Towards Unified Multi-granularity Text Detection with Interactive Attention

Xingyu Wan [1] Chengquan Zhang † [1] Pengyuan Lyu [1] Sen Fan [1] Zihan Ni [1] Kun Yao [1] Errui Ding [1]
Jingdong Wang [1]

Abstract

Existing OCR engines or document image analysis systems typically rely on training separate models for text detection in varying scenarios and granularities, leading to significant computational complexity and resource demands. In this paper, we introduce "Detect Any Text" (DAT), an advanced paradigm that seamlessly unifies scene text detection, layout analysis, and document page detection into a cohesive, end-to-end model. This design enables DAT to efficiently manage text instances at different granularities, including *word, line, paragraph* and *page*. A pivotal innovation in DAT is the across-granularity interactive attention module, which significantly enhances the representation learning of text instances at varying granularities by correlating structural information across different text queries. As a result, it enables the model to achieve mutually beneficial detection performances across multiple text granularities. Additionally, a prompt-based segmentation module refines detection outcomes for texts of arbitrary curvature and complex layouts, thereby improving DAT's accuracy and expanding its real-world applicability. Experimental results demonstrate that DAT achieves state-of-the-art performances across a variety of text-related benchmarks, including multi-oriented/arbitrarily-shaped scene text detection, document layout analysis and page detection tasks.

1. Introduction

Text detection serves as the cornerstone for parsing and understanding the content of texts in natural scenes and electronic documents. Existing document image analysis systems typically categorize text-related detection tasks into separate modules, such as scene text detection, document layout analysis, and document page detection. Within this context, scene text detection focuses on localizing individual text instances, which may be multi-oriented (Karatzas et al., 2015; Yao et al., 2012) or arbitrarily-shaped (Ch'ng et al., 2020; Liu et al., 2019), and primarily involves detecting elements at the word-level or text line-level. Document layout analysis delves into examining geometric structures at the paragraph-level. It involves classifying fine-grained categories within these structures, but does not extend to analyzing their sub-level elements. Document page detection addresses the identification of the most salient page body in natural scenes, typically utilizing image segmentation techniques (Chen et al., 2018; Kirillov et al., 2023).

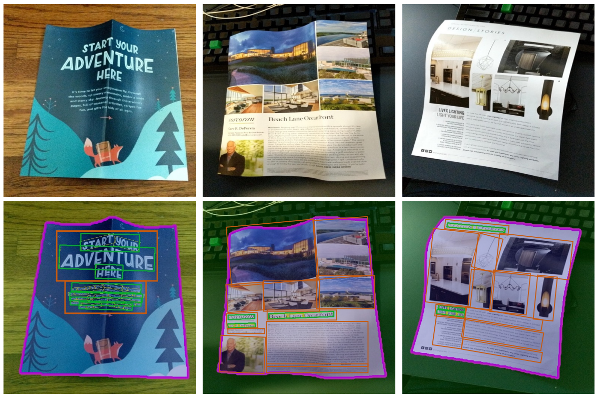

Figure 1. Illustration of the structural correlations among multi-granularity text instances, i.e., word(annotated with yellow polygons), text-line(annotated with green polygons), paragraph(annotated with brown polygons) and page(annotated with magenta contours). The blurred small text instances are ignored.

To achieve state-of-the-art (SOTA) results in the aforementioned tasks, current methods necessitate training separate models for each task using diverse datasets, which leads to considerable computational complexity and resource demands. Moreover, while these tasks involve representation learning at varying text granularities, there is often a lack of attention to the intrinsic correlations among these multi-granularity text instances, as illustrated in Figure 1.

[1] Baidu, Beijing, China. Correspondence to: Chengquan Zhang <zhangchengquan@baidu.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

HierText (Long et al., 2022) is the first to propose the unified framework for scene text detection and layout analysis, claiming that such a combination can benefit both tasks. However, it has two limitations: (1) It does not fully explore the intrinsic correlations of multi-granularity texts during representation learning. The framework primarily relies on word- or line-based methods for paragraph construction through online clustering, this bottom-up unidirectional approach neglects the potential influence of paragraph-level representations on sub-level elements. (2) The training methodology suffers from limited generalizability due to its reliance on a cascading bottom-up design. This design necessitates ground-truth labeling at all text granularities for each training sample, which restricts its applicability to other prevalent datasets.

To overcome these limitations, we introduce DAT, a unified multi-granularity text detection paradigm for detecting text instances at multiple granularities. Unlike the bottom-up, cascaded framework of HierText, DAT incorporates an interactive attention module within its Transformer decoder. This module facilitates the transmission of learned query embeddings across adjacent granularities during representation learning. In order to enable parallel training using datasets with incomplete-granularity annotations, we design a multi-granularity detection framework with a mixed-granularity training strategy. Additionally, to facilitate arbitrarily-shaped text localization and accurate document page segmentation, we follow SAM (Kirillov et al., 2023) and introduce a prompt-based mask decoder to perform foreground-background segmentation of the multi-granularity text instances.

A key feature of DAT is across-granularity representation learning within the proposed interactive attention module. This module effectively correlates the structural information among text queries of different granularities, enriching the understanding and integration of textual instance representations from both bottom-up and top-down perspectives. This attention mechanism not only elevates the accuracy in text detection but also allows for a more nuanced analysis of texts, regardless of their complexity or format. This innovative use of interactive attention significantly enhances the versatility and effectiveness of DAT, making it suitable for a wide range of text detection and understanding scenarios.

Our contributions can be summarized as follows: (1) We propose an innovative interactive across-granularity attention module tailored for the representation learning of text instances across varying granularities. (2) We design a multi-granularity text detection framework with a mixed-granularity training strategy, which addresses the limitation of previous methods that required full annotations at all text levels. The resulted model substantially improves detection performances across all text granularities, and outperforms other SOTA single-task models in text detection benchmarks across multiple granularities. (3) We introduce a prompt-based mask decoder to perform fine-grained text segmentation, which significantly improves the detection performances of arbitrarily-shaped texts, complex layouts and page bodies.

2. Related Works

2.1. Text Detection

Scene Text Detection. Scene text detection has evolved considerably, primarily divided into two categories: regression-based (or point-based) and segmentation-based approaches. Regression-based methods (He et al., 2021; Liu et al., 2020; Wang et al., 2019b; Zhu et al., 2021b; Ye et al., 2023) directly regress bounding boxes or polygon points around the text regions, demonstrating their efficiency in detecting texts of varying complexity. To enhance the detection accuracy of texts with arbitrary curvature in complex scenes, the number of regressed points are usually augmented in this line of works. Segmentation-based methods (Liao et al., 2020; 2022; Tian et al., 2019; Wang et al., 2019a; Xie et al., 2019; Qin et al., 2023; Wang et al., 2020a;b) frame text detection as a segmentation problem at different levels, e.g., pixel level (Liao et al., 2020; 2022; Tian et al., 2019; Wang et al., 2019a; Xie et al., 2019; Qin et al., 2023), segment level (Baek et al., 2019; Tang et al., 2019), and contour level (Wang et al., 2020a;b), which usually involve grouping algorithms as post-processing stages. This line of works excels at delineating arbitrarily-shaped text by analyzing fine details of contours. Datasets for scene text detection tasks are typically annotated at the word or text line level granularity.

Document Layout Analysis. Recent advancements in document layout analysis are marked by the development of comprehensive datasets. Notable examples include PubLayNet (Zhong et al., 2019), DocBank (Li et al., 2020), and DocLayNet (Pfitzmann et al., 2022), which offer diverse annotations covering a range of documents from magazines to technical papers. M6Doc (Cheng et al., 2023) is the first dataset to include Chinese examples and blend both real-world and born-digital files, presenting the most fine-grained categories for layout analysis. Therefore, we use M6Doc to validate the effectiveness of our model in document layout analysis. The annotation granularity of these datasets is at the paragraph level.

Additionally, HierText (Long et al., 2022) first proposed a unified framework for scene text detection and layout analysis, claiming that such a combination can simultaneously benefit both tasks. However, due to insufficient representation learning strategy and cascading bottom-up design, its applicability remains confined to specific dataset and scenarios.

Document Page Detection. The objective of document page detection (or page frame detection) is to accurately capture the clean and actual contour regions of text page in natural scenes or scanned documents. Traditional approaches (Shafait et al., 2007; Stamatopoulos et al., 2010; Shafait et al., 2008; Reza et al., 2019) primarily utilize detection-based strategies, which involve identifying text regions such as text lines and subsequently employing post-processing techniques to amalgamate these regions into unified page areas. Modern document image dewarping methods (Das et al., 2019; Xie et al., 2021; Ma et al., 2022; Xue et al., 2022) based on deep-learning typically employ image segmentation techniques (Chen et al., 2018; Kirillov et al., 2023) to extract the accurate edges of document pages and exclude background information, followed by subsequent rectification processes. This page segmentation is performed on commonly used document dewarping datasets, such as DIW (Ma et al., 2022) and Doc3D (Das et al., 2019), which are annotated at the page level.

2.2. Transformer-based Object Detection

Recent advancements in text detection have been significantly influenced by the evolution of Transformer-based object detection algorithms. A pivotal development in this field was marked by DETR (Carion et al., 2020), which introduced a novel one-to-one label assignment strategy and eliminated the need for manually designed components like non-maximum suppression (NMS). Subsequent methods have delved deeper into the evolution of decoder queries within DETR (Carion et al., 2020). Deformable DETR (Zhu et al., 2021a) proprosed a deformable attention module that focuses on specific sampling points around a reference point. DN-DETR (Li et al., 2022) introduced a denoising training method by bringing noisy annotations and boxes to the decoder. DINO (Zhang et al., 2022) advanced this field further by introducing mixed query selection and contrastive denoising modules. Most recently, Group-DETR (Chen et al., 2023) proposed to learn group-wise object queries for one-to-many label assignment, enhancing both detection accuracy and training efficiency.

Different from Group-DETR (Chen et al., 2023), our proposed DAT model adopts multiple groups of object queries to enable parallel training for multi-granularity text detection, and to facilitate the correlation of intrinsic structural information across different text granularities. In the DAT decoder, each group of object queries is distinctly defined by text instances at varying granularities, including *word, line, paragraph* and *page*.

3. Method

3.1. Preliminaries

We frame multi-granularity text detection task as two hierarchical branches, i.e., text detection (DET) and text segmentation (SEG), as demonstrated in Figure 2. Given an input image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$, the output of DET branch is defined as $\mathbf{Y}^{DET} = \{(b_j, c_j)\}_{j=1}^{n}$, where $b_j \in \mathbb{R}^4$ denotes the polygon points coordinates of j-th located text instance, and $c_j \in \mathcal{C}^{DET}$ denotes the corresponding assigned class label. Here, the category vocabulary $\mathcal{C}^{DET} = \mathcal{C}^{word} + \mathcal{C}^{line} + \mathcal{C}^{para} + \mathcal{C}^{page}$ is composed of four granularity levels. Specifically, $\mathcal{C}^{para}$ denotes a multi-class vocabulary, while $\mathcal{C}^{word}$, $\mathcal{C}^{line}$, and $\mathcal{C}^{page}$ represent binary classifications. The output of SEG branch is defined as $\mathbf{Y}^{SEG} = \{m_j\}_{j=1}^{n}$, where $m_j \in \mathbb{R}^{1\times H\times W}$ denotes the predicted mask of j-th detected text instance. The forward propagation for DET ans SEG branches are formulated as follows:

$$\mathbf{Y}^{DET} = f_{\mathcal{H}}(f_{dec}(f_{enc}(\mathbf{F})|\mathbf{A}, \mathbf{Q})) \tag{1}$$

$$\mathbf{Y}^{SEG} = f_{\mathcal{M}}(f_{fpn}(\mathbf{F}), \mathbf{Y}^{DET}) \tag{2}$$

For DET branch in Eq.(1), the Transformer encoder $f_{enc}(\cdot)$ first aggregates the multi-scale image features $\mathbf{F}$ using multi-head self-attention, and the Transformer decoder $f_{dec}(\cdot)$ takes the aggregated image embedding, attention mask $\mathbf{A}$ and group queries $\mathbf{Q}$ as inputs to conduct interactive feature learning and global reasoning about text instances. The output $\mathbf{Y}^{DET}$ is obtained through a multi-task detection head $f_{\mathcal{H}}(\cdot)$ for each granularity. For SEG branch in Eq.(2), we adopt a FPN network $f_{fpn}(\cdot)$ to obtain a fused image feature from $\mathbf{F}$. The task-agnostic mask decoder $f_{\mathcal{M}(\cdot)}$ takes the fused image feature and detection output $\mathbf{Y}^{DET}$ as inputs to conduct detection-conditioned image segmentation.

3.2. Multi-granularity Detection Framework

We employ the advanced Transformer-based object detection algorithm DINO (Zhang et al., 2022) to construct our text detection framework. To enable parallel training and inference for multi-granularity text instances, we initialize a set of learnable query embeddings for each granularity of text instance separately, forming group queries $\mathbf{Q} = \{(\mathbf{Q}_k^{word}, \mathbf{Q}_k^{line}, \mathbf{Q}_k^{para}, \mathbf{Q}_k^{page})\}_{k=1}^{N_q}$ that serve as inputs to the Transformer decoder as in Eq.(1). Here N_q is the query number of each group, which is same for all text granularities. As in Figure 2, each layer of Transformer decoder is composed of three components: 1) group-wise self-attention module with non-shared parameters for learning text queries at each granularity, 2) an interactive across-granularity attention module for correlating the intrinsic structural information between different text queries (introduced in 3.3), 3) a parameter-shared cross-attention module and feed-forward

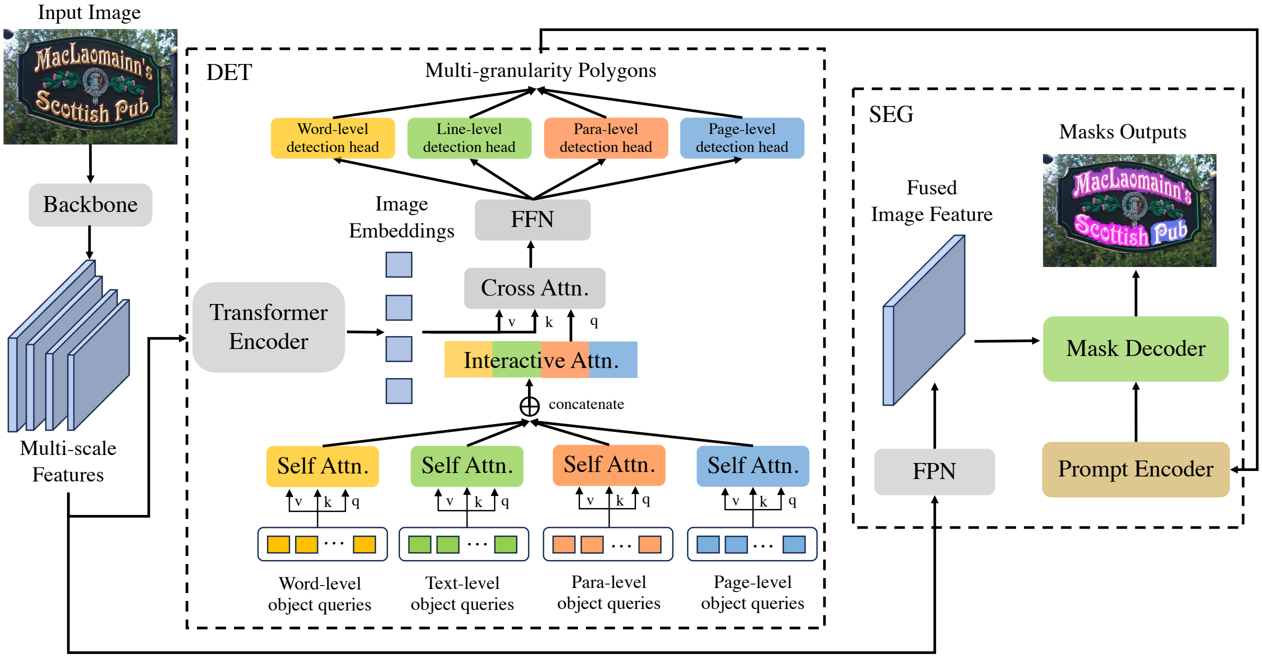
Input Image
Backbone
Multi-scale Features
DET
Multi-granularity Polygons
Word-level detection head
Line-level detection head
Para-level detection head
Page-level detection head
FFN
Image Embeddings
Transformer Encoder
Cross Attn.
v
k
q
Interactive Attn.
concatenate
Self Attn.
Self Attn.
Self Attn.
Self Attn.
Word-level object queries
Text-level object queries
Para-level object queries
Page-level object queries
SEG
Masks Outputs
Fused Image Feature
Mask Decoder
FPN
Prompt Encoder

Figure 2. Network structure of Detect Any Text (DAT). "DET" illustrates the multi-granularity detection framework with a single layer of Transformer decoder network, where the residual connection and norm layers are omitted for simplicity. "SEG" illustrates the model pipeline of prompt-based segmentation module.

network (FFN) for global reasoning about text instances at each text granularity. Additionally, a group-wise multi-task detection head $f_{\mathcal{H}} = \{f_{\mathcal{H}}^{word}, f_{\mathcal{H}}^{line}, f_{\mathcal{H}}^{para}, f_{\mathcal{H}}^{page}\}$ is added to the FFN. Here $f_{\mathcal{H}}$ for each granularity is composed of a box regression head, a box classification head and a polygon regression head, in which the polygon regression head is only added to the last layer of Transformer decoder for efficiency.

Mixed-granularity Training. The training target for optimizing multi-granularity text detection task is defined as follow:

$$\mathcal{L}^{DET} = \sum_{t=1}^{4}(\omega_t \times \mathcal{L}_t(\mathbf{Y}_t^{DET}, \hat{\mathbf{Y}}_t^{DET})) \qquad (3)$$

The subscript t in Eq.(3) refers to four different tasks of optimizing *word, line, paragraph, page* granularities. $\mathbf{Y}_t^{DET}$ is the output prediction for granularity t, and $\hat{\mathbf{Y}}_t^{DET}$ is the corresponding training label generated from ground-truth(GT). The loss function of each granularity is composed of l_1 loss for polygon regression, l_1 and GIoU (Rezatofighi et al., 2019) losses for box regression, and focal loss (Lin et al., 2017) for classification. The loss weights ω_t for multi-task learning are defined as follow:

$$\omega_t = \begin{cases} 0, & \text{if } \hat{N}_t = 0; \\ 1, & \text{if } \hat{N}_t = 1 \text{ and } \sum_{t=1}^{4} \hat{N}_t = 1; \\ \frac{1}{\sum_{t=1}^{4} \hat{N}_t}, & \text{if } \hat{N}_t = 1 \text{ and } \sum_{t=1}^{4} \hat{N}_t > 1. \end{cases} \qquad (4)$$

Here $\hat{N}_t$ is a binary indicator $(0/1)$ representing whether the label of granularity t is annotated in the GT. It is worth mentioning that the loss weights ω_t for each text granularity are dynamically adjusted within each training batch.

Discussion. Such framework design leverages the power of parallel training on diverse datasets, even those with limited annotation granularities such as single-granularity annotations or incomplete labeling schemes. Notably, the resulting model generates multi-granularity text detection outputs in one-forward-propagation, leading to significantly improved efficiency in text-related systems. Moreover, our model is capable of generating high-quality multi-granularity pseudo labels for incomplete-granularity annotated datasets. The detailed results of generated pseudo labels are present in Sec 4.4.

3.3. Across-granularity Representation Learning with Interactive Attention Module

As shown in Figure 1, text instances in natural scenes or document images are normally (but not necessarily) correlated to each other structurally among different granularities. Most existing approaches overlooked the correlation of these intrinsic linked multi-level texts, while we argue that such intrinsic correlations can be useful to facilitate a deeper understanding and integration of textual instance representations. Motivated by this, we introduce an across-granularity interactive attention module to text detection decoder, facilitating the transmission of learned query embeddings across adjacent granularities during representation learning. As shown in Figure 3, after group-wise self-attention for each level of query embeddings, we concatenate them to form a global query embedding $\mathbf{Q}_g \in \mathbb{R}^{4N_q \times c}$, here c is the embedding dimensions for each query. We employ a global attention mask $\mathbf{A}$ with interaction factor $\mathcal{I}$ to conduct across-granularity self-attention for global query embedding $\mathbf{Q}_g$. The attention mask $\mathbf{A}$ is a binary matrix with a shape of $4N_q \times 4N_q$ (here we omit the batch size and numbers of attention heads for simplicity). In this layer of global self-attention, the interactions between query embeddings of different granularities depend on the weight parameters at corresponding positions in the attention mask $\mathbf{A}$. When $\mathcal{I} = 1$, the global query embedding is enabled to interactive across different levels of query embeddings only in adjacent granularities, i.e., the interactions of word-to-line, line-to-para, para-to-page from bottom-up, and page-to-para, para-to-line, line-to-word from top-down. When $\mathcal{I}$ is increased to 2 and 3, more extensive cross-granularity interactions are allowed during global self-attention. After this interactive across-granularity attention computation, we extract each group of query embeddings from the corresponding positions of global query embedding $\mathbf{Q}_g$ to obtain the updated $\mathbf{Q} = \{(\mathbf{Q}_k^{word}, \mathbf{Q}_k^{line}, \mathbf{Q}_k^{para}, \mathbf{Q}_k^{page})\}_{k=1}^{N_q}$ for the subsequent cross-attention.

3.4. Prompt-based Segmentation Module

To address the problem of arbitrarily-shaped text localization and accurate document page segmentation, we introduce a hierarchical prompt-based segmentation module to perform foreground-background segmentation of the multi-granularity text instances. As illustrated in Figure 2, the learnable parameters for segmentation module is composed of a FPN layer for extracting a fused image feature, a Prompt Encoder for representing multi-granularity polygons from detection module, and a Mask Decoder for generating fine-grained masks of given text regions. Following SAM (Kirillov et al., 2023), we initialize a group of learnable embeddings to sum with the positional encodings of each polygon coordinates for encoding multi-granularity polygons within Prompt Encoder. The Mask Decoder includes blocks of prompt self-attention, two-way cross-attention, up-sampling, and MLP modules.

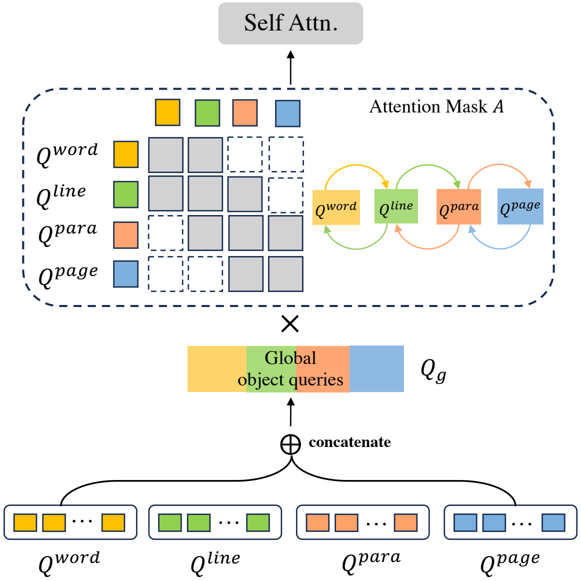

Figure 3. Illustration of across-granularity representation learning with interactive attention module (interaction factor $\mathcal{I} = 1$).

The introduced mask decoder can perform more fine-grained text contour segmentation by using the multi-granularity detection results as prompts, which can significantly improve the detection of curved and arbitrarily-shaped texts, as well as the segmentation of complex layouts and page bodies.

The training target for optimizing segmentation module is a linear combination of mean-square-error(MSE) and dice losses (Milletari et al., 2016) for mask prediction and MSE loss for intersection-over-union(IoU) prediction:

$$\begin{aligned}\mathcal{L}^{SEG} &= 5 \times \mathcal{L}_{mse}(\mathbf{Y}^{SEG}, \hat{\mathbf{Y}}^{SEG}) \\ &+ \mathcal{L}_{dice}(\mathbf{Y}^{SEG}, \hat{\mathbf{Y}}^{SEG}) \\ &+ \mathcal{L}_{iou}(\mathbf{Y}^{SEG}, \hat{\mathbf{Y}}^{SEG}) \qquad (5)\end{aligned}$$

4. Experiments

4.1. Experiment Setup

Datasets and Evaluation Protocol. Our experimental framework utilized popular benchmarks corresponding to each level of text granularity. For word detection, we used the ICDAR2015 (Karatzas et al., 2015) and Total-Text (Ch'ng et al., 2020) datasets; for line detection, CTW1500 (Liu et al., 2019) and MSRA-TD500 (Yao et al.,

Table 1. Results for DAT and other SOTA models on benchmark test sets of scene text detection, layout analysis, and document page segmentation. “WORD, LINE, PARA, PAGE” indicate the word detection, text-line detection, layout analysis and page segmentation respectively. “P, R, F” stand for Precision, Recall, Fscore metrics. The best and secone-best metrics are highlighted in **bold** and blue.

METHOD	WORD						LINE						PARA	PAGE
	ICDAR2015			TOTAL-TEXT			CTW1500			MSRA-TD500			M6DOC	DIW
	P	R	F	P	R	F	P	R	F	P	R	F	MAP	MIOU
SAM (KIRILLOV ET AL., 2023)	-	-	-	-	-	-	-	-	-	-	-	-	-	84.1
DEEPLABV3+ (CHEN ET AL., 2018)	-	-	-	-	-	-	-	-	-	-	-	-	-	98.61
M6DOC (CHENG ET AL., 2023)	-	-	-	-	-	-	-	-	-	-	-	-	63.8	-
HIERTEXT (LONG ET AL., 2022)	-	-	-	85.49	**90.53**	87.94	84.56	87.44	85.97	88.04	87.44	87.70	-	-
SIR (QIN ET AL., 2023)	90.4	85.4	87.8	90.9	85.6	88.2	87.4	83.7	85.5	93.6	86.0	89.6	-	-
DPTEXT-DETR (YE ET AL., 2023)	-	-	-	91.8	86.4	89.0	91.7	86.2	88.8	-	-	-	-	-
UNITS (KIL ET AL., 2023)	**94.0**	91.0	92.5	-	-	89.8	-	-	-	-	-	-	-	-
ESTEXTSPOTTER (HUANG ET AL., 2023)	92.5	89.6	91.0	92.0	88.1	90.0	91.5	88.6	90.0	92.9	86.3	89.5	-	-
DAT-DET (OURS)	90.87	94.51	**92.66**	93.98	88.17	90.98	89.25	89.28	89.26	**95.11**	86.63	**90.67**	-	-
DAT-SEG (OURS)	87.46	**95.76**	91.42	**95.04**	89.16	**92.01**	**92.51**	**90.94**	**91.72**	92.74	**88.60**	90.62	**65.7**	**98.65**

2012) were employed; M6Doc (Cheng et al., 2023) facilitated our document layout analysis; and DIW (Ma et al., 2022) was the choice for page detection. Notably, ICDAR2015 and MSRA-TD500 are multi-oriented datasets annotated with quadrilateral points, Total-Text and CTW-1500 feature arbitrarily shaped texts annotated with polygon points outlining text contours. M6Doc offers a fine-grained layout analysis with 74 categories, and DIW is recognized for document dewarping, annotated with foreground page masks. For evaluation metrics, we report Precision, Recall, and F1-Score (abbreviated as “P, R, F”) for word and line detection tasks. The mean Average Precision (mAP) metric is used for layout analysis, and mean Intersection Over Union (mIoU) is used for page segmentation tasks.

Implementation Details. We adopted the Swin Transformer Large (SwinL) (Liu et al., 2021) pretrained on ImageNet-22K (Deng et al., 2009) as our initialization backbone network. For comprehensive benchmark evaluations, we trained our DAT model on a diverse set of datasets: ICDAR2015 (Karatzas et al., 2015), Total-text (Ch’ng et al., 2020), Curved SynthText (Liu et al., 2020), ICDAR-MLT (Nayef et al., 2017), ArT (Chng et al., 2019), CTW1500 (Liu et al., 2019), MSRA-TD500 (Yao et al., 2012), M6Doc (Cheng et al., 2023), DIW (Ma et al., 2022), and Doc3d (Das et al., 2019). Notably, ICDAR-MLT and ArT are multilingual datasets, with annotations for Chinese and Japanese texts at the text-line level, and annotations for other languages at the word level. To accommodate these datasets within our DAT framework, we implemented a masking strategy. Specifically, for images annotated with Chinese or Japanese texts, we masked out texts of other languages, categorizing these images under the text-line level for training. Conversely, images devoid of these languages were categorized under the word level. Further elaboration on training settings is detailed in the Appendix.

4.2. Main Results

As outlined in Section 3.3, the DAT model consists of two key branches: a text detection branch (DAT-DET) and a hierarchical segmentation branch (DAT-SEG). Table 1 demonstrates that our “all-in-one” DAT model consistently outperforms single-task models, achieving SOTA performances in all text-related tasks: scene text detection, document layout analysis, and page segmentation.

Scene Text Detection. The DAT-DET model outperformed the previous SOTA methods UNITS (Kil et al., 2023) and SIR (Qin et al., 2023) on ICDAR2015 and MSRA-TD500 multi-oriented datasets. The DAT-SEG model further boosted the F-score by 2.01 and 1.72 points on Total-Text and CTW-1500 datasets espectively, outperforming ESTextSpotter (Huang et al., 2023). Our approach obtained the highest precision on arbitrarily shaped datasets and the highest recall on multi-oriented datasets, validating the effectiveness of our multi-granularity text detection framework. Notably, the DAT-SEG model showed a slight decrease in performance on ICDAR2015 and MSRA-TD500 datasets compared to the DAT-DET model. This is attributed to the quadrilateral-based annotations used in these datasets, which does not favor the more refined outcomes of DAT-SEG. We discuss this further with visualizations in Sec 4.4.

Document Layout Analysis. For paragraph-level document layout analysis, the DAT-SEG model significantly improved performance on M6Doc dataset (Cheng et al., 2023), currently the most fine-grained dataset with 74 categories. The model’s mAP saw an increase from 63.8 to 65.7, a notable gain of 1.9 points. This improvement indicates that our model, leveraging interactive information from multiple granularities, is adept at learning more discriminative and nuanced paragraph features, enhancing its overall document layout analysis capability.

Document Page Detection. We implemented SAM (Kir-

Table 2. Ablation study on impact of each text granularity. “WORD, LINE, PARA, PAGE” indicate the word detection, text-line detection, layout analysis and page detection respectively. “P, R, F” stand for Precision, Recall, Fscore metrics.

MODEL	WORD						LINE						PARA	PAGE
	ICDAR 15			TOTAL-TEXT			CTW1500			MSRA-TD500			M6DOC	DIW
	P	R	F	P	R	F	P	R	F	P	R	F	MAP	MIOU
SINGLE GRANULARITY BASELINE	72.90	91.19	81.02	89.82	89.66	89.74	81.05	80.02	80.53	91.46	83.87	87.50	62.0	**98.67**
WORD + LINE	82.37	**97.16**	89.15	88.29	**90.97**	89.61	85.73	80.48	83.02	92.78	84.95	88.69	-	-
WORD + LINE + PARA	88.99	94.17	91.51	91.56	90.20	90.88	88.90	88.92	88.91	94.94	86.48	90.51	67.8	-
WORD + LINE + PARA + PAGE (DAT)	**90.87**	94.51	**92.66**	**93.98**	88.17	**90.98**	**89.25**	**89.28**	**89.26**	**95.11**	**86.63**	**90.67**	**70.5**	98.65

Table 3. Analysis of different attention modules. The best results are highlighted in **bold**.

MODEL	WORD						LINE						PARA	PAGE
	ICDAR 15			TOTAL-TEXT			CTW1500			MSRA-TD500			M6DOC	DIW
	P	R	F	P	R	F	P	R	F	P	R	F	MAP	F@0.9
w/o INTERACTIVE ATTENTION	76.32	87.67	81.60	85.80	83.79	84.78	77.63	78.39	78.01	92.23	82.03	86.83	54.9	85.31
BOTTOM-UP ATTENTION	81.56	95.62	88.03	90.44	90.20	90.32	83.60	83.21	83.40	88.77	**89.86**	89.31	63.0	89.05
INTERACTIVE ATTENTION ($\mathcal{I} = 1$)	**90.88**	94.51	**92.66**	**93.98**	88.17	90.98	**89.25**	**89.28**	**89.26**	**95.11**	86.63	**90.67**	70.5	**94.15**
INTERACTIVE ATTENTION ($\mathcal{I} = 2$)	83.47	**97.01**	89.73	90.76	**91.37**	**91.06**	86.48	86.11	86.30	90.28	**89.86**	90.07	**71.2**	91.00
INTERACTIVE ATTENTION ($\mathcal{I} = 3$)	76.99	89.74	82.88	86.69	84.42	85.54	75.66	79.01	77.30	89.57	84.48	86.96	61.2	87.25

illov et al., 2023) on DIW dataset with official pre-trained weights, employing its default Automatic Mask Generation configuration. The largest mask area from panoptic segmentation results was chosen as the final page segmentation outcome. In contrast, DeepLabv3+ (Chen et al., 2018) was fine-tuned on DIW and Doc3d training sets for a fair comparison. As shown in the last column of Table 1, our DAT-SEG model outperformed these algorithms on DIW dataset. This success is attributed to our robust representation learning for individual text elements and accurate page segmentation via edge features.

4.3. Ablation Study

4.3.1. IMPACT OF EACH TEXT GRANULARITY

Word + Line. Table 2 reveals that the $word + line$ model saw an 8.13-point F-score increase at the word-level on ICDAR2015 dataset, compared to the $baseline$ model (Table 2 row 1). However, on Total-Text dataset, a minor decrease of 0.13-point in F-score was observed. This drop is primarily attributed to a rise in false positives due to the integration of line-level features, as indicated by a higher recall but lower precision. On CTW1500 and MSRA-TD500 datasets, the model registered F-score improvements of 2.49 and 1.19 points respectively, showcasing the efficacy of word-level features in supporting line-level detection.

Word + Line + Paragraph. As detailed in Table 2, incorporating paragraph granularity tasks led to significant enhancements across various granularity benchmarks. Notably, the $word + line + para$ model showed a remarkable 5.89-point increase in F-score on CTW1500 dataset over the $word + line$ model. Furthermore, on M6Doc dataset, this model exhibited an impressive mAP improvement from 62.0 to 67.8 (+5.8), confirming our hypothesis that text line distributions are beneficial for detailed layout analysis and that layout structures can guide the localizations of words and text lines.

Word + Line + Paragraph + Page. Our full DAT model (Table 2 row 4), which includes page-level granularity, achieved the highest performance metrics in text detection tasks across word, line, and paragraph granularities. This highlights the value of page-level granularity in providing top-down guidance for sub-level text detection. On DIW dataset, the full model experienced a slight mIoU decrease from 98.67 to 98.65 compared to the $baseline$ model, due to the relatively simpler task of page segmentation. Nevertheless, the $baseline$ model's robust performance on this dataset already outperformed DeepLabv3+ (Chen et al., 2018) (98.61). The introduction of page detection in our full model leverages top-down feature learning at the page level, significantly enhancing sub-level text detection tasks.

4.3.2. ANALYSIS OF INTERACTIVE ATTENTION MODULE

Without Interactive Attention. The first row of Table 3 presents the model's performance without the interactive attention module. This model relies solely on group-wise self-attention and global cross-attention within the Transformer decoder. The evaluation results show that, while the model was trained using datasets of various granularities, it only attained sub-optimal detection outcomes. This underlines the importance of interactive attention in enhancing the model's learning capabilities for better performance.

Bottom-up Attention. Similar to HierText (Long et al.,

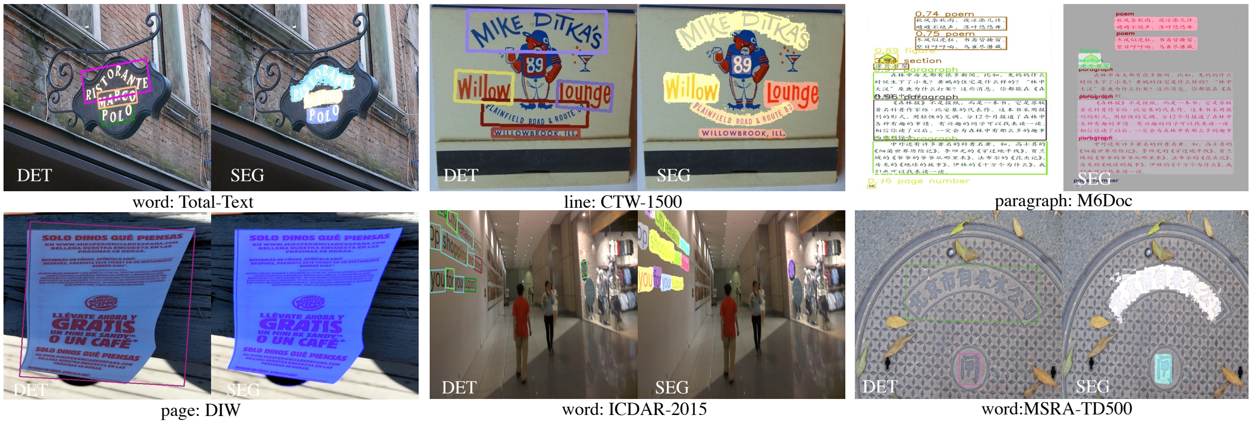

Figure 4. Visualization results of DAT on each granularity of benchmark datasets. “DET” and “SEG” indicate text detection and segmentation results respectively. For multi-oriented datasets ICDAR-2015 and MSRA-TD500, the DAT-SEG model further refined detection results, particularly for curved texts. However, a slight decline in benchmark evaluation results occurred due to the quadrilateral-based annotations.

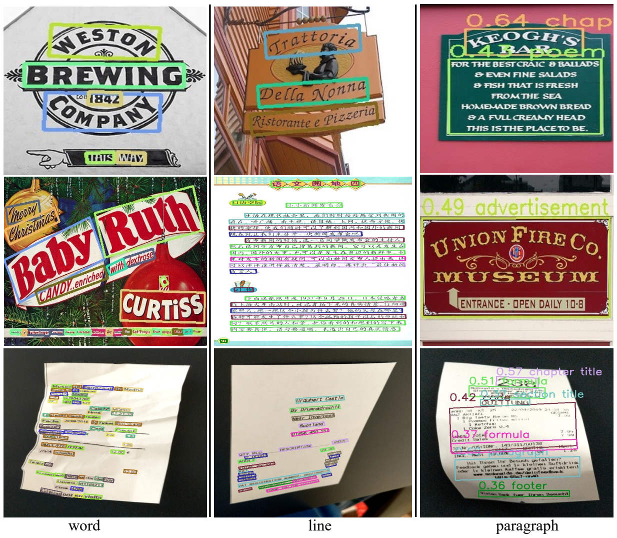

Figure 5. Multi-granularity pseudo labels produced by DAT. From left to right: text detection results at the word, line and paragraph levels. Note that these datasets do not have the corresponding GT annotations for these specific granularities.

2022), we constructed a bottom-up attention scheme to investigate the effects of unidirectional interactive attention. As shown in the second row of Table 3, incorporating bottom-up attention significantly improved text detection across all granularities compared to models without interactive attention. Notably, this approach improved the F-score on Total-Text from 87.94 to 90.32 (+2.38), and on MSRA-TD500 from 87.70 to 89.31 (+1.61), surpassing original HierText metrics. Unlike HierText, our method does not require hierarchical text annotations and benefits single granularity detection by increasing training data volume.

Interactive Attention. We observed the highest detection performance metrics at the line and page granularities when $\mathcal{I} = 1$ as demonstrated in Table 3. Increasing $\mathcal{I}$ to 2 improved performance on Total-Text and M6Doc but resulted in a decline in other datasets such as ICDAR2015, CTW1500, and DIW, where F-scores dropped by 2.93, 2.96, and 3.15, respectively. This change indicates a trade-off between higher recall and lower precision, suggesting that $\mathcal{I} = 2$ model introduced more false positives, reducing text detection accuracy. When $\mathcal{I}$ was further increased to 3, there was a sharp drop in performance across all granularities, nearly mirroring the model without interactive attention. This implies that a completely unrestricted information interaction (no attention mask) is detrimental to feature learning across granularities due to difficulty in distinguishing relevant features from noise. Therefore, we selected $\mathcal{I} = 1$ for our experimental benchmark evaluations on public datasets.

4.4. Qualitative Results

Figure 4 presents the qualitative results of DAT across different text granularities. Our model shows notable detection performances in arbitrarily-shaped datasets Total-Text and CTW-1500. The model also demonstrates its proficiency in fine-grained paragraph classification on M6Doc dataset, as well as accurate page segmentation on DIW dataset. For the multi-oriented datasets ICDAR-2015 and MSRA-TD500 (the last two blocks in Figure 4, the segmentation results output by the DAT model after introducing the prompt-based segmentation module further optimize the text contours within the detected polygons, especially for curved texts. This demonstrates the effectiveness of the proposed segmentation module for the overall multi-granularity detection framework. Thanks to the multi-granularity detection framework design and the across-granularity interactive attention module, our DAT model is capable of generating high-quality pseudo labels for incomplete-granularity annotated datasets as demonstrated in Figure 5, more detailed analyses are provided in the Appendix.

4.5. Discussion

Computational Cost Analysis. The number of parameters (Params) and GFLOPS of our proposed DAT-DET model are 228.29M and 394 respectively, and the complete DAT model (DET+SEG) has a Params of 284.65M and GFLOPS of 474. Our method has approximately 2 times the GFLOPS of single-task SOTA method DPText-DETR (Ye et al., 2023) (GFLOPS=249), but the training/testing speeds remain competitive or even faster than SOTA methods. For instance, the training and testing FPS of our DAT-DET model are 1.4 and 3.57, respectively. In contrast, the training and testing FPS of DPText-DETR (Ye et al., 2023) are 0.56 and 3.84, respectively. Moreover, our approach achieves the detection of text at four different granularities using a single unified model, with the only cost being a slight increase in GPU memory usage (31.21G). In contrast, previous SOTA methods dedicated to single-task operations would require training and testing separate models for multi-granularity tasks, essentially necessitating four times the amount of time for training and testing. These analyses not only highlight our model's superior efficiency and effectiveness but also underscore its innovation in handling multi-granularity text detection tasks within a single framework.

Limitations. The limitations of our DAT method can be summarized as follows: (1) The number of parameters and GFLOPS of our DAT model are relatively larger than previous single-task text detection models, requiring more GPU memory and a longer training cycle, but the unified framework is still more cost-effective than the sum of the three independent task models, as discussed above. (2) Our model shows a low utilization rate of multilingual training data (such as MLT and ArT). Due to the different annotation granularities of different languages in existing multilingual datasets (e.g., English annotations at the word level, Chinese at the line level), we have not yet been able to clarify whether the annotation granularity of other languages is word or line, except for Chinese and English. As a result, we have only used incomplete training samples when training the MLT set (described in Section 4.1). In the future, we will explore the possibility of unified training that includes more languages with different granularities.

5. Conclusion

In this paper, we introduce a novel multi-granularity text detection paradigm, termed as "DAT". Inspired by the inherent structural relationships among different text granularities in natural scenes, we propose a bi-directional interactive attention module within the text detection decoder to bolster representation learning across all granularities. Particularly, our approach is distinguished by its independence from complete granularity data annotations and the capability of parallel training of a single model for concurrent text detection tasks across word, line, paragraph, and page granularities within a unified detection framework. Our extensive experiments on public datasets reveal that DAT significantly enhances text detection performance at all levels of granularity, establishing a new benchmark for State-of-the-Art (SOTA) in multi-granularity text detection models. Additionally, we have integrated a prompt-based segmentation module to accurately localize arbitrarily-shaped texts and segment document pages. These innovative designs allow our model to outperform other SOTA single-task models across a variety of benchmarks, including scene text detection, document layout analysis, and page segmentation.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Baek, Y., Lee, B., Han, D., Yun, S., and Lee, H. Character region awareness for text detection. In *CVPR*, pp. 9365–9374, 2019.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In *ECCV*, pp. 213–229. Springer, 2020.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pp. 801–818, 2018.

Chen, Q., Chen, X., Wang, J., Zhang, S., Yao, K., Feng, H., Han, J., Ding, E., Zeng, G., and Wang, J. Group detr: Fast detr training with group-wise one-to-many assignment. In *ICCV*, pp. 6633–6642, 2023.

Cheng, H., Zhang, P., Wu, S., Zhang, J., Zhu, Q., Xie, Z., Li, J., Ding, K., and Jin, L. M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In *CVPR*, pp. 15138–15147, 2023.

Chng, C. K., Liu, Y., Sun, Y., Ng, C. C., Luo, C., Ni, Z., Fang, C., Zhang, S., Han, J., Ding, E., et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *ICDAR*, pp. 1571–1576. IEEE, 2019.

Ch'ng, C.-K., Chan, C. S., and Liu, C.-L. Total-text: toward orientation robustness in scene text detection. *IJDAR*, 23 (1):31–52, 2020.

Das, S., Ma, K., Shu, Z., Samaras, D., and Shilkrot, R. Dewarpnet: Single-image document unwarping with stacked 3d and 2d regression networks. In *ICCV*, pp. 131–140, 2019.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. Ieee, 2009.

He, M., Liao, M., Yang, Z., Zhong, H., Tang, J., Cheng, W., Yao, C., Wang, Y., and Bai, X. Most: A multi-oriented scene text detector with localization refinement. In *CVPR*, pp. 8813–8822, 2021.

Huang, M., Zhang, J., Peng, D., Lu, H., Huang, C., Liu, Y., Bai, X., and Jin, L. Estextspotter: Towards better scene text spotting with explicit synergy in transformer. In *ICCV*, pp. 19495–19505, 2023.

Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V. R., Lu, S., et al. Icdar 2015 competition on robust reading. In *ICDAR*, pp. 1156–1160. IEEE, 2015.

Kil, T., Kim, S., Seo, S., Kim, Y., and Kim, D. Towards unified scene text spotting based on sequence generation. In *CVPR*, pp. 15223–15232, 2023.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. In *ICCV*, pp. 4015–4026, 2023.

Li, F., Zhang, H., Liu, S., Guo, J., Ni, L. M., and Zhang, L. Dn-detr: Accelerate detr training by introducing query denoising. In *CVPR*, pp. 13619–13627, 2022.

Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., and Zhou, M. Docbank: A benchmark dataset for document layout analysis. *arXiv preprint arXiv:2006.01038*, 2020.

Liao, M., Wan, Z., Yao, C., Chen, K., and Bai, X. Real-time scene text detection with differentiable binarization. In *AAAI*, volume 34, pp. 11474–11481, 2020.

Liao, M., Zou, Z., Wan, Z., Yao, C., and Bai, X. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *TPAMI*, 45(1):919–931, 2022.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *ICCV*, pp. 2980–2988, 2017.

Liu, Y., Jin, L., Zhang, S., Luo, C., and Zhang, S. Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recognition*, 90:337–345, 2019.

Liu, Y., Chen, H., Shen, C., He, T., Jin, L., and Wang, L. Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In *CVPR*, pp. 9809–9818, 2020.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pp. 10012–10022, 2021.

Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., and Raptis, M. Towards end-to-end unified scene text detection and layout analysis. In *CVPR*, pp. 1049–1059, 2022.

Ma, K., Das, S., Shu, Z., and Samaras, D. Learning from documents in the wild to improve document unwarping. In *ACM SIGGRAPH*, pp. 1–9, 2022.

Milletari, F., Navab, N., and Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, pp. 565–571. Ieee, 2016.

Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., Luo, Z., Pal, U., Rigaud, C., Chazalon, J., et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *ICDAR*, volume 1, pp. 1454–1459. IEEE, 2017.

Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A. S., and Staar, P. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *KDD*, pp. 3743–3751, 2022.

Qin, X., Lyu, P., Zhang, C., Zhou, Y., Yao, K., Zhang, P., Lin, H., and Wang, W. Towards robust real-time scene text detection: From semantic to instance representation learning. In *ACM Multimedia*, pp. 2025–2034, 2023.

Reza, M. M., Rakib, M. A., Bukhari, S. S., and Dengel, A. A robust page frame detection method for complex historical document images. In *ICPRAM*, pp. 556–564, 2019.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, pp. 658–666, 2019.

Shafait, F., Van Beusekom, J., Keysers, D., and Breuel, T. M. Page frame detection for marginal noise removal from scanned documents. In *Image Analysis: 15th Scandinavian Conference*, pp. 651–660. Springer, 2007.

Shafait, F., Van Beusekom, J., Keysers, D., and Breuel, T. M. Document cleanup using page frame detection. *IJDAR*, 11:81–96, 2008.

Stamatopoulos, N., Gatos, B., and Georgiou, T. Page frame detection for double page document images. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pp. 401–408, 2010.

Tang, J., Yang, Z., Wang, Y., Zheng, Q., Xu, Y., and Bai, X. Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping. *Pattern recognition*, 96:106954, 2019.

Tian, Z., Shu, M., Lyu, P., Li, R., Zhou, C., Shen, X., and Jia, J. Learning shape-aware embedding for scene text detection. In *CVPR*, pp. 4234–4243, 2019.

Wang, F., Chen, Y., Wu, F., and Li, X. Textray: Contour-based geometric modeling for arbitrary-shaped scene text detection. In *ACM Multimedia*, pp. 111–119, 2020a.

Wang, W., Xie, E., Song, X., Zang, Y., Wang, W., Lu, T., Yu, G., and Shen, C. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In *ICCV*, pp. 8440–8449, 2019a.

Wang, X., Jiang, Y., Luo, Z., Liu, C.-L., Choi, H., and Kim, S. Arbitrary shape scene text detection with adaptive text region representation. In *CVPR*, pp. 6449–6458, 2019b.

Wang, Y., Xie, H., Zha, Z.-J., Xing, M., Fu, Z., and Zhang, Y. Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection. In *CVPR*, pp. 11753–11762, 2020b.

Xie, E., Zang, Y., Shao, S., Yu, G., Yao, C., and Li, G. Scene text detection with supervised pyramid context network. In *AAAI*, volume 33, pp. 9038–9045, 2019.

Xie, G.-W., Yin, F., Zhang, X.-Y., and Liu, C.-L. Document dewarping with control points. In *ICDAR*, pp. 466–480. Springer, 2021.

Xue, C., Tian, Z., Zhan, F., Lu, S., and Bai, S. Fourier document restoration for robust document dewarping and recognition. In *CVPR*, pp. 4573–4582, 2022.

Yao, C., Bai, X., Liu, W., Ma, Y., and Tu, Z. Detecting texts of arbitrary orientations in natural images. In *CVPR*, pp. 1083–1090. IEEE, 2012.

Ye, M., Zhang, J., Zhao, S., Liu, J., Du, B., and Tao, D. Dptext-detr: Towards better scene text detection with dynamic points in transformer. In *AAAI*, volume 37, pp. 3241–3249, 2023.

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., and Shum, H.-Y. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

Zhong, X., Tang, J., and Yepes, A. J. Publaynet: largest dataset ever for document layout analysis. In *ICDAR*, pp. 1015–1022. IEEE, 2019.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021a.

Zhu, Y., Chen, J., Liang, L., Kuang, Z., Jin, L., and Zhang, W. Fourier contour embedding for arbitrary-shaped text detection. In *CVPR*, pp. 3123–3131, 2021b.

A. Training Details

Our multi-granularity text detection framework was implemented on 8 NVIDIA A100 GPUs. During the model training phase, the batch size is set to 8 (1 per single GPU). The query number N_q of each group (Sec 3.2) is set to 900. We train our full DAT model using public datasets of all granularities, with total 120 epochs. The base learning rate is 1×10^{-4} and reduced to 1×10^{-5} at the 66-th epoch and 1×10^{-6} at the 99-th epoch. For our proposed mixed-granularity training(Sec 3.2), the weight of l_1 loss is 5.0, the weight of GIoU loss is 2.0, and the weight of focal loss is 1.0. We choose AdamW with a weight decay parameter of 1×10^{-4} as our optimizer. The number of both encoder and decoder layers is set to 6. The size of fused image feature after the FPN layer (Figure 2) is $\frac{1}{8}$ of the original input image size. We adopt multiple data augmentation strategies for training DAT-DET module including: 1) randomly flipping the image with a probability of 0.3; 2) randomly rotating the image within a range of -45 to 45 degrees with a probability of 0.5; 3) random color distortion with a probability of 0.1; 4) randomly cropping the image with a probability of 0.3; 5) randomly resizing the shorter size of input images within a range of 480 to 800 with an interval of 32, while constraining the longer size within 1333 pixels. No additional data augmentation strategies were used when training the DAT-SEG module. During the model testing phase, we uniformly resize the input images to 800 × 1333 in height and width.

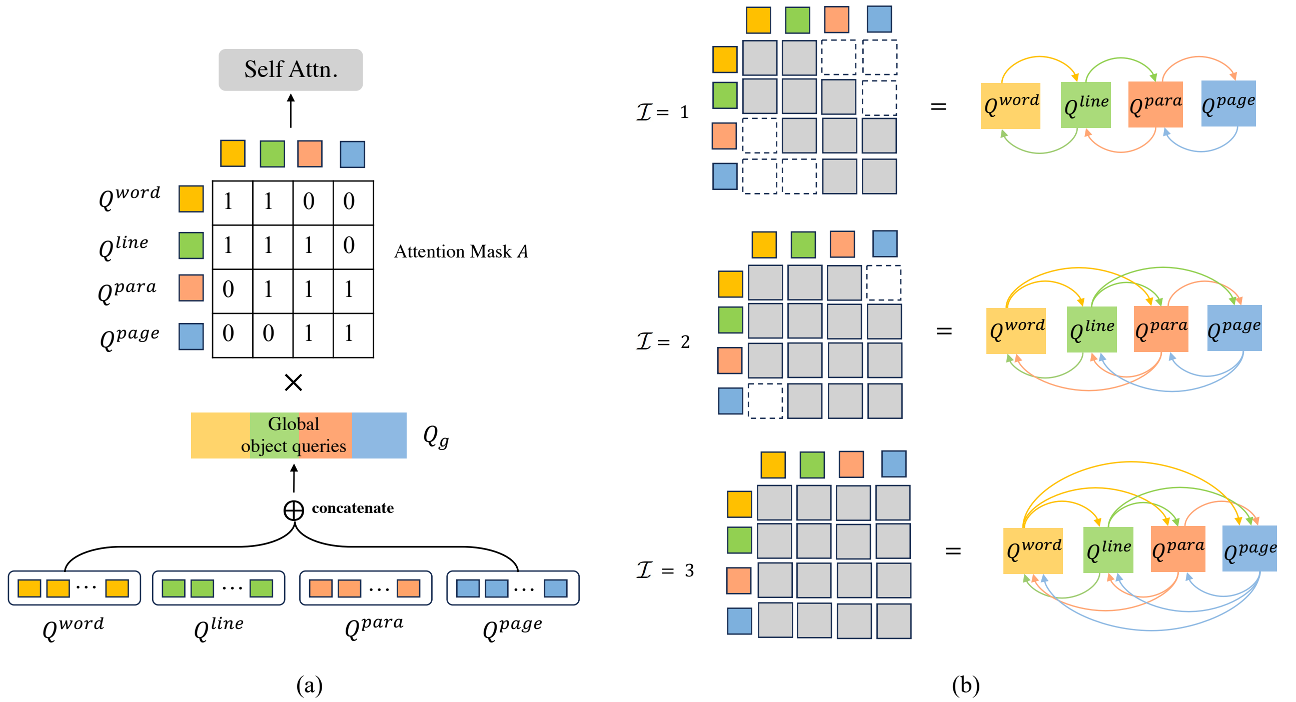

Figure 6. Illustration of interactive attention module with different interaction factors $\mathcal{I}$.

B. Explanation for the interactive attention module with different interaction factors

Within our proposed multi-granularity text detection framework, the feature interactions between text queries of different granularities are facilitated through a global self-attention layer, which is guided by an attention mask **A**. This interactive can be adjusted based on the interactive factor $\mathcal{I}$. Specifacally, as illustrated in Figure 6, when $\mathcal{I} = 1$, the global query embedding is enabled to interactive across different levels of query embeddings only in adjacent granularities, i.e., the interactions of word-to-line, line-to-para, para-to-page from bottom-up, and page-to-para, para-to-line, line-to-word from top-down. When $\mathcal{I}$ is increased to 2 and 3, more extensive cross-granularity interactions are allowed during global self-attention. Specifically, the bi-directional interactions between word-para, line-page are enabled when $\mathcal{I} = 2$, and word-page is enabled when $\mathcal{I} = 3$. It is worth mentioning that when $\mathcal{I} = 3$, query embeddings of different granularities are fully connected, meaning no mask is applied to this global self-attention module.

The proposed across-granularity attention module can effectively correlate the intrinsic structural information among text queries by learning representations from other granularity of query embeddings and enabling the duplicate removal for instances. By doing so, it facilitates a deeper understanding and integration of textual instance representations from bottom-up and top-down, ranging from individual words to entire page.

C. Qualitative results of multi-granularity pseudo labels

Our proposed multi-granularity text detection framework, equipped with a mixed-granularity training strategy, supports parallel training using datasets with incomplete-granularity annotations. More importantly, after training on multi-granularity public datasets, the resulting DAT model is capable of generating high-quality pseudo labels for various text granularities. This feature significantly enhances the model's utility and applicability, especially in scenarios where comprehensive annotations are not readily available. We provided some qualitative results of multi-granularity pseduo labels produced by our DAT-DET model as shown in Figure 7. It is worth mentioning that the Total-Text dataset is annotated at the word granularity, while the CTW-1500 and MSRA-TD500 datasets are annotated at the line granularity. Additionally, the M6Doc dataset is annotated at the paragraph granularity, the DIW and Doc3D datasets are annotated at the page granularity. Figure 7 illustrates pseudo labels of text detection branch at different text granularities: the first row shows word-level pseudo labels, the second row presents line-level pseudo labels, and the third row features paragraph-level pseudo labels. Figure 8 further visualizes the multi-granularity pseudo labels of segmentation task produced by our DAT-SEG model. Thanks to our well-designed multi-granularity detection framework and interactive cross-granularity representation learning, our model is capable of producing quite promising text detection results at the word, line, paragraph and page-level without the need for corresponding annotated data for training. It also demonstrates strong multi-granularity text detection and segmentation capabilities in complex scenarios, such as dense text lines (M6Doc) and rich texts in natural scenes (DIW & Doc3D).

Our failure cases primarily focus on blurred small text instances and extremely severe occlusions, as shown in the partial image regions of Figure 8 in our paper.

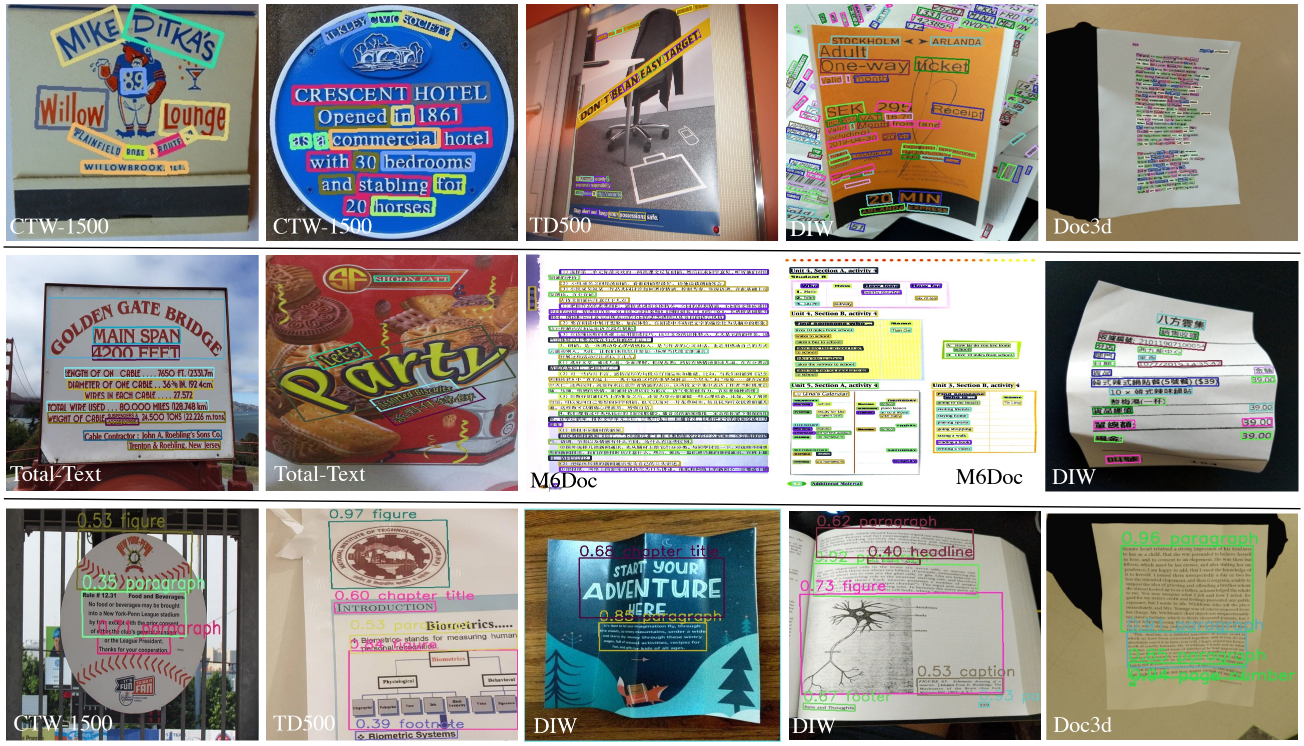

Figure 7. Qualitative results of multi-granularity pseudo labels on public benchmarks produced by our DAT-DET model. From top to bottom: visualization results of the produced pseudo labels at word, line, paragraph levels respectively.

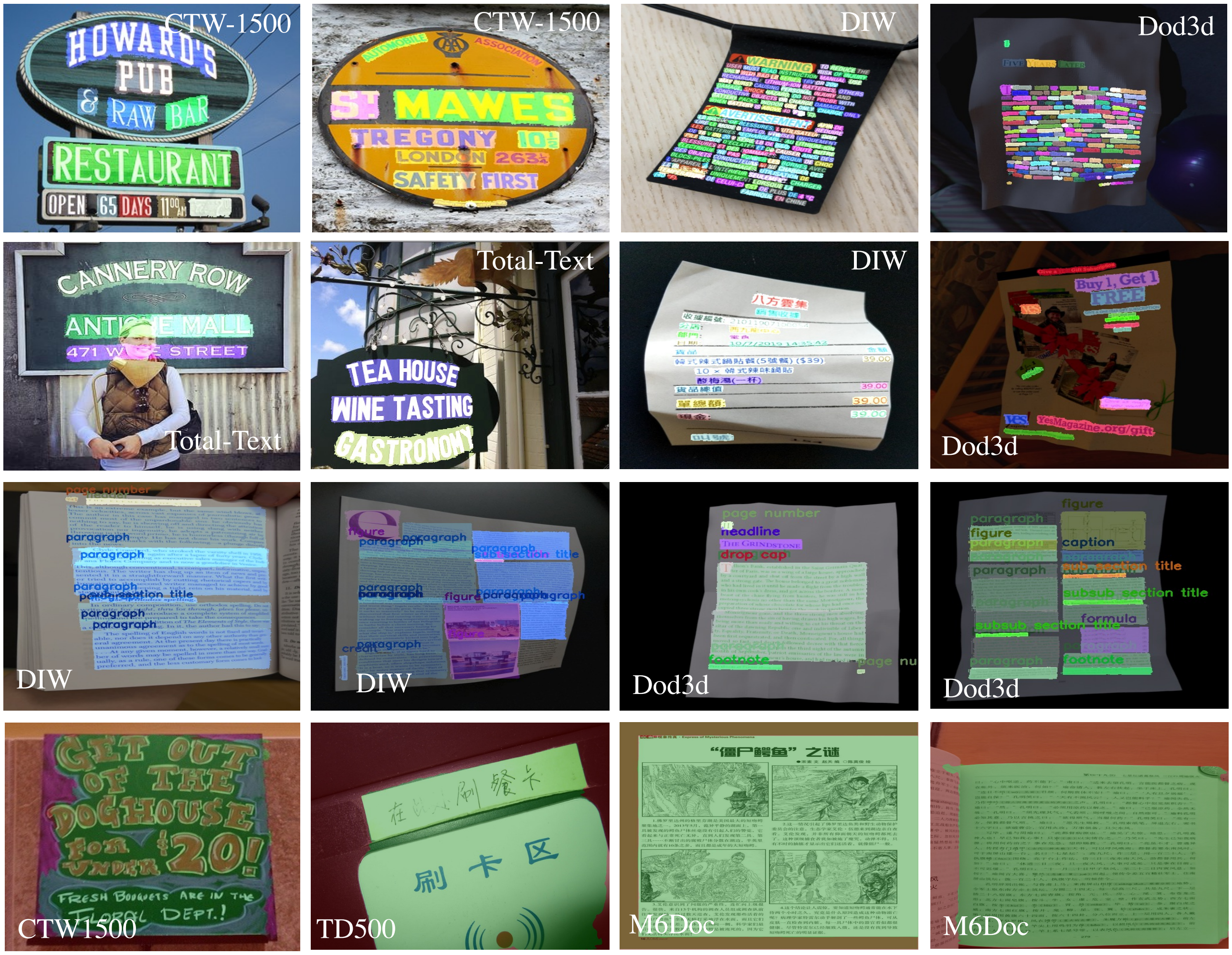

Figure 8. Qualitative results of multi-granularity pseudo labels on public benchmarks produced by our DAT-SEG model. From top to bottom: visualization results of the produced pseudo labels at word, line, paragraph, page levels respectively.