng Ch'ng and Chee Seng Chan. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 935–942. IEEE, 2017. 2, 3, 8
- [11] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1571–1576. IEEE, 2019. 2, 3
- [12] Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. Icdar2017 competition on recognition of documents with complex layouts-rdcl2017. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1404–1410. IEEE, 2017. 2
- [13] Christian Clausner, Christos Papadopoulos, Stefan Pletschacher, and Apostolos Antonacopoulos. The enp image and ground truth dataset of historical newspapers. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 931–935. IEEE, 2015. 2
- [14] Pengwen Dai, Sanyi Zhang, Hua Zhang, and Xiaochun Cao. Progressive contour regression for arbitrary-shape scene text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7393–7402, June 2021. 1, 2, 8
- [15] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 4
- [16] Robert M Haralick. Document image understanding: Geometric and logical layout. In *CVPR*, volume 94, pages 385–390, 1994. 3
- [17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2, 6
- [18] Minghang He, Minghui Liao, Zhibo Yang, Humen Zhong, Jun Tang, Wenqing Cheng, Cong Yao, Yongpan Wang, and Xiang Bai. Most: A multi-oriented scene text detector with localization refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8813–8822, 2021. 1, 8
- [19] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Angelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160. IEEE, 2015. 2, 3, 4
- [20] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 6
- [21] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019. 4
- [22] Ilya Krylov, Sergei Nosov, and Vladislav Sovrasov. Open images v5 text annotation and yet another mask text spotter. In *Asian Conference on Machine Learning*, pages 379–389. PMLR, 2021. 2, 3
- [23] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5
- [24] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasic, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallochi, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 3
- [25] Chae Young Lee, Youngmin Baek, and Hwalsuk Lee. Tederal: A fair evaluation metric for scene text detectors. In *2019*

- International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 7, pages 14–17. IEEE, 2019. 4
- [26] Joonho Lee, Hideaki Hayashi, Wataru Ohyama, and Seiichi Uchida. Page segmentation using a convolutional neural network with trainable co-occurrence features. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1023–1028. IEEE, 2019. 1, 2
- [27] Liangcheng Li, Feiyu Gao, Jiajun Bu, Yongpan Wang, Zhi Yu, and Qi Zheng. An end-to-end ocr text re-organization sequence learning for rich-text detail image comprehension. In *European Conference on Computer Vision*, pages 85–100. Springer, 2020. 2
- [28] Minghui Liao, Guan Pang, Jing Huang, Tal Hassner, and Xiang Bai. Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 706–722. Springer, 2020. 2
- [29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 7
- [30] Yuliang Liu, Lianwen Jin, Zecheng Xie, Canjie Luo, Shuaítiao Zhang, and Lele Xie. Tightness-aware evaluation protocol for scene text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9612–9620, 2019. 4
- [31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2
- [32] Shangbang Long, Xin He, and Cong Yao. Scene text detection and recognition: The deep learning era. *International Journal of Computer Vision*, 129(1):161–184, 2021. 2
- [33] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. Textsnake: A flexible representation for detecting text of arbitrary shapes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 20–36, 2018. 2
- [34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 7
- [35] Ning Lu, Wenwen Yu, Xianbiao Qi, Yihao Chen, Ping Gong, Rong Xiao, and Xiang Bai. Master: Multi-aspect non-local network for scene text recognition. *Pattern Recognition*, 117:107980, 2021. 4
- [36] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 5
- [37] Nibal Nayef, Yash Patel, Michal Busta, Pinaki Nath Chowdhury, Dimosthenis Karatzas, Wafa Khelif, Jiri Matas, Umapada Pal, Jean-Christophe Burie, Cheng-lin Liu, et al. Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1582–1587. IEEE, 2019. 2, 3
- [38] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1454–1459. IEEE, 2017. 2, 3, 8
- [39] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 4
- [40] Siyang Qin, Alessandro Bissacco, Michalis Raptis, Yasuhisa Fujii, and Ying Xiao. Towards unconstrained end-to-end text spotting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4704–4714, 2019. 1
- [41] Zobeir Raisi, Mohamed A Naei, Georges Younes, Steven Wardell, and John S Zelek. Transformer-based text detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3162–3171, 2021. 2
- [42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 2
- [43] Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel, and Sheraz Ahmed. Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 1162–1167. IEEE, 2017. 2
- [44] Tao Sheng, Jie Chen, and Zhouhui Lian. Centripetaltext: An efficient text instance representation for scene text detection. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 8
- [45] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2035–2048, 2018. 4
- [46] Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge Belongie, Shijian Lu, and Xiang Bai. Icdar2017 competition on reading chinese text in the wild (rctw-17). In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1429–1434. IEEE, 2017. 4
- [47] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 1
- [48] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF Conference on Computer*

- Vision and Pattern Recognition (CVPR)*, pages 8802–8812, June 2021. 2, 3, 8
- [49] Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1557–1562. IEEE, 2019. 2, 3, 8
- [50] Jun Tang, Zhibo Yang, Yongpan Wang, Qi Zheng, Yongchao Xu, and Xiang Bai. Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping. *Pattern recognition*, 96:106954, 2019. 2
- [51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2, 5, 7, 8
- [52] Bala R Vatti. A generic solution to polygon clipping. *Communications of the ACM*, 35(7):56–63, 1992. 2
- [53] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5463–5474, 2021. 2, 4, 5, 6
- [54] Renshen Wang, Yasuhisa Fujii, and Ashok C Popat. Post-ocr paragraph recognition by graph convolutional networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 493–502, 2022. 1, 3, 6, 7
- [55] Wenhui Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. Shape robust text detection with progressive scale expansion network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9336–9345, 2019. 2, 8
- [56] Mark Weber, Huiyu Wang, Siyuan Qiao, Jun Xie, Maxwell D Collins, Yukun Zhu, Liangzhe Yuan, Dahun Kim, Qihang Yu, Daniel Cremers, et al. Deeplab2: A tensorflow library for deep labeling. *arXiv preprint arXiv:2106.09748*, 2021. 6
- [57] Guanghui Xu, Shuaicheng Niu, Mingkui Tan, Yucheng Luo, Qing Du, and Qi Wu. Towards accurate text-based image captioning with content diversity exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12637–12646, 2021. 1, 2
- [58] Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C Lee Giles. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5315–5324, 2017. 1
- [59] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Detecting texts of arbitrary orientations in natural images. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1083–1090. IEEE, 2012. 2, 3, 8
- [60] Liu Yuliang, Jin Lianwen, Zhang Shuaitao, and Zhang Sheng. Detecting curve text in the wild: New dataset and new solution. *arXiv preprint arXiv:1712.02170*, 2017. 2, 3, 8
- [61] Shi-Xue Zhang, Xiaobin Zhu, Chun Yang, Hongfa Wang, and Xu-Cheng Yin. Adaptive boundary proposal network for arbitrary shape text detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1305–1314, October 2021. 1, 8
- [62] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE, 2019. 1, 2, 4, 6
- [63] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560, 2017. 2
- [64] Yiqin Zhu, Jianyong Chen, Lingyu Liang, Zhanghui Kuang, Lianwen Jin, and Wayne Zhang. Fourier contour embedding for arbitrary-shaped text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3123–3131, June 2021. 8