DTDT: Highly Accurate Dense Text Line Detection in Historical Documents via Dynamic Transformer

Haiyang Li[1], Chongyu Liu[1], Jiapeng Wang[1], Mingxin Huang[1], Weiying Zhou[1], and Lianwen Jin[1,2(✉)]

[1] South China University of Technology, Guangzhou, China
eelwjin@scut.edu.cn
[2] SCUT-Zhuhai Institute of Modern Industrial Innovation, Zhuhai, China

Abstract. Text detection in historical documents is challenging owing to the dense distribution of texts with diverse scales and complex layouts, resulting in low detection accuracy under high Intersection over Union (IoU) conditions. Historical document digitization requires highly accurate detection results to preserve the contents completely. In this paper, we present an end-to-end text detection framework, namely **D**ynamic **T**ext **D**etection **T**ransformer (**DTDT**), for dense text detection in historical documents under high accuracy requirements. We introduce a deformable convolution-based dynamic encoder to strengthen the text representation ability at different scales. In addition, the parallel dynamic attention heads are designed to facilitate better interaction between the box and mask branches to obtain accurate text detection results. Experiments on the MTHv2 and ICDAR 2019 HDRC-CHINESE (short for "IC19 HDRC") datasets show that the proposed DTDT method achieves state-of-the-art performance. Furthermore, our DTDT achieves competitive results in layout analysis on SCUT-CAB benchmark, demonstrating its excellent generalization capabilities.

Keywords: Text Detection · Detection Transformer · Historical Document Understanding

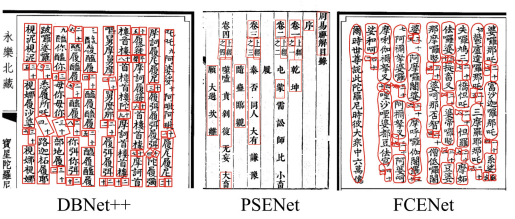
1 Introduction

Historical document digitization, which facilitates the preservation and understanding of the knowledge and insights that are contained in ancient books, has attracted increasing research attention [2,6,10,29,38]. The aim of text line detection, which is a critical step of historical document digitization, is to locate text instances. Accurate text detection is beneficial for subsequent tasks such as text recognition and ancient book restoration. Moreover, accurate text line detection results can effectively reduce the difficulty of layout analysis, which aims to locate and categorize document elements such as figures, tables and paragraphs.

With the rapid development of deep learning, scene text detection methods have made significant success on various benchmarks [19,47,51,57]. However, it is difficult for these methods to perform well on complex historical documents

G. A. Fink et al. (Eds.): ICDAR 2023, LNCS 14187, pp. 381–396, 2023.
https://doi.org/10.1007/978-3-031-41676-7_22

with dense text alignment. Figure 1 (a) presents the results of the scene text detection methods DBNet++ [19], PSENet [47] and FCENet [57] for historical documents. It can be observed that many of the detection results of these methods overlap with neighboring texts and do not closely match the texts, and also suffer from missed and false detections. We summarize the reasons for the insufficient generalization ability of scene text detection methods for these historical documents as follows: (1) As illustrated in Fig. 1 (b), the text distribution in the historical documents is denser than scene text images. For example, MTHv2 [29] contains an average of 33 text instances, while there are only seven text instances per image on SCUT-CTW1500 [52]. (2) Significant degradation of historical documents, including stains, seal noise, ink seepage, and breakage, makes it difficult for scene text detection methods [19,24,26,47,48,54,57] to obtain accurate detection results, which are essential for the subsequent text recognition. Figure 1 (c)-(f) show examples of the degradation of ancient documents.



(a) Inaccurate detection results from scene text detection methods

(b) Number of text instances in historical document (left) vs. scene text image (right)

(c) Stains (d) Seal noise (e) Ink seepage (f) Breakage

Fig. 1. (a) Inaccurate detection results of scene text detection methods on historical document images, (b) comparison of the number of texts of historical document and scene text image, and (c)–(f) degradation phenomena such as stains, seal noise, ink seepage, and breakage.

In this paper, to alleviate the problem of insufficient detection accuracy and difficulty in generalizing to complex layout scenarios with dense text distribution by previous methods, we propose the **D**ynamic **T**ext **D**etection **T**ransformer (**DTDT**) to adapt to the dense and multi-scale characteristics of historical document texts and to meet the requirements of high accuracy. Firstly, for the dense and multi-scale text arrangement, we present a deformable convolution-based dynamic encoder to fuse the adjacent scale features of the feature pyramid

with dynamic attention, which leverages spatial attention, channel attention, and multi-scale feature aggregation to pay attention to text features at different scales. Second, to meet the high accuracy detection requirements, we introduce a parallel dynamic attention head using a dynamic attention module to fuse the Region of Interest (RoI) and image features, and make the box and mask branches interact effectively. The parallel dynamic attention head facilitates the mutual interaction of dual-path branch information and precisely detects text regions in a continuously refined manner. Furthermore, we employ the spatial attention transform (SAT) mask head [30] to suppress background noise in the feature maps. Discrete cosine transform (DCT) is also used to encode the text masks as compact vectors for the accurate representation of text in arbitrary shapes. We conduct experiments on the historical document datasets MTHv2, IC19 HDRC and SCUT-CAB, illustrating the strong robustness and generalization ability of our model.

The contributions of this paper are summarized as follows:

- We propose an end-to-end text detection model named DTDT, which is based on a dynamic Transformer for the accurate detection of dense texts in historical documents with complex layouts.
- We introduce a deformable convolution-based dynamic encoder using dynamic attention to improve the detection performance of text at different scales, and present parallel dynamic attention heads with shared image features for joint detection and segmentation.
- We adopt the SAT mask head to suppress the background noise and employ DCT to encode arbitrary-shaped text masks while maintaining a low training complexity.
- DTDT achieves state-of-the-art results with F-measure of 97.90% and 96.62% for MTHv2 and IC19 HDRC, respectively. Furthermore, it obtains competitive results for layout analysis on SCUT-CAB, illustrating its outstanding generalization capabilities.

2 Related Work

2.1 Regression-Based Methods

Regression-based methods directly regress the bounding boxes of the text. [17] modified the aspect ratios of anchors based on SSD [23] to accommodate the scale characteristics of text lines. TextBoxes++ [32] regressed the quadrilateral vertices to detect multi-oriented text. EAST [54] generated rotated rectangles and quadrilaterals directly at the pixel level. To avoid the learning confusion caused by the order of points, OBD [24] decomposed the order of the quadrilateral label points into key edges comprising four invariant points and included a key edge module for learning the bounding boxes. To prevent entangled vertices from interfering with the learning process, DCLNet [1] regressed each side that is disentangled from the quadrilateral contour. The above methods are mainly for horizontal and multi-oriented text, and their performance degrades when

dealing with irregular text. To tackle the issue of irregular text detection, TextRay [46] represented arbitrary-shaped text in the polar system using a uniform geometric encoding. FCENet [57] mapped the text border to the Fourier domain to obtain Fourier contour embedding that fits curved text contours. Regression-based methods enjoy simple post-processing algorithms, but a complex representation design is required to fit arbitrary-shaped text. The one-stage methods [17,32,54] are slightly less accurate because they only regress once, and the two-stage methods [10,24,29] usually require the manual setting of the anchor to accommodate the multi-scale text distribution. In contrast, our method performs multiple iterations of the learnable query boxes to obtain more accurate results and proposes a dynamic encoder to fuse multi-scale features to better adapt to the textual characteristics of ancient documents.

2.2 Segmentation-Based Methods

In segmentation-based methods, text detection is considered as a segmentation problem. TextSnake [26] described the text as a series of ordered overlapping disks. PAN [48] adopted a lightweight segmentation head and a learnable post-processing method known as pixel aggregation. DBNet [18] provided differentiable binarization by adding the binarization step to the network for training. DBNet++ [19] extended DBNet by introducing an adaptive scale fusion module to enhance the scale robustness. To better distinguish adjacent text, PSENet [47] generated text segmentation maps in a progressive scale expansion manner. SAE [43] mapped pixels to an embedding space, drawing closer to pixels belonging to the same text and vice versa to divide the adjacent text more effectively. Although segmentation-based methods can be adapted to curved text, they require complex post-processing and are sensitive to background noise, and are more computationally intensive for ancient text detection owing to the dense text. Therefore, our method uses DCT to encode individual text instances to obtain a lightweight mask to reduce computational complexity. The SAT mask head is used to suppress noise in historical documents with complex layouts.

2.3 Transformer-Based Methods

Transformer [44] has attracted increasing attention in scene text detection. Raisi et al. [34] proposed a Transformer-based architecture for detecting multi-oriented text in scene images and a loss function for the rotated text detection problem. Tang et al. [41] adopted Transformer to model the relationship between a few sampled features to decode control points. DPText-DETR [51] used explicit box coordinates to generate and subsequently dynamically update position queries. The lack of interaction between the branches of the decoding the control points and those for detecting the bounding boxes prevents them from achieving better performance. Our DTDT explicitly establishes the interaction of the box and mask information for accurate text detection using the dynamic attention module.

3 Methodology

3.1 Overall Architecture of DTDT

As illustrated in Fig. 2, our proposed DTDT consists of three components: Backbone, Dynamic Encoder and Dynamic Decoder. The backbone network is composed of Swin Transformer (Swin-T) [25] and feature pyramid network (FPN) [20] to extract feature maps at different stages of the input image. The dynamic encoder applies dynamic attention to the features at different scales and fuses adjacent layer features to enhance multi-scale feature representation. The sum of the image features P extracted from x^{DE} and position embeddings E is fed into the Transformer encoder for self-attention learning to obtain enhanced features Z. Based on Sparse R-CNN [40], the RoI features U_t^{box} and U_t^{mask} together with the enhanced image features Z_{t-1} are fed into the dynamic attention module [9] of the box and mask branches, respectively, to obtain the object features O_t^{box} and O_t^{mask} for the prediction of the class, bounding box, and mask of each text instance. Finally, the output of the previous layer will be continuously refined in the dynamic decoder with parallel dynamic attention heads to obtain accurate results.

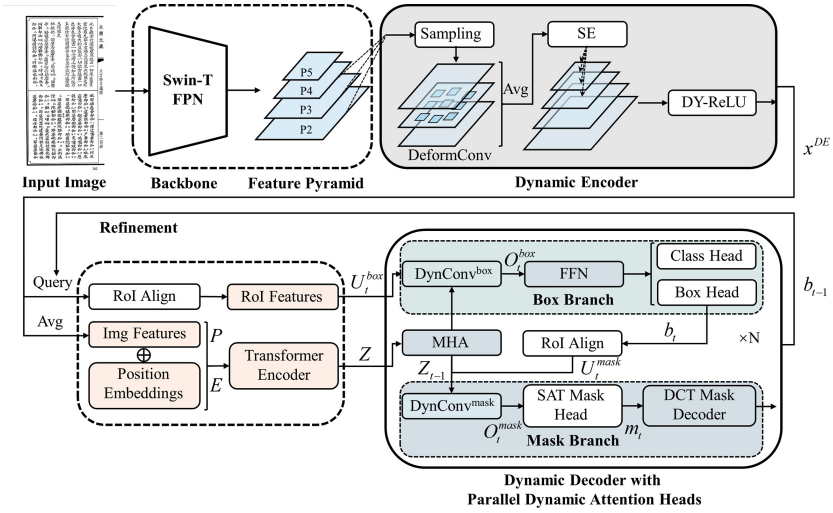


Fig. 2. Framework of proposed DTDT model. Our model consists of three components: the backbone, the dynamic encoder, and the dynamic decoder with parallel dynamic attention heads. MHA denotes the multi-head attention and FFN denotes the feedforward network.

3.2 Dynamic Encoder

In general, large and small objects are assigned to high-level and low-level feature maps to extract the RoI features, respectively. However, this may not be optimal

[22] as other unused feature maps may contain information that helps to improve the final prediction. Therefore, inspired by recent research on dynamic encoder [7,33], we introduce a dynamic encoder to perform multi-scale feature fusion on adjacent feature maps, which is depicted in the upper right part of Fig. 2. The process is divided into three steps. First, given a set of features $P = \{P_2, ..., P_k\}$ $(k = 5)$ from the feature pyramid, deformable aggregation, which consists of several deformable convolution layers [55] on each feature map and an averaging operator, is performed to simulate the spatial attention for specific regions on P_i.

This process can be formulated as follows:

$$s_i = Offset_i(P_i) \tag{1}$$

$$\begin{aligned} P_i^* = \{&DeformConv_{i-1}(Downsample(P_{i-1}), s_i), \\ &DeformConv_i(P_i, s_i), \\ &DeformConv_{i+1}(Upsample(P_{i+1}), s_i)\} \end{aligned} \tag{2}$$

$$P_i^{'} = Avg(P_i^*), \tag{3}$$

where the offset s_i that corresponds to the feature map P_i is learned using a 3×3 convolution $Offset_i$ for deformed sampling locations. The neighboring feature maps P_{i-1} and P_{i+1} are downsampled and upsampled, respectively, to the same size as P_i. Deformable convolution is performed on the sampled feature maps and P_i, and each feature map focuses on the specific position s_i that is learned from the middle layer to avoid conflicts during feature aggregation. $P_i^{'}$ is obtained by averaging each term of P_i^*.

Second, $P_i^{'}$ is used for channel attention learning with the squeeze and excitation (SE) module [13]:

$$P_i^{''} = SE(P_i^{'}). \tag{4}$$

Finally, we use the DY-ReLU [5] activation function, whose parameters are dynamically generated from the input elements to improve the feature representation capability:

$$P_i^o = DY\text{-}ReLU(P_i^{''}). \tag{5}$$

3.3 Parallel Dynamic Attention Heads

The feature maps from the dynamic encoder are cropped and aligned using RoIAlign [12] to obtain the RoI features $U \in \mathbb{R}^{k \times d \times l \times l}$ via k learnable query boxes b_t $(t = 0)$, where d is the channel dimension, and l denotes the output resolution after the pooling. The feature maps of each layer are averaged and summed to obtain the image features $P \in \mathbb{R}^{k \times d}$, which are summed with the learnable position embeddings $E \in \mathbb{R}^{k \times d}$ to be fed into the Transformer encoder and MHA module to obtain $Z_{t-1} \in \mathbb{R}^{k \times d}$. We design parallel dynamic attention heads with the RoI features U and enhanced image features Z_{t-1}, as indicated in the bottom right part of Fig. 2.

Existing methods [24,28,29] use the RoI features that are obtained from the box branch to predict the mask directly, which ignores the interaction between

the box and mask branches. As illustrated in Fig. 3 (b), we use the dynamic attention module, namely *DynConv*, for more effective interaction of the box and mask branches, thereby enabling improved results. The box branch employs $DynConv_t^{box}$ to fuse the RoI features U_t^{box} and enhanced image features Z_{t-1} to extract object features O_t^{box} for classification and bounding box regression. The mask branch leverages the RoI features U_t^{mask} that are extracted from the predicted box b_t and the enhanced image features Z_{t-1} for further fusion in $DynConv_t^{mask}$ to obtain the final detection results m_t. The above process is expressed by Eqs. 6 and 7, where $\mathcal{P}^{box}$ and $\mathcal{P}^{mask}$ denote a pooling operator for the extraction of RoI features U_t^{box} and U_t^{mask}, respectively. $\mathcal{B}_t$ denotes the box head that is stacked by three linear layers. $\mathcal{M}_t$ indicates the SAT mask head. x^{DE} is the output feature map of the dynamic encoder.

$$\begin{aligned} U_t^{box} &= \mathcal{P}^{box}(x^{DE}, b_{t-1}), \\ O_t^{box} &= DynConv_t^{box}(U_t^{box}, Z_{t-1}), \\ b_t &= \mathcal{B}_t(FFN(O_t^{box})), \end{aligned} \tag{6}$$

$$\begin{aligned} U_t^{mask} &= \mathcal{P}^{mask}(x^{DE}, b_t), \\ O_t^{mask} &= DynConv_t^{mask}(U_t^{mask}, Z_{t-1}), \\ m_t &= \mathcal{M}_t(O_t^{mask}). \end{aligned} \tag{7}$$

The above process offers two advantages: (1) it provides the mask information obtained from the supervision of the mask branch to the box branch, and (2) the collaborative interaction between the box and the mask branches is improved. Moreover, we employ the SAT [30] mask head, which has been demonstrated as effective for dense instance segmentation and exploits spatial attention to suppress noise. The implementation details of the SAT mask head are illustrated in Fig. 3 (a). Average and max pooling operations are carried out along the channel axis of the object features $O_t^{mask} \in \mathbb{R}^{14\times14\times C}$ that are obtained by $DynConv_t^{mask}$ to generate the pooling features $P_{avg}, P_{max} \in \mathbb{R}^{14\times14\times1}$, which are stacked along the channel, where C denotes the channel dimension. Subsequently, a 3×3 convolution layer is applied and the features are normalized with a sigmoid function. Finally, element-wise multiplication is performed on the object feature O_t^{mask}. A mask feature of length 40 is obtained using two convolution and linear layers.

3.4 DCT Mask Representation

The direct prediction of the two-dimensional binary grid incurs a high computational cost for large resolutions. However, fine-grained features cannot be captured on a small scale. Therefore, we apply DCT [39] to transform the text mask encoding into the frequency domain. As the energy is concentrated in the low-frequency part, we keep this part to produce a compact vector as a predictive object to accurately represent the text shape. The flow of the DCT encoding and inverse DCT (IDCT) decoding is depicted in Fig. 4.

We resize the ground truth mask $M_{gt} \in \mathbb{R}^{H \times W}$ to $M \in \mathbb{R}^{K \times K}$ during training, where H and W are the height and width of M_{gt}, and K denotes the mask size. We apply two-dimensional DCT transforms M to obtain $M_{DCT} \in \mathbb{R}^{K \times K}$.

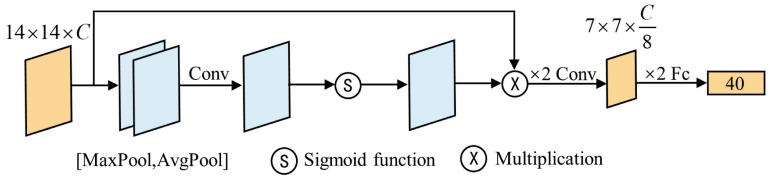
$$M_{DCT}(u,v) = \frac{2}{K} C(u)C(v) \sum_{x=0}^{K-1} \sum_{y=0}^{K-1} M(x,y) cos \frac{(2x+1)u\pi}{2K} cos \frac{(2y+1)v\pi}{2K}, \quad (8)$$

where $C(w) = \frac{1}{\sqrt{2}}$ for $w = 0$ and $C(w) = 1$ otherwise.

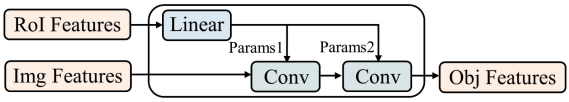
The first N-dimensional vector V is sampled from the M_{DCT} in a "zig-zag" manner to obtain the one-dimensional mask representation. We extend V to $M_{dct} \in \mathbb{R}^{K \times K}$ by filling in zeros at the end during inference and apply two-dimensional IDCT processes V to obtain $M_{IDCT} \in \mathbb{R}^{K \times K}$.

$$M_{IDCT}(x,y) = \frac{2}{K} C(u)C(v) \sum_{u=0}^{K-1} \sum_{v=0}^{K-1} M_{dct}(u,v) cos \frac{(2x+1)u\pi}{2K} cos \frac{(2y+1)v\pi}{2K} \quad (9)$$

Finally, M_{IDCT} is resized to $M_{rec} \in \mathbb{R}^{H \times W}$ using bilinear interpolation. It is worth noting that the time complexity of DCT and IDCT is $O(nlogn)$ [11].



(a) Implementation details of SAT mask head



(b) Implementation details of dynamic attention module

Fig. 3. (a) Structure of SAT mask head. (b) Dynamic attention module applied to box and mask branches.

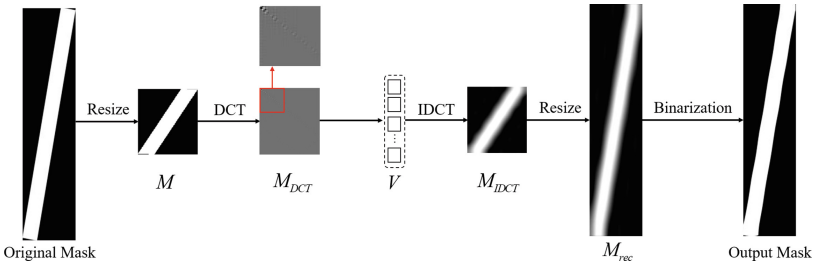


Fig. 4. DCT encoding and IDCT decoding.

3.5 Loss Function

We adopt the Hungarian algorithm [15] to match the predicted and ground truth boxes. DTDT applies a set prediction loss to the set of predictions of the categories, box coordinates, and mask representations. The total loss function can be formulated as follows:

$$\mathcal{L} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{box}\mathcal{L}_{box} + \lambda_{mask}\mathcal{L}_{mask}. \tag{10}$$

$\mathcal{L}_{box}$ is defined as:

$$\mathcal{L}_{box} = \lambda_{L1}\mathcal{L}_{L1} + \lambda_{giou}\mathcal{L}_{giou}. \tag{11}$$

$\mathcal{L}_{mask}$ is defined as:

$$\mathcal{L}_{mask} = \mathcal{L}_{L2} + \mathcal{L}_{dice}. \tag{12}$$

In the above equations, $\mathcal{L}_{cls}$ is the focal loss [21], and $\mathcal{L}_{L1}$ and $\mathcal{L}_{giou}$ are the L1 loss and the generalized IoU loss [37], respectively. $\mathcal{L}_{L2}$ is the L2 loss of the one-dimensional mask embedding before DCT decoding and $\mathcal{L}_{dice}$ is the dice loss [31] of the two-dimensional mask after IDCT decoding. λ_{cls}, λ_{box}, λ_{mask}, λ_{L_1} and $\lambda_{L_{giou}}$ are set to 2, 1, 5, 5 and 2, respectively.

4 Experiments

4.1 Datasets

MTHv2 [29] is a Chinese historical document dataset consisting of 2,399 training images and 800 testing images. The dataset includes character-level and line-level quadrilateral annotations.

ICDAR 2019 HDRC-CHINESE [38] is a large historical documents dataset of structured Chinese family records that are annotated using line-level quadrilaterals. We randomly used 10,715 images for training and 1,000 for testing among the 11,715 available images.

SCUT-CAB [6] is a complex layout analysis dataset of Chinese historical documents containing 3,200 training images and 800 testing images. SCUT-CAB contains two subsets: SCUT-CAB-Logical and SCUT-CAB-Physical, which have 27 and 4 categories, respectively. All text instances are annotated using quadrilaterals.

4.2 Implementation Details

We used Swin-T [25], pre-trained on ImageNet [8] as the backbone. The number of learnable proposal boxes was set to 500. The number of iterations was set to four to improve the accuracy. We selected a mask size of 80 $\times$ 80 and a 40-dimensional DCT mask vector. We trained DTDT for 90k iterations with a batch size of eight on two NVIDIA RTX A6000 GPUs. We used AdamW [27] as the optimizer and set an initial learning rate of $2.5e^{-5}$ and a weight decay of $1e^{-4}$. The learning rate was divided by 10 at 50% and 70% of the total number of iterations. We applied data augmentation methods including random cropping and multi-scale training. The maximum image scale was set to 1333 $\times$ 800.

4.3 Comparison with Previous Methods

We compared our method with previous state-of-the-art methods on MTHv2 and IC19 HDRC. Tables 1 and 2 display the quantitative experimental results. Figure 5 shows the qualitative results for MTHv2. Furthermore, by modifying the number of categories in the class head, we applied DTDT to the SCUT-CAB dataset to validate the potential of our method in the task of ancient book layout analysis.

Text Line Detection. The results in Tables 1 and 2 demonstrate the high accuracy and robustness of our method on these two datasets. Our method achieved an F-measure of 97.90% on MTHv2, which was 0.18% higher than the

Table 1. Detection results on MTHv2 dataset. "P", "R", and "F" indicate the precision, recall, and F-measure, respectively. **Bold** indicates the best performance. Underline indicates second best.

Method	IoU=0.5			IoU=0.6	IoU=0.7	IoU=0.8	Post-processing
	P	R	F	F	F	F	
Projection analysis [29]	–	–	69.22	66.87	60.97	–	–
EAST [54]	–	–	95.04	91.55	80.35	–	–
Ma et al. [29]	–	–	97.72	97.26	96.03	–	–
Mask R-CNN [12]	**98.17**	95.98	97.06	96.67	95.51	90.23	–
FCENet [57]	95.16	92.82	93.97	91.30	86.51	73.86	–
OBD [24]	97.83	97.43	97.63	97.32	**96.31**	90.78	–
Deformable DETR [56]	97.92	94.64	96.25	95.62	93.80	84.22	–
DBNet++ [19]	96.20	94.93	95.56	77.01	36.15	18.70	0.015s
PSENet [47]	93.97	87.84	90.80	88.65	83.68	70.96	0.022s
PAN [48]	97.18	93.14	95.12	92.55	84.63	62.74	0.011s
TextSnake [26]	95.07	89.00	91.94	90.92	89.36	84.58	0.497s
DTDT(Ours)	97.94	**97.86**	**97.90**	**97.41**	95.98	**91.18**	**0.008s**

Table 2. Detection results on IC19 HDRC dataset.

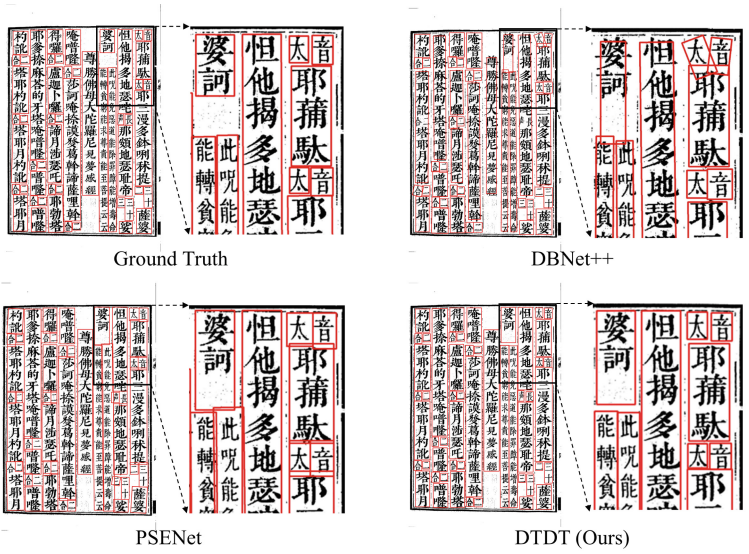
Method	IoU=0.5			IoU=0.6	IoU=0.7	IoU=0.8	Post-processing
	P	R	F	F	F	F	
Mask R-CNN [12]	96.54	96.21	96.37	94.66	88.80	70.01	–
FCENet [57]	93.63	91.50	92.55	87.74	77.25	52.12	–
OBD [24]	94.56	**97.02**	95.77	93.91	86.83	64.18	–
Deformable DETR [56]	94.43	95.72	94.57	92.55	86.27	**71.96**	–
DBNet++ [19]	96.37	95.73	96.05	90.64	75.57	48.51	0.021s
PSENet [47]	91.57	88.57	90.04	83.02	68.42	42.19	0.026s
PAN [48]	95.11	92.84	93.96	88.68	71.27	31.65	**0.012s**
TextSnake [26]	82.90	72.22	77.19	73.40	68.41	51.54	0.512s
DTDT(Ours)	**96.89**	96.35	**96.62**	**95.15**	**90.10**	71.42	0.016s

second best score when the IoU threshold was 0.5. Only three methods maintained performance above 90% when the IoU was 0.8, and our method is the best. Analogous results were obtained for IC19 HDRC. Our method obtained an F-measure of 96.62%, outperforming the second best method by 0.25%. Our method remained robust under high IoU requirements without much performance degradation compared to other methods. Our DTDT still yielded high accuracy when the IoU threshold was between 0.5 and 0.8. The post-processing times for the segmentation-based methods and our DTDT are given in Tables 1 and 2, and the results illustrate the rapidity of IDCT decoding.

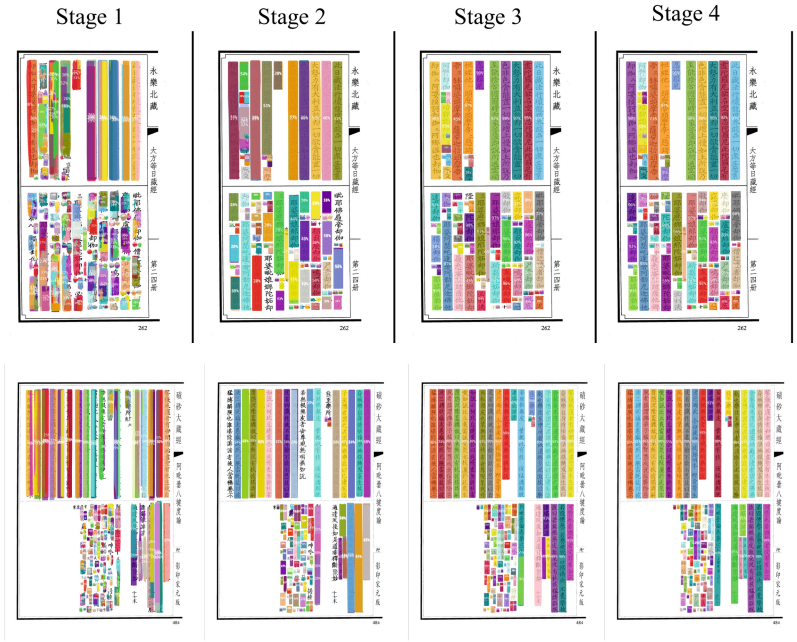
Layout Analysis Experiments. Table 3 presents the experimental results for the ancient book layout analysis on SCUT-CAB dataset [6]. The results show that our method could achieve results that are comparable to those of other methods in the physical and logical layout analysis tasks. Our model achieved the best AP75 and AP results on the physical layout analysis task, demonstrating the effectiveness of DTDT. In the logical analysis task, DTDT yielded the second best performance, which was slightly lower than that of Deformable DETR.

Table 3. AP50, AP75, and AP of each model on SCUT-CAB testing sets. AP refers to average precision, AP50 and AP75 are the average precision at IoU = 0.5 and 0.75, respectively.

Method	Physical						Logical					
	Objection Detection			Instance Segmentation			Object Detection			Instance Segmentation		
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Anchor-based one-stage												
RetinaNet [21]	91.5	82.9	74.7	91.5	81.6	73.8	78.3	61.2	55.1	78.3	61.7	55.0
YOLOv3 [35]	87.6	82.5	75.9	87.1	79.1	73.1	71.4	59.3	52.7	71.4	59.3	52.7
GFL [16]	92.6	74.8	73.7	92.6	73.2	72.4	78.1	57.8	54.1	78.1	58.8	53.5
Anchor-free one-stage												
FCOS [42]	83.2	76.0	68.9	83.1	74.7	68.1	74.1	54.4	50.2	74.0	53.4	49.1
FoveaBox [14]	91.3	82.5	74.6	91.3	80.0	73.1	80.4	60.2	54.9	80.3	60.3	54.3
Anchor-based multi-stage												
Faster R-CNN [36]	91.3	86.1	77.5	91.0	83.4	75.3	77.4	61.3	54.9	77.3	60.6	54.2
Cascade R-CNN [3]	91.4	87.8	79.9	91.4	84.8	77.4	77.5	62.3	55.9	77.5	60.9	55.4
Mask R-CNN [12]	92.1	87.7	79.1	91.7	87.2	79.5	78.5	61.9	55.1	77.7	63.1	55.3
Cascade Mask R-CNN [3]	92.1	88.6	80.9	92.1	88.4	81.0	78.0	62.7	56.8	77.9	61.8	56.3
HTC [4]	92.8	89.4	81.4	92.8	88.8	81.0	80.1	65.2	58.3	80.0	63.1	58.0
SCNet [45]	**94.1**	89.0	81.3	**94.1**	89.1	82.0	83.6	67.3	60.2	83.6	68.0	60.3
Pure Instance Segmentation												
SOLO [49]	90.7	81.6	75.2	91.2	84.3	76.7	73.8	57.7	51.6	73.2	57.8	51.5
SOLOv2 [50]	91.5	81.6	75.1	92.2	85.1	78.7	76.4	53.2	50.5	77.0	59.7	53.9
Query-based												
Deformable DETR [56]	92.7	87.9	81.0	92.5	85.1	78.8	**84.6**	**69.8**	**61.6**	**84.6**	**69.9**	**61.1**
QueryInst [9]	91.7	87.1	79.3	91.2	86.7	79.2	80.4	65.7	58.5	80.4	65.3	58.1
Multi-modality based												
VSR [53]	90.4	85.5	78.5	90.4	84.5	78.2	78.3	61.6	55.7	78.2	61.1	55.1
DTDT(Ours)	94.0	**90.0**	**83.0**	94.0	**89.6**	**82.7**	81.1	68.0	60.8	81.1	67.8	60.4



(a) Comparison of detection results on historical documents from MTHv2 dataset



(b) Qualitative results at each stage: masks from DTDT on MTHv2 dataset

Fig. 5. (a) Visualization results of our method and other scene text detection methods. Our method achieved a higher detection accuracy. (b) Qualitative experimental results for the four stages of two example images. The different colors are used to distinguish the detection results of each text instance of the model.

4.4 Ablation Study

We performed an ablation study on MTHv2 to verify the effectiveness of our proposed method. The quantitative results for different settings are presented in Table 4. The DCT resulted in a 2.78% improvement, indicating that the text shape can be more represented accurately using DCT masks. The dynamic encoder achieved performance improvements of 0.12% and 0.23% in the precision and recall, respectively, on the MTHv2 dataset, indicating its ability to improve the network's adaptation to multi-scale text. The parallel dynamic attention heads resulted in a 0.12% improvement in the F-measure. The design of the parallel dynamic attention heads provides better interaction and collaboration between the box and mask branches, facilitating the benefits of the two branches. The SAT mask, which achieved an F-measure of 97.90%, has a certain ability to suppress noise.

Table 4. Detection results for different settings of DCT, dynamic encoder, parallel dynamic attention heads, and SAT mask head on MTHv2 dataset. "DE" indicates dynamic encoder and "PDAH" indicates parallel dynamic attention heads.

DCT	DE	PDAH	SAT	P	R	F	△F
–	–	–	–	92.36	97.33	94.78	–
✓	–	–	–	97.83	97.28	97.56	↑2.78
✓	✓	–	–	97.95	97.51	97.73	↑0.17
✓	✓	✓	–	97.89	97.80	97.85	↑0.12
✓	✓	✓	✓	**97.94**	**97.86**	**97.90**	↑0.05

5 Conclusions

We proposed DTDT, which is a highly accurate text line detection method for dense text distribution of historical documents. We introduced a dynamic encoder to improve the representation ability of multi-scale text and parallel dynamic attention heads to facilitate the mutual benefits of the box and mask branches for generating more accurate text masks. The experiments demonstrated that our method achieved state-of-the-art results on historical document datasets such as MTHv2 and IC19 HDRC, and achieved comparable results on the layout analysis dataset SCUT-CAB. The potential of DTDT for text detection in modern documents and other scenarios will be explored further in future research.

Acknowledgements. This research is supported in part by NSFC (Grant No.: 61936003), Zhuhai Industry Core and Key Technology Research Project (no. 2220004002350), and Science and Technology Foundation of Guangzhou Huangpu Development District (No. 2020GH17) and GD-NSF (No.2021A1515011870).

References

1. Bi, Y., Hu, Z.: Disentangled contour learning for quadrilateral text detection. In: WACV, pp. 909–918 (2021)
2. Boillet, M., Kermorvant, C., Paquet, T.: Robust text line detection in historical documents: learning and evaluation methods. IJDAR **25**(2), 95–114 (2022)
3. Cai, Z., Vasconcelos, N.: Cascade R-CNN: delving into high quality object detection. In: CVPR, pp. 6154–6162 (2018)
4. Chen, K., et al.: Hybrid task cascade for instance segmentation. In: CVPR, pp. 4974–4983 (2019)
5. Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., Liu, Z.: Dynamic ReLU. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12364, pp. 351–367. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58529-7_21
6. Cheng, H., Jian, C., Wu, S., Jin, L.: SCUT-CAB: a new benchmark dataset of ancient Chinese books with complex layouts for document layout analysis. In: Porwal, U., Fornés, A., Shafait, F. (eds.) ICFHR 2022. LNCS, vol. 13639, pp. 436–451. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-21648-0_30
7. Dai, X., Chen, Y., Yang, J., Zhang, P., Yuan, L., Zhang, L.: Dynamic DETR: end-to-end object detection with dynamic attention. In: ICCV, pp. 2988–2997 (2021)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR, pp. 248–255. IEEE (2009)
9. Fang, Y., et al.: Instances as queries. In: ICCV, pp. 6910–6919 (2021)
10. Grüning, T., Leifert, G., Strauß, T., Michael, J., Labahn, R.: A two-stage method for text line detection in historical documents. Int. J. Doc. Anal. Recogn. (IJDAR) **22**(3), 285–302 (2019). https://doi.org/10.1007/s10032-019-00332-1
11. Haque, M.: A two-dimensional fast cosine transform. IEEE Trans. Acoust., Speech, Signal Process. **33**(6), 1532–1539 (1985)
12. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. In: ICCV, pp. 2961–2969 (2017)
13. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR, pp. 7132–7141 (2018)
14. Kong, T., Sun, F., Liu, H., Jiang, Y., Li, L., Shi, J.: FoveaBox: beyound anchor-based object detection. IEEE Trans. Image Process. **29**, 7389–7398 (2020)
15. Kuhn, H.W.: The Hungarian method for the assignment problem. Nav. Res. Logist. Q. **2**(1–2), 83–97 (1955)
16. Li, X., et al.: Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) NIPS 2020. LNCS, vol. 33, pp. 21002–21012. Curran Associates Inc, Red Hook, NY, USA (2020). https://doi.org/10.5555/3495724.3497487
17. Liao, M., Shi, B., Bai, X., Wang, X., Liu, W.: Textboxes: a fast text detector with a single deep neural network. In: AAAI (2017)
18. Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X.: Real-time scene text detection with differentiable binarization. In: AAAI, pp. 11474–11481 (2020). https://doi.org/10.1609/aaai.v34i07.6812
19. Liao, M., Zou, Z., Wan, Z., Yao, C., Bai, X.: Real-time scene text detection with differentiable binarization and adaptive scale fusion. TPAMI (2022)
20. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR, pp. 2117–2125 (2017)

21. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: ICCV, pp. 2980–2988 (2017)
22. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: CVPR, pp. 8759–8768 (2018)
23. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
24. Liu, Y., Zhang, S., Jin, L., Xie, L., Wu, Y., Wang, Z.: Omnidirectional scene text detection with sequential-free box discretization. In: IJCAI, pp. 3052–3058 (2019)
25. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: ICCV, pp. 10012–10022 (2021)
26. Long, S., Ruan, J., Zhang, W., He, X., Wu, W., Yao, C.: TextSnake: a flexible representation for detecting text of arbitrary shapes. In: ECCV, pp. 20–36 (2018)
27. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
28. Lyu, P., Liao, M., Yao, C., Wu, W., Bai, X.: Mask TextSpotter: an end-to-end trainable neural network for spotting text with arbitrary shapes. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 71–88. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_5
29. Ma, W., Zhang, H., Jin, L., Wu, S., Wang, J., Wang, Y.: Joint layout analysis, character detection and recognition for historical document digitization. In: ICFHR, pp. 31–36 (2020)
30. Mao, Q., Sun, L., Wu, J., Gao, Y., Wu, X., Qiu, L.: SATMask: spatial attention transform mask for dense instance segmentation. In: DSC, pp. 592–598 (2022)
31. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 3DV, pp. 565–571 (2016)
32. Minghui Liao, B.S., Bai, X.: Textboxes++: a single-shot oriented scene text detector. IEEE Trans. Image Process. **27**(8), 3676–3690 (2018)
33. Mishra, S.K., Sinha, S., Saha, S., Bhattacharyya, P.: Dynamic convolution-based-encoder decoder framework for image captioning in Hindi. ACM Trans. Asian Low-Resour. Lang. Inf. Process. **22**(4), 1–18 (2023)
34. Raisi, Z., Naiel, M.A., Younes, G., Wardell, S., Zelek, J.S.: Transformer-based text detection in the wild. In: CVPR Workshops, pp. 3162–3171 (2021)
35. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
36. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) NIPS 2015. LNCS, vol. 28. Curran Associates, Inc. (2015)
37. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: a metric and a loss for bounding box regression. In: CVPR, pp. 658–666 (2019)
38. Saini, R., Dobson, D., Morrey, J., Liwicki, M., Simistira Liwicki, F.: ICDAR 2019 historical document reading challenge on large structured Chinese family records. In: ICDAR, pp. 1499–1504. IEEE (2019)
39. Shen, X., et al.: DCT-Mask: discrete cosine transform mask representation for instance segmentation. In: CVPR, pp. 8720–8729 (2021)
40. Sun, P., et al.: Sparse R-CNN: end-to-end object detection with learnable proposals. In: CVPR, pp. 14454–14463 (2021)
41. Tang, J., et al.: Few could be better than all: feature sampling and grouping for scene text detection. In: CVPR, pp. 4563–4572 (2022)

42. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: fully convolutional one-stage object detection. In: ICCV, pp. 9627–9636 (2019)
43. Tian, Z., Shu, M., Lyu, P., Li, R., Zhou, C., Shen, X., Jia, J.: Learning shape-aware embedding for scene text detection. In: CVPR, pp. 4234–4243 (2019)
44. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) NIPS 2017. LNCS, vol. 30, pp. 5998–6008. Curran Associates, Inc. (2017). https://doi.org/10.5555/3295222.3295349
45. Vu, T., Kang, H., Yoo, C.D.: SCNet: training inference sample consistency for instance segmentation. In: AAAI, pp. 2701–2709 (2021)
46. Wang, F., Chen, Y., Wu, F., Li, X.: TextRay: contour-based geometric modeling for arbitrary-shaped scene text detection. In: ACM MM, pp. 111–119 (2020)
47. Wang, W., Xie, E., Li, X., Hou, W., Lu, T., Yu, G., Shao, S.: Shape robust text detection with progressive scale expansion network. In: CVPR, pp. 9336–9345 (2019)
48. Wang, W., et al.: Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In: ICCV, pp. 8440–8449 (2019)
49. Wang, X., Kong, T., Shen, C., Jiang, Y., Li, L.: SOLO: segmenting objects by locations. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12363, pp. 649–665. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58523-5_38
50. Wang, X., Zhang, R., Kong, T., Li, L., Shen, C.: SOLOv2: dynamic and fast instance segmentation. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) NIPS 2020. LNCS, vol. 33, pp. 17721–17732. Curran Associates Inc, Red Hook, NY, USA (2020)
51. Ye, M., Zhang, J., Zhao, S., Liu, J., Du, B., Tao, D.: DPText-DETR: towards better scene text detection with dynamic points in transformer. In: AAAI (2023)
52. Yuliang, L., Lianwen, J., Shuaitao, Z., Sheng, Z.: Detecting curve text in the wild: new dataset and new solution. arXiv preprint arXiv:1712.02170 (2017)
53. Zhang, P., et al.: VSR: a unified framework for document layout analysis combining vision, semantics and relations. In: Lladós, J., Lopresti, D., Uchida, S. (eds.) ICDAR 2021. LNCS, vol. 12821, pp. 115–130. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86549-8_8
54. Zhou, X., et al.: East: an efficient and accurate scene text detector. In: CVPR (2017)
55. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: more deformable, better results. In: CVPR, pp. 9308–9316 (2019)
56. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: ICLR (2021)
57. Zhu, Y., Chen, J., Liang, L., Kuang, Z., Jin, L., Zhang, W.: Fourier contour embedding for arbitrary-shaped text detection. In: CVPR, pp. 3123–3131 (2021)