

TextDCT: Arbitrary-Shaped Text Detection via Discrete Cosine Transform Mask

Yuchen Su, Zhiwen Shao, Yong Zhou, Fanrong Meng, Hancheng Zhu, Bing Liu, and Rui Yao

Abstract—Arbitrary-shaped scene text detection is a challenging task due to the variety of text changes in font, size, color, and orientation. Most existing regression based methods resort to regress the masks or contour points of text regions to model the text instances. However, regressing the complete masks requires high training complexity, and contour points are not sufficient to capture the details of highly curved texts. To tackle the above limitations, we propose a novel light-weight anchor-free text detection framework called TextDCT, which adopts the discrete cosine transform (DCT) to encode the text masks as compact vectors. Further, considering the imbalanced number of training samples among pyramid layers, we only employ a single-level head for top-down prediction. To model the multi-scale texts in a single-level head, we introduce a novel positive sampling strategy by treating the shrunk text region as positive samples, and design a feature awareness module (FAM) for spatial-awareness and scale-awareness by fusing rich contextual information and focusing on more significant features. Moreover, we propose a segmented non-maximum suppression (S-NMS) method that can filter low-quality mask regressions. Extensive experiments are conducted on four challenging datasets, which demonstrate our TextDCT obtains competitive performance on both accuracy and efficiency. Specifically, TextDCT achieves F-measure of 85.1 at 17.2 frames per second (FPS) and F-measure of 84.9 at 15.1 FPS for CTW1500 and Total-Text datasets, respectively.

Index Terms—Arbitrary-shaped scene text detection, discrete cosine transform, single-level head, segmented non-maximum suppression.

I. INTRODUCTION

OVER the past few years, scene text detection has attracted much attention in the computer vision community, due to its applications in many fields like scene understanding [2], product search [3], and autonomous driving [4]. Benefited from the development of deep learning, scene text detection has made great progress [5]–[11]. However, due to unconstrained text variations in font, size, color, and

Manuscript received January, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62106268, and in part by the High-Level Talent Program for Innovation and Entrepreneurship (ShuangChuang Doctor) of Jiangsu Province under Grant JSSCBS20211220. It was also supported in part by the National Natural Science Foundation of China under Grants 62101555 and 62172417, in part by the Natural Science Foundation of Jiangsu Province under Grants BK20201346 and BK20210488, in part by the Six Talent Peaks Project in Jiangsu Province under Grant 2015-DZXX-010, and in part by the Fundamental Research Funds for the Central Universities under Grant 2021QN1072. (Yuchen Su and Zhiwen Shao contributed equally to this work. Corresponding author: Zhiwen Shao.)

Y. Su, Z. Shao, Y. Zhou, F. Meng, H. Zhu, B. Liu, and R. Yao are with the School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China, and also with the Engineering Research Center of Mine Digitization, Ministry of Education of the People's Republic of China, Xuzhou 221116, China (e-mail: {yuchen_su; zhiwen_shao; yzhou; mengfr; zhuhancheng; liubing; ruiyao}@cumt.edu.cn).

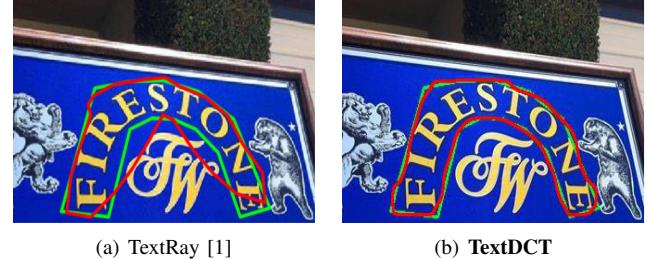


Fig. 1. Comparison of the text detection results of an example image for TextRay [1] and our TextDCT. Ground-truth contours are represented in green, and the detection results are represented in red. It can be seen that TextRay fails to model highly curved texts.

orientation, arbitrary-shaped scene text detection remains a challenge.

Current scene text detection methods based on deep learning can be divided into two categories: regression based approaches [12]–[14] and segmentation based approaches [15]–[19]. Due to the prediction for each pixel, segmentation based approaches do not need to explicitly process complex curved texts. However, such approaches are sensitive to noises, so they usually depend on the pre-training on a large dataset. Besides, pixel-level processing significantly increases the computational cost and the post-processing steps are typically very complicated. In contrast, regression based methods are often more concise and are easier to train. However, there are still two main problems unresolved for regression based methods.

On one hand, designing a compact text mask representation that can fit diverse geometry variances of arbitrary-shaped text instances is challenging. Because of the high complexity of directly regressing arbitrary-shaped text masks, most of the existing regression based methods regress contour point sequences of texts. However, point sequences are not sufficient to capture the details of highly curved texts, in which the represented text contour is usually unsatisfactory, as shown in Fig. 1(a).

On the other hand, state-of-the-art regression based methods rely heavily on the divide-and-conquer strategy in feature pyramid networks (FPN) [20] to regress multi-scale texts. However, all training samples are subject to the same supervision, leading to an imbalanced supervision issue among different pyramid layers, especially for single-stage detectors. Specifically, the number of training samples in P3 layer is 256 times of that in P7 layer of FPN. Thus, connecting multi-level heads together will cause extremely imbalanced learning among samples on different layers, as the shallow layers like P3 receive much more supervision than deeper layers.

To tackle these two problems, we propose a novel arbitrary-shaped scene text detection framework, namely TextDCT. First, inspired by the recent instance segmentation work [21], we model the high-resolution text instance mask in the frequency domain instead of the spatial domain via discrete cosine transform (DCT), and keep its low-frequency components as the mask representation, which has low training complexity. Due to the energy concentration characteristics of DCT, in which most of the natural signal energy is concentrated in the low-frequency components, so the mask representation by this transformation has a high quality.

Moreover, although recent research efforts [22], [23] can mitigate the imbalanced supervision problem by independent supervision on different pyramid layers, they do not go beyond the divide-and-conquer, which makes its structure more complex in single-stage detectors. Intuitively, using only a single-level head can solve this problem and the model is much simpler. However, directly replacing multi-level heads with a single-level head can lead to a dramatic model performance drop. The first reason is that the single-level head is not scale-aware and spatial-aware, in which the receptive field of single-level feature cannot match large range of text scales. Since arbitrary-shaped texts usually appear in vastly different fonts, rotations, and shapes, the detection performance for multi-scale text instances using a single-level head is limited. The second one is that a positive sampling strategy suitable for single-level prediction is required. If using the popular “center sampling” positive sampling strategy [24], the masks of different text instances whose center points are in the same region are easily confused.

Considering the above issues, we design a feature awareness module (FAM) to achieve spatial-awareness and scale-awareness by fusing rich contextual information, capturing larger receptive field and focusing on more significant features. We also introduce a text kernel sampling (TKS) strategy for the single-level prediction by treating the shrunk text kernel region as the positive samples, which can adaptively adjust the number of positive samples to balance text regression at different scales. Besides, based on the separation of text kernels from each other, we propose segmented non-maximum suppression (S-NMS) to effectively suppress false positives, especially for long text instances.

Our TextDCT is a single-shot, anchor-free, and light-weighted framework that performs joint optimization of text/non-text kernel classification, text bounding box regression, and text mask regression. The main contributions of this work are summarized as follows:

- We propose a novel scene text detection framework by employing DCT to represent text masks, which can accurately approximate arbitrary-shaped text instances while having low training complexity.
- We design a single-level head for top-down text prediction, with a feature awareness module, a text kernel sampling strategy, and a segmented non-maximum suppression method. These strategies are beneficial for avoiding imbalanced supervision, processing multi-scale text variations, and suppressing false positives.

- Extensive experiments demonstrate that our TextDCT achieves competitive performance on both accuracy and efficiency. Particularly, TextDCT obtains F-measure of 85.1 at 17.2 frames per second (FPS) and F-measure of 84.9 at 15.1 FPS for CTW1500 and Total-Text datasets, respectively.

II. RELATED WORK

Current scene text detection methods can be divided into two categories: segmented based scene text detection and regression based scene text detection. Besides, we discuss some current false positive suppressing methods.

A. Segmentation Based Scene Text Detection

Inspired by semantic segmentation methods [25], [26], some works regard the arbitrary-shaped text detection as a segmentation problem, which represents complex text instances with pixel-level prediction and rebuilds text instances through specific post-processing. Pixellink [27] first predicted the linkage relationship between pixels, then extracted the text bounding boxes by separating the links belonging to different text instances. TextSnake [28] treated text instances as a sequence of overlapping disks, and predicted text area, text center line, and some geometric attributes of disks to rebuild the text instances. To effectively split the close text instances, PSENet [15] detected different scale kernels in each text instance, and adopted a progressive scaling method to gradually expand the text kernels to obtain the final detections. Besides, TextField [29] first learnt a direction field containing both text mask and relative position information away from text boundary, then linked neighbor pixels to generate candidate text instances. Tian *et al.* [30] assumed each text instance as a cluster and performed a two-step clustering strategy to segment dense text instances, where pixels of the same text usually lie in the same cluster. CRAFT [31] detected character-level text regions by using two-dimensional Gaussian segmentation labels and exploring affinity between characters. DBNet [17] introduced a differentiable binarization processing module that gives a high threshold for text boundaries to distinguish adjacent texts. Moreover, TextBPN [32] first obtained boundary proposals based on the distance field map and classification map, then deformed the boundary proposals into more accurate text boundaries by the adaptive boundary deformation module.

Most of these methods can adapt to curved texts, but are sensitive to text-like background noises and holes, resulting in false positive cases. In contrast, our TextDCT can deal with false positives well by the S-NMS strategy.

B. Regression Based Scene Text Detection

Regression based methods typically rely on the object detection pipeline with bounding box regression, which is usually easier to train than segmentation based methods. CTPN [33] first used a modified pipeline of Faster R-CNN [34] to detect a set of partial text components with a fixed-size width, then connected them within different instances. RRPN [35] also used a modified pipeline of Faster R-CNN, which adopted

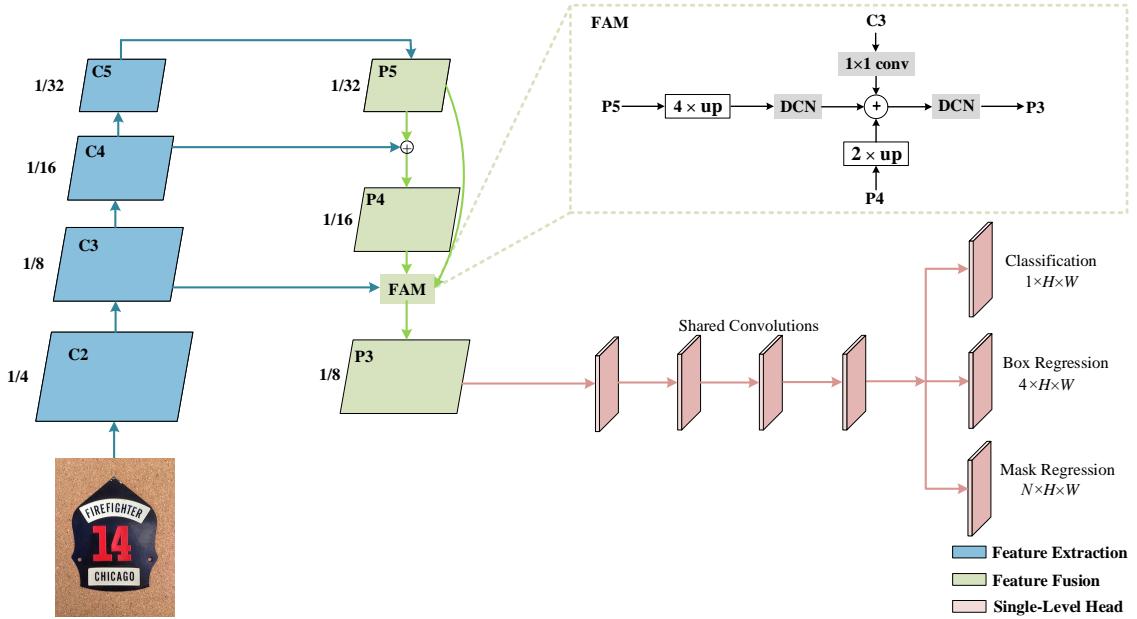


Fig. 2. The overview of our TextDCT, which is composed of three modules: (a) feature extraction as the backbone, (b) feature fusion via feature pyramid with FAM, and (c) single-level head for three-branch joint optimization. The classification branch and the box regression branch are used for predicting text kernels and text boxes, and the mask regression branch is used for predicting the mask vectors transformed by DCT. Here “ $1/k$ ” denotes $1/k$ of the input image size, DCN denotes deformable convolution, “+” denotes element-wise sum of feature maps, and “ $k \times$ up” denotes up-scaling the feature map with a factor k .

rotated proposals to detect multi-oriented texts. TextBoxes [36] modified the shapes of convolutional filters and increased the proportion of default boxes of SSD [37] to adapt to the aspect ratio of texts. TextBoxes++ [5] extended TextBoxes by applying quadrilateral regression to effectively detect multi-oriented text. EAST [38] adopted an anchor-free single-stage detection framework to directly regress the offsets between points within text instances and the corresponding bounding boxes or four corner points. MOST [39] proposed a text feature alignment module (TFAM) and a position-aware non-maximum suppression (PA-NMS) module to achieve accurate detection of long texts. However, the above regression based text representations are horizontal or oriented rectangles, which have limited capacity to model irregular texts.

Recently, several works [1], [13], [14], [40]–[43] focused on irregular text detection. CTD-TLOC [40] regressed the offsets between the top-left point of the bounding box and the key points on text contours, then smoothed the offsets with a recurrent neural network (RNN). LOMO [13] introduced a shape representation module that uses center lines, text regions, and border offsets to represent texts, then proposed an iterative refinement module to regress long texts. DRRG [14] treated text instances as a series of combinations of small rectangular components and introduced a graph convolutional network to learn the linkage relationships of the text components. TextRay [1] formulated the text contours in the polar system, and regressed the distance from the polar coordinate to the point where the emitted N-rays intersect with the text boundary. ABCNet [41] adopted an anchor-free network to predict a series of points, then used Bernstein polynomial to transform these points into Bezier curves that fit the text contours. PCR [42] proposed the contour location mechanism

(CLM) to regress text contours progressively. FCENet [43] treated text contours as periodic functions, and used the discrete Fourier transform (DFT) to convert text contours to Fourier eigenvectors. Unlike FCENet using DFT to encode text contours, our method adopts DCT to encode text masks as compact vectors.

C. False Positive Suppression

To suppress false positives, SPCNet [19] designed a text context module (TCM) and a re-score mechanism to improve the accuracy of the scores of tilted texts. ContourNet [44] designed a local orthogonal texture-aware module (LOTM), which considers the local texture information in two orthogonal directions simultaneously. TextRay [1] designed a central-weighted training strategy to give beneficial gradients for long text instances. ABCNet [41] used the center-ness branch to get better text bounding boxes. TextFuseNet [45] fused global-level, char-level and word-level features of texts to strengthen the capability of distinguishing texts and non-texts. PCR [42] designed a contour localization mechanism to re-score the localized contours. However, these methods increase the training complexity or the post-processing time, which brings additional computational burden. In this paper, we propose the S-NMS to suppress false positives in long texts with low overhead.

III. ARBITRARY-SHAPED TEXT DETECTION VIA DISCRETE COSINE TRANSFORM MASK

A. Overview

Our TextDCT is a single-shot, anchor-free detection framework that can fit arbitrary-shaped texts. As shown in Fig. 2,

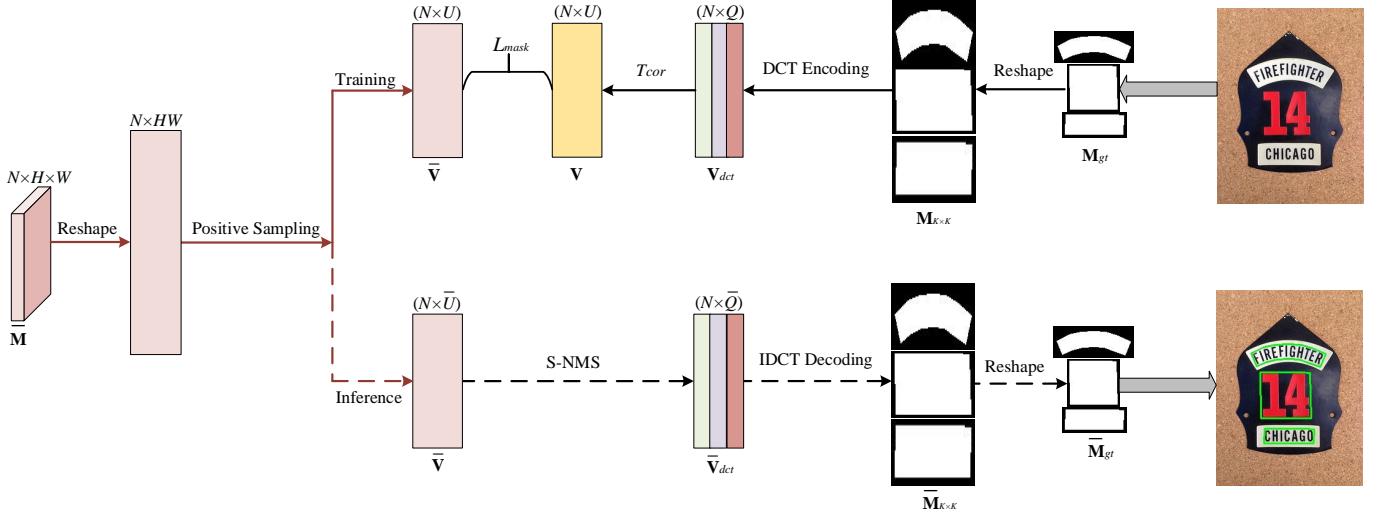


Fig. 3. The learning of DCT mask representation. $\bar{\mathbf{M}}$ is the output of the mask regression branch in Fig. 2, and $\bar{\mathbf{V}}$ are predicted mask vectors corresponding to the positive sample points in $\bar{\mathbf{M}}$. L_{mask} is the loss for mask regression task. T_{cor} refers to the one-to-one correspondence between the frequency-domain ground-truth and each positive sample point. Both $M_{K \times K}$ and $\bar{M}_{K \times K}$ have a uniform size, which are beneficial for DCT encoding and IDCT decoding. M_{gt} and M_{gt} denote the predicted text masks and the ground-truth text masks, respectively. In the shown example image, the number of text instances is $Q = 3$.

this framework mainly contains three parts: feature extraction, feature fusion, and three-branch joint optimization based on the single-level head. In the feature extraction module, we utilize ResNet50 [46] as the backbone to generate shared feature maps with different receptive fields. We first adopt FPN [20] to generate fused features with strong representation, and then add an efficient feature awareness module (FAM) between P4 and P3 of FPN to achieve spatial-awareness and scale-awareness for the single-level head. In the three-branch joint optimization module, we share the first four convolutions to improve the correlation between the classification and regression tasks, and then use three convolutions to predict the probability of each position corresponding to the text kernel, the regression value of each position corresponding to the text bounding box, and the text mask, respectively.

B. DCT Mask Representation

Most of the existing contour points modeling methods have a limited capacity to represent highly curved or long texts, while the directly regressed high-resolution masks contain redundant information since the discriminative pixels are mainly distributed along the text boundaries. To obtain low-complexity and high-quality text mask representations, we encode the text masks into compact vectors by DCT. As shown in Fig. 3, the output $\bar{\mathbf{M}} \in \mathbb{R}^{N \times H \times W}$ of the mask regression branch in Fig. 2 is first reshaped to be the size of $N \times HW$, where H and W are the height and width of $\bar{\mathbf{M}}$, and N denotes the vector dimension at each point in $\bar{\mathbf{M}}$. Then, we regress only the mask vectors $\bar{\mathbf{V}} \in \mathbb{R}^{N \times U}$ corresponding to the positive samples in $\bar{\mathbf{M}}$, where U is the number of positive sample points.

During training, the generation process of the ground-truth mask vector corresponding to each point consists of three steps. Firstly, we reshape the ground-truth mask M_{gt} of the

j -th text instance to be $M_{K \times K}^j \in \mathbb{R}^{K \times K}$, where $K \times K$ is the uniform mask size, $j = 1, \dots, Q$, and Q is the number of text instances in the input image.

Secondly, we employ two-dimensional DCT to encode $M_{K \times K}^j$ into the frequency domain:

$$\begin{aligned} M_{dct}^j(u, v) &= \frac{2}{K} C(u) C(v) \left[\sum_{x=0}^{K-1} \sum_{y=0}^{K-1} M_{K \times K}^j(x, y) \cos \frac{(2x+1)u\pi}{2K} \cos \frac{(2y+1)v\pi}{2K} \right], \end{aligned} \quad (1)$$

where $C(w) = 1/\sqrt{2}$ for $w = 0$, and $C(w) = 1$ otherwise. Due to the energy concentration characteristics of DCT, in which most of the natural signal energy is concentrated in the low-frequency components, the compact vector $V_{dct}^j \in \mathbb{R}^N$ can be sampled from the first N -dimensional vector of the M_{dct}^j in a “zigzag”.

Thirdly, we assign V_{dct}^j to $V^i \in \mathbb{R}^N$ by T_{cor} , where $i = 1, \dots, U$. T_{cor} is a label assignment strategy that matches the frequency-domain labels to corresponding positive sample points.

During inference, there are three steps for the prediction of text instance masks. Firstly, we employ a segmented non-maximum suppression (S-NMS) strategy to obtain \bar{V}_{dct} from the prediction $\bar{\mathbf{V}}$, which will be elaborated in Sec. III-D. Secondly, we expand \bar{V}_{dct}^j to K^2 -dimensions by backward complementing 0, and then transform \bar{V}_{dct}^j to $\bar{M}_{dct}^j \in \mathbb{R}^{K \times K}$ by reshaping the size of \bar{V}^j to $K \times K$. Thirdly, \bar{M}_{dct}^j can be generated by two-dimensional inverse discrete cosine transform (IDCT):

$$\begin{aligned} \bar{M}_{K \times K}^j(x, y) &= \frac{2}{K} \left[\sum_{u=0}^{K-1} \sum_{v=0}^{K-1} C(u) C(v) \right. \\ &\quad \left. \bar{M}_{dct}^j(u, v) \cos \frac{(2x+1)u\pi}{2K} \cos \frac{(2y+1)v\pi}{2K} \right]. \end{aligned} \quad (2)$$

Note that the computational cost of DCT and IDCT are negligible, since the complexity of computation through fast cosine transform (FCT) [47] is only $O(n \log n)$.

C. Single-Level Prediction

State-of-the-art regression based text detectors often adopt the divide-and-conquer strategy, which introduces an imbalanced supervision issue and causes a more complex structure in single-stage detectors. To suppress this issue, our TextDCT framework uses a single-level prediction structure. Two keys to the single-level prediction lie in how to design the input feature of the head so that the head is scale-aware and spatial-aware, and how to design a positive sample assignment strategy.

1) Feature Awareness Module: As shown in Fig. 2, we use the pyramid feature P3 as the input of the single-level head. Due to the diversity of text scale variations, the receptive field of P3 can only cover a limited scale range. As shown in Fig. 4(a), extreme long texts cannot be accurately regressed using P3 as the head's input, so we design a feature awareness module (FAM).

The structure of FAM is shown in Fig. 2, which consists of a skip connection and two deformable convolutions. Specifically, we first add a skip connection from P5 to enrich the contextual information of P3. Since direct element-wise sum of P3 and upsampled P5 leads to misaligned contexts of fused features, which may harm the prediction around text boundaries, we add a deformable convolution in the skip connection to learn offsets from spatial differences between P3 and unsampled P5, so as to obtain the aligned feature with contextual information. Besides, considering deformable convolution has the ability to focus on salient regions from various text appearances, we add a deformable convolution after the fused feature to capture more significant features. By obtaining features with rich contextual information and adaptively adjusting the receptive fields to achieve spatial-awareness and scale-awareness for the single-level head, our proposed FAM can suppress the limitations of the divide-and-conquer strategy.

2) Positive Sampling Strategy: The definition of positive samples is crucial for text detection, and balanced positive and negative samples are beneficial for achieving accurate classification results. Most existing anchor-free single-stage detectors adopt the "center sampling" strategy [24] to define positive samples. However, this strategy is not suitable for the case of single-level head, as potentially numerous ambiguous points harm the accuracy of regressing text shapes, as shown in Fig. 4(c). Besides, it is unreasonable to give the same number of positive samples to texts of different sizes. Therefore, we propose a text kernel sampling (TKS) strategy by treating the text kernel as the positive sample region, in which the text kernel can be obtained by Vatti clipping algorithm [48]. Since each point in the text kernel is located inside the text instance, there are few ambiguous samples, and the number of positive samples changes adaptively with the text scales. In this way, we can effectively alleviate the bias between texts with different scales during training. As shown in Fig. 4(d), our proposed TKS can generate more high-quality predictions.



Fig. 4. Illustration of the superiority of our proposed FAM and TKS. The comparison between (a) and (b) shows that adding FAM can better regress long texts, and the comparison of (c) and (d) shows that TKS is more effective for single-level prediction.



Fig. 5. Visualization of false positive suppression strategies. (a) The results of using NMS. (b) The results of using S-NMS. The red contour in (a) represents the false positive detection.

D. Segmented Non-Maximum Suppression

As shown in Fig. 2, a given image goes through our TextDCT, in which the classification scores predicted by the classification branch and the text boxes regressed by the box regression branch can be combined to obtain text box predictions. Typically, non-maximum suppression (NMS) is used to remove duplicated box predictions. However, the points near the text kernel edges of long text instances are generally far from the centers of the texts, making it hard to perceive the comprehensive text shape information accurately. Therefore, the regressed boxes corresponding to these points may be inaccurate and may have a low intersection over union (IoU) with the boxes corresponding to the points near the centers of the texts, which makes it difficult to be filtered by the NMS and thus leads to false positives, as illustrated in Fig. 5.

TextRay [1] and ABCNet [41] address this problem by using a center-weighted training strategy and adding a centerness branch, respectively, in which the training complexity is increased. Instead, we adopt a segmented non-maximum suppression (S-NMS) strategy. In particular, benefited from the strategy of using text kernels as positive samples, each text kernel can be easily distinguished. For each text kernel, we firstly select only the box corresponding to the point with the highest classification score, which allows us to filter out other boxes that may be of low quality directly without considering the IOU threshold. Then, we use NMS to filter the remaining potentially duplicate boxes.

Note that, as shown in Fig. 6(a), there are still duplicate

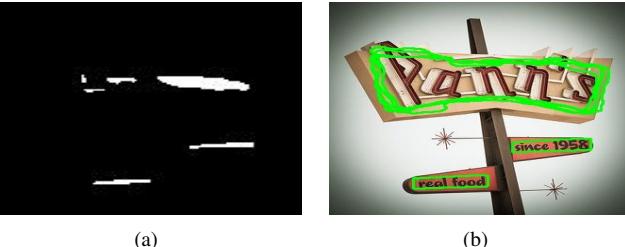


Fig. 6. Visualization of the effectiveness of the regression based methods. (a) The results of predicted text kernel. (b) The final detection results corresponding to the highest classification scores for all text kernels.

boxes after the first step of the S-NMS because the widespread holes in the text may cause the text kernel to be split into multiple parts. Due to the regression task prefers to focus on the edge parts, as shown in Fig. 6(b), where most of the points belonging to the same kernel but split into different parts can still regress to the same text instance.

E. Full Loss Function

In our TextDCT framework, the full loss function is formulated as

$$\mathcal{L} = L_{cls} + \lambda_1 L_{box} + \lambda_2 L_{mask}, \quad (3)$$

where λ_1 and λ_2 are the trade-off factors of the loss function. L_{cls} , L_{box} and L_{mask} denote classification loss, bounding box regression loss and mask vector regression loss of the three branches, respectively.

$$L_{cls} = 1 - \frac{2 \sum_i P_{ker}(i) G_{ker}(i)}{\sum_i P_{ker}(i) + \sum_i G_{ker}(i)}, \quad (4)$$

$$L_{box} = 1 - IOU + \frac{|C - (A \cup B)|}{|C|}, \quad (5)$$

where we choose Dice Loss to optimize L_{cls} , $G_{ker}(i)$ and $P_{ker}(i)$ denote the i -th pixel value in the ground-truth and the prediction of text kernels, respectively. Besides, we use GIoU loss for L_{box} following ABCNet [41], where A, B are ground truth and predicted bounding boxes, respectively. C is the smallest convex box enclosing both A and B and $IOU = |A \cap B| / |A \cup B|$.

Our mask vector regression loss is defined as

$$L_{mask} = \mathbb{1}^{text} \sum_i^N d_{mask}(\bar{v}_i, v_i), \quad (6)$$

where v_i, \bar{v}_i represent the i -th element in ground-truth and prediction mask vectors, respectively. $\mathbb{1}^{text}$ is the indicator function for text kernels. d_{mask} is the vector loss, in which we choose smooth-L1 loss for its effectiveness and stability in training.

F. Post-Processing

The post-processing of our TextDCT model consists of four steps. Firstly, the locations of the positive samples are obtained according to the classification threshold τ_a , and then their corresponding predicted text boxes and text mask

vectors are obtained. Secondly, the S-NMS is utilized to filter the overlapped text boxes. Thirdly, the text mask vectors corresponding to the remaining text boxes are transformed to 2D text masks by IDCT, and then the spatial resolutions of these text masks are resized to the same sizes of the spatial resolutions of their corresponding text boxes. Finally, these resized text masks are binarized by a threshold τ_b , and then are projected into the complete map with the same spatial resolution as the input image based on the locations of their corresponding text boxes.

IV. EXPERIMENTS

A. Datasets and Settings

1) **Datasets:** We evaluate our TextDCT on four popular text detection datasets: CTW1500, Total-Text, ICDAR2015, and MLT.

- **CTW1500** [40] is a challenging dataset that contains irregular-shaped and multi-oriented texts. It consists of 1,000 training images and 500 test images. Texts in this dataset predominantly suffer from blurring, low resolution and perspective distortion, and the text regions are all annotated by 14 key points.
- **Total-Text** [49] consists of 1,555 images (1,255 for training and 300 for testing). It contains horizontal, multi-oriented, and curved texts. Unlike CTW1500, all texts in Total-Text are annotated by world-level polygons with the adaptive number of vertices.
- **ICDAR2015** [50] consists of 1,000 natural images for training and 500 images for testing, including many multi-oriented and street-viewed text instances. The ground-truth of each text is annotated with eight coordinates to enclose the text in a clockwise way.
- **MLT** [51] is proposed on ICDAR 2017 Competition, which involves multi-lingual, multi-script and multi-oriented scene texts. It contains 9,000 images for training (7,200 training images and 1,800 validation images) and 9,000 images for testing.

2) **Implementation Details:** The backbone of our TextDCT framework is the pre-trained ResNet50 on ImageNet [52]. During training, stochastic gradient descent (SGD) is adopted as an optimizer with a batch size of 2 for the CTW1500 and Total-Text datasets and 4 for the MLT and ICDAR2015 datasets. The model is trained up to 120,000 iterations with the initial learning rate of 0.001, in which the learning rate is decreased to 0.0001 at the 90,000-th iteration. In addition, we employ a pre-trained model trained with 180,000 iterations on the SynthText [53] dataset. Following other methods [17], [41], [54], the text regions labeled as “DO NOT CARE” are ignored during training. Our TextDCT is trained on an NVIDIA Tesla V100 GPU. It takes about 10 training hours on CTW1500 and Total-Text datasets, and takes about 22 training hours on ICDAR2015 and MLT datasets.

We set K to 64, N to 300, τ_a to 0.9, τ_b to 0.35, and the shrinking rate of the text kernels to 0.5 for all datasets. Besides, we use the multi-scale training strategy, in which the short side of the input images is set to 640, 672, 704, 736, 768, 800, 832, 864, and 896, while the long side is

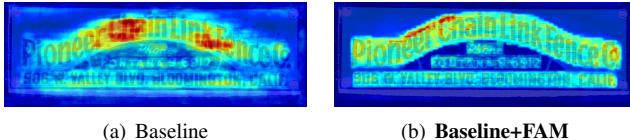


Fig. 7. Visualization of the feature heatmaps of an example image for Baseline and Baseline+FAM. For better viewing, we take channel-wise maximum value at each pixel of P3 to obtain the feature heatmap, and overlay the feature heatmap on the example image.

TABLE I
PERFORMANCE OF USING DIFFERENT PYRAMID LAYERS AS HEAD'S INPUT ON CTW1500.

Head's input	Recall	Precision	F-measure
P5	65.5	78.7	71.5
P4	67.8	78.0	72.6
P3 (Baseline)	70.7	77.6	74.0
Muti-level	77.4	81.2	79.2

maintained to 1,024 on CTW1500 and Total-Text datasets, and is maintained to 1,600 on ICDAR2015 and MLT datasets. In Eq. (3), $\lambda_1 = 1$, and $\lambda_2 = 1$. For data augmentation, we perform random horizontal flip, random crop, color jitter and contrast jitter for input images. In random cropping, we only crop the non-text regions, in which the crop area is larger than half of the original image area. During inference, the input images are resized as $1,056 \times 1,440$, $1,024 \times 2,000$, $1,500 \times 2,600$, and $1,600 \times 2,000$ for CTW1500, Total-Text, ICDAR2015, and MLT, respectively. In the following sections, we omit % for simplicity in the recall (R), precision (P), and F-measure (F) results.

B. Ablation Study

In this section, we conduct ablation studies on both Total-Text and CTW1500 datasets to validate main modules in our TextDCT.

1) *Input of the head*: Table I shows the results of a baseline model using different pyramid feature layers on CTW1500 dataset. The baseline model is TextDCT without FAM and S-NMS, in which the first four convolutions of the head are not shared, and the positive sampling strategy is “center sampling” [24]. Compared to multi-level head inputs, using only a single-level head has a significant performance drop, especially for the recall rate. This is mainly because the single-level head without FAM is not scale-aware and spatial-aware, and the “center sampling” strategy is not suitable for single-level prediction. Although the performance of using multi-level heads is higher than that of single-level head, the imbalanced supervision problem limits its performance. Besides, the F-measure of P3 is 1.4 and 2.5 higher than that of P4 and P5, respectively. It demonstrates that P3 makes a good trade-off between semantic information and boundary information compared to P4 and P5.

2) *FAM and TKS*: Table II shows the results of using FAM and TKS over the baseline model on CTW1500 and Total-Text datasets. We can see that the baseline model with FAM achieves the gain of 5.5 on CTW1500 in terms of F-measure. In Fig. 7, the baseline model with FAM has high responses

TABLE II
PERFORMANCE GAINS OF FAM AND TKS ON CURVE TEXTS.

Dataset	Method	Recall	Precision	F-measure
CTW1500	Baseline	70.7	77.6	74.0
	Baseline+FAM	78.0	81.1	79.5
	Baseline+TKS	76.8	80.1	78.4
	Baseline+FAM+TKS	80.8	82.6	81.7
Total-Text	Baseline	73.7	79.9	76.7
	Baseline+FAM	78.0	82.2	80.1
	Baseline+TKS	77.7	81.9	79.7
	Baseline+FAM+TKS	80.3	84.0	82.1

TABLE III
PERFORMANCE GAINS OF DIFFERENT COMPONENTS IN THE FAM ON CTW1500. SC, DCN1, DCN2 DENOTE THE SKIP CONNECTION OF P5 TO P3 AND THE TWO DEFORMABLE CONVOLUTIONS IN THE FAM, RESPECTIVELY. DCN3 DENOTES REPLACING THE 1×1 CONVOLUTION IN FAM WITH A DEFORMABLE CONVOLUTION

SC	DCN1	DCN2	DCN3	R	P	F
-	-	-	-	76.6	80.6	78.6
✓	-	-	-	76.3	81.4	78.8
✓	✓	-	-	80.1	82.9	81.5
-	-	✓	-	80.4	84.0	82.2
-	-	✓	✓	79.8	84.6	82.2
✓	✓	✓	-	81.5	84.7	83.1

on all the text regions, while the baseline model highlights on a few non-text regions and misses some text regions. This demonstrates that FAM is beneficial for accurately perceiving the overall layout of the texts with different sizes and shapes. Besides, using TKS also significantly improves the model performance, with recall and F-measure improved by 6.1 and 4.4 on CTW1500, respectively. When using both FAM and TKS, the performance is further improved on both CTW1500 and Total-Text datasets.

3) *Components of FAM*: Here we investigate the effectiveness of each component in FAM. FAM consists of a skip connection from P5 to P3 and two deformable convolutions. As shown in Table III, compared to the model of TextDCT without FAM, further adding a skip connection from P5 to P3 almost has no effect on the performance. However, the combination of skip connection and DCN1 improves recall and F-measure by 3.8 and 2.7, respectively, which shows that the features with rich contextual information aligned by deformable convolution are helpful for detecting multi-scale texts. Besides, if only adding DCN2 after P3, the F-measure can reach 82.2, which shows that adaptively adjusting the receptive field to focus on important regions can significantly improve the performance of the single-level prediction model. We also implement a variant by using both DCN2 and DCN3, in which DCN3 denotes replacing the 1×1 convolution in FAM with a deformable convolution. We can observe that the use of DCN3 almost brings no impact on the performance. Therefore, the performance gain by FAM is mainly due to the rich aligned contextual information and the adaptively adjusted receptive field, instead of the deformable convolution itself. When SC, DCN1, and DCN2 are combined together, our TextDCT achieves the best performance.

TABLE IV

ABLATION STUDY OF THE RESOLUTION OF MASK REPRESENTATION AND THE DIMENSION OF MASK VECTOR ON CTW1500. “*” MEANS WITHOUT DCT.

Resolution	Dim	Recall	Precision	F-measure	IOU
32×32	1,024*	79.7	85.0	82.3	95.6
32×32	300	80.0	84.6	82.3	95.0
64×64	4,096*	77.0	84.3	80.5	97.8
64×64	100	81.0	83.6	82.3	94.2
64 × 64	300	81.5	84.7	83.1	97.4
64×64	500	81.0	85.3	83.1	97.6
128×128	300	80.5	85.6	83.0	97.4

TABLE V

PERFORMANCE OF SHARING THE FIRST FOUR CONVOLUTIONS IN THE HEAD ON CTW1500. “GAP” REFERS TO THE ABSOLUTE VALUE OF THE DIFFERENCE BETWEEN THE CLASSIFICATION SCORE AND THE LOCATION SCORE.

Shared	R	P	F	gap
None	80.6	84.7	82.6	0.22
Regression(Box+Mask)	80.6	84.9	82.7	0.22
Classification+Regression	81.5	84.7	83.1	0.19

4) *DCT mask representation*: To investigate the effectiveness of the DCT mask representation, we implement different variants on CTW1500 dataset, as presented in Table IV. IOU refers to the intersection over union between reconstructed mask \tilde{M}_{gt} and ground-truth M_{gt} , which is used to evaluate the quality of the mask representation. Without the DCT mask representation, a high dimension makes the model difficult to optimize, leading to the performance degradation. On the other hand, a low resolution leads to significant reconstruction errors, which makes the model difficult to fit complex text shapes accurately.

Firstly, by comparing the 4,096-dimensional mask vector without DCT and the 300-dimensional DCT vector under the same resolution of 64×64 , we can see that the DCT mask representation using lower training complexity can obtain higher recall, precision, and F-measure results. Secondly, we compare the effect of different resolutions for the 300-dimensional DCT mask vector. As the resolution increases from 32×32 to 64×64 , the IOU increases from 95.0 to 97.4, and the F-measure gains 0.8 improvement. When the resolution is further increased to 128×128 , the results are almost unchanged. Therefore, we set the resolution to 64×64 in our TextDCT. Finally, we compare different dimensions of the DCT vectors. When the dimension is increased from 300 to 500, the IOU only grows 0.2 and the F-measure remains unchanged at 83.1. The is due to the energy concentration characteristics of DCT, in which the former dimension is more important than the latter dimension for reconstructed masks.

5) *Shared head convolutions*: Here we investigate the effect of sharing head convolutions. The results on CTW1500 dataset are shown in Table V, in which “gap” refers to the absolute value of the difference between the classification score and the location score. The location score means the intersection over union between the predicted box and the ground-truth box. In Table V, we can see that only sharing the first four

TABLE VI

PERFORMANCE OF DIFFERENT NMS TYPES IN THE POST-PROCESSING ON CTW1500.

Type	Recall	Precision	F-measure	FPS
NMS	81.8	82.6	82.2	17.0
K-NMS	81.9	82.9	82.4	16.8
S-NMS	81.5	84.7	83.1	17.2

convolutions in the heads of the box regression and the mask regression branches has little effect on the performance. If further sharing the first four convolutions in the heads of both the classification and regression branches, the F-measure is improved from 82.6 to 83.1. This is mainly due to the 0.03 reduction of “gap”, resulting in a higher-quality regression to the positive sample with the highest classification score.

6) *S-NMS*: In Table VI, we filter the redundant boxes using NMS, S-NMS, and kernel-level NMS (K-NMS) in our TextDCT, respectively. K-NMS is a two-stage NMS, which first uses NMS for each text kernel corresponding to the regression box separately, and then uses NMS for the remaining boxes. Compared with NMS, K-NMS has a slightly different filtering order, resulting in slightly different final results (82.4 vs. 82.2 in terms of F-measure). Compared with NMS and KNMS, S-NMS can effectively improve the performance. Note that there may be adhesions between the predicted adjacent text kernels. Since S-NMS only selects the text box corresponding to the highest classification score for the adhesion region, the recall is decreased slightly. Although with 0.3 recall rate drops on the CTW1500, S-NMS exceeds NMS by 0.9 in terms of F-measure, in which F-measure gives a convincing measurement result by balancing precision and recall.

C. Comparison with State-of-the-Art Methods

For a fair comparison, we only evaluate our model on a single scale for all datasets. Furthermore, for the text spotting model [41], [64], [65], we only show the detection results without recognition branch.

1) *Evaluation on Long Curved Text Benchmark*: The comparison results on long curved text dataset CTW1500 are given in Table VII, from which we can see that TextDCT achieves competitive performance and speed. Without pre-training, TextDCT can still achieve competitive results with 81.5, 84.7, and 83.1 in recall, precision, and F-measure, respectively.

On one hand, TextDCT outperforms at least 0.6 than segmentation based methods [15], [17], [29], [32], [54] in F-measure, which are more sensitive to noise and require pre-training on additional datasets to achieve good results. For example, the F-measure of PSENet [15] after pre-training on the Synthtext [53] dataset is 4.2 higher than that without pre-training. Although the precision of DBNet [17] is 1.9 higher than TextDCT, some small texts may be filtered out in the post-processing of DBNet, so its recall rate is much lower than our TextDCT (80.2 vs. 85.3).

On the other hand, for regression based methods, EAST [38] only regresses quadrilaterals so it cannot adapt to curved texts

TABLE VII

COMPARISONS WITH STATE-OF-THE-ART WORKS ON CTW1500 AND TOTAL-TEXT. EXT DENOTES EXTRA TRAINING DATA, WHERE SYN, MLT, ART, AND MIXT DENOTE SYNTHTEXT [53], ICDAR2017-MLT [51], ICDAR-ART [55], AND A MIXED DATASET COMPOSED OF SYNTHCURVE [41], COCO-TEXT [56], AND ICDAR2019-MLT [57], RESPECTIVELY. † INDICATES EVALUATING THE PERFORMANCE OF TOTAL-TEXT WITH THE IOU@0.5 [40], AND THE DEFAULT EVALUATION MATRIX OF TOTAL-TEXT IS DETEVAL [49].

Method	Paper	CTW1500					Total-Text				
		Ext	Recall	Precision	F-measure	FPS	Ext	Recall	Precision	F-measure	
TextSnake [28]	ECCV'18	Syn	85.3	67.9	75.6	-	Syn	74.5	82.7	78.4	-
SegLink++ [58]	PR'19	Syn	79.8	82.8	81.3	-	Syn	80.9	82.1	81.5	-
TextField† [29]	TIP'19	Syn	79.8	83.0	81.4	6.0	Syn	79.9	81.2	80.6	6.0
MSR [12]	IJCAI'19	Syn	78.3	85.0	81.5	4.3	Syn	74.8	83.8	79.0	4.3
LOMO [13]	CVPR'19	Syn	69.6	89.2	78.4	4.4	Syn	75.7	86.6	81.6	4.4
PSENet-1s† [15]	CVPR'19	MLT	79.7	84.8	82.2	3.9	MLT	78.0	84.0	80.9	3.9
ATRR [59]	CVPR'19	-	80.2	80.1	80.1	10.0	-	76.2	80.9	78.5	-
CRAFT [31]	CVPR'19	Syn	81.1	86.0	83.5	-	Syn	79.9	87.6	83.6	-
PAN [54]	ICCV'19	Syn	81.2	86.4	83.7	39.8	Syn	81.0	89.3	85.0	39.6
Mask-TTD [60]	TIP'20	-	79.0	79.7	79.4	-	-	74.5	79.1	76.7	-
DBNNet [17]	AAAI'20	Syn	80.2	86.9	83.4	22.0	Syn	82.5	87.1	84.7	32.0
TextRay† [1]	MM'20	ArT	80.4	82.8	81.6	-	ArT	77.9	83.5	80.6	-
ABCNet† [41]	CVPR'20	MixT	78.5	84.4	81.4	-	MixT	81.3	87.9	84.5	-
ContourNet [44]	CVPR'20	-	84.1	83.7	83.9	4.5	-	83.9	86.9	85.4	3.8
DRRG [14]	CVPR'20	MLT	83.0	85.9	84.5	-	MLT	84.9	86.5	85.7	-
ReLaText [61]	PR'21	Syn	83.3	86.2	84.8	10.6	Syn	83.1	84.8	84.0	3.2
OPMP [7]	TMM'21	-	80.8	85.1	82.9	1.4	Syn	82.7	87.6	85.1	1.4
Dai <i>et al.</i> [9]	TMM'21	-	80.4	86.2	83.2	0.6	-	81.2	85.4	83.2	0.7
PCR [42]	CVPR'21	MLT	82.3	87.2	84.7	-	MLT	82.0	88.5	85.2	-
FCENet [43]	CVPR'21	-	83.4	87.6	85.5	-	-	82.5	89.3	85.8	-
TextBPN [32]	ICCV'21	Syn	81.4	87.8	84.5	12.1	Syn	84.6	90.2	87.3	12.6
TextDCT	Ours	-	81.5	84.7	83.1	17.3	-	80.5	85.8	83.0	15.2
TextDCT	Ours	Syn	85.3	85.0	85.1	17.2	Syn	82.7	87.2	84.9	15.1

TABLE VIII
COMPARISONS WITH STATE-OF-THE-ART WORKS ON ICDAR2015 AND MLT.

Method	Paper	ICDAR2015					MLT				
		Ext	Recall	Precision	F-measure	FPS	Ext	Recall	Precision	FPS	
RRPN [35]	TMM'18	-	73.0	82.0	77.0	-	-	55.5	71.2	62.4	-
Lyu <i>et al.</i> [62]	CVPR'18	Syn	70.7	94.1	80.7	3.6	Syn	55.6	83.8	66.8	5.7
SPCNet [19]	AAAI'19	MLT	85.8	88.7	87.2	-	Syn	73.4	76.9	70.0	-
ATRR [59]	CVPR'19	-	86.0	89.2	87.6	-	-	-	-	-	-
PSENet-1s [58]	CVPR'19	MLT	84.5	86.9	85.7	1.6	-	68.2	73.7	70.8	-
LOMO [13]	CVPR'19	Syn	83.5	91.3	87.2	3.4	Syn	60.6	78.8	68.5	-
CRAFT [31]	CVPR'19	Syn	84.3	89.8	86.9	8.6	Syn	68.2	80.6	73.9	8.6
DBNNet [17]	AAAI'20	Syn	83.2	91.8	87.3	12	Syn	67.9	83.1	74.7	19
DRRG [14]	CVPR'20	MLT	84.7	88.5	86.6	-	-	61.0	75.0	67.3	-
ContourNet [44]	CVPR'20	-	86.1	87.6	86.9	3.5	-	-	-	-	-
Boundary [10]	MM'20	Syn	82.2	88.1	85.0	-	-	-	-	-	-
MS-CAFA [6]	TMM'20	-	82.7	86.2	84.4	0.5	-	66.8	79.5	72.6	0.5
R-Net [8]	TMM'21	Syn	82.8	88.7	85.6	21.4	Syn	64.5	70.4	67.3	-
TextMountain [63]	PR'21	Syn	84.1	87.3	85.7	10.4	-	55.2	80.8	65.6	6.0
FCENet [43]	CVPR'21	-	82.6	90.1	86.2	-	-	-	-	-	-
MOST [39]	CVPR'21	Syn	87.3	89.1	88.2	10.0	Syn	72.0	82.0	76.7	10.1
TextDCT	Ours	-	83.7	86.9	85.3	7.6	-	68.0	79.0	73.3	11.6
TextDCT	Ours	Syn	84.8	88.9	86.8	7.5	Syn	67.7	83.9	74.9	11.8

(60.4 in F-measure). Although LOMO [13] strives to achieve accurate text region representation by iterative optimization module, TextDCT obtains more higher performance (85.1 vs. 80.8). Moreover, compared with FCENet [43], TextDCT is more capable of modeling extremely long texts, which will be explored in detail in Sec. IV-D.

2) *Evaluation on Curved Text Benchmark:* The comparison results on curved text dataset Total-Text are shown in Table VII, where we evaluate using the protocol in [49].

We can observe that TextDCT in a single testing scale achieves competitive performance and speed (84.9 in F-measure and 15.1 in FPS). As a single-stage regression based method, TextDCT significantly outperforms existing regression based methods [1], [13], [65]. As a two-stage method,

TextSpotter [65] performs end-to-end text detection and recognition by modifying the mask branch of Mask-RCNN [66], which has much lower detection capability than the single-stage TextDCT (84.4 vs. 78.5 in F-measure). Compared with SPCNet [19], which is a two-stage regression based model with a text context module (TCM) and a re-score mechanism to improve model performance, TextDCT outperforms it by 4.2 and 2.0 in recall and F-measure respectively. Besides, as a single-stage model with a single-level head, TextDCT has a much simpler pipeline than these two-stage models. CRAFT [31] additionally uses character-level annotations to supervise the learning process, which greatly increases the difficulty of annotating datasets. Compared to CRAFT, our TextDCT trains

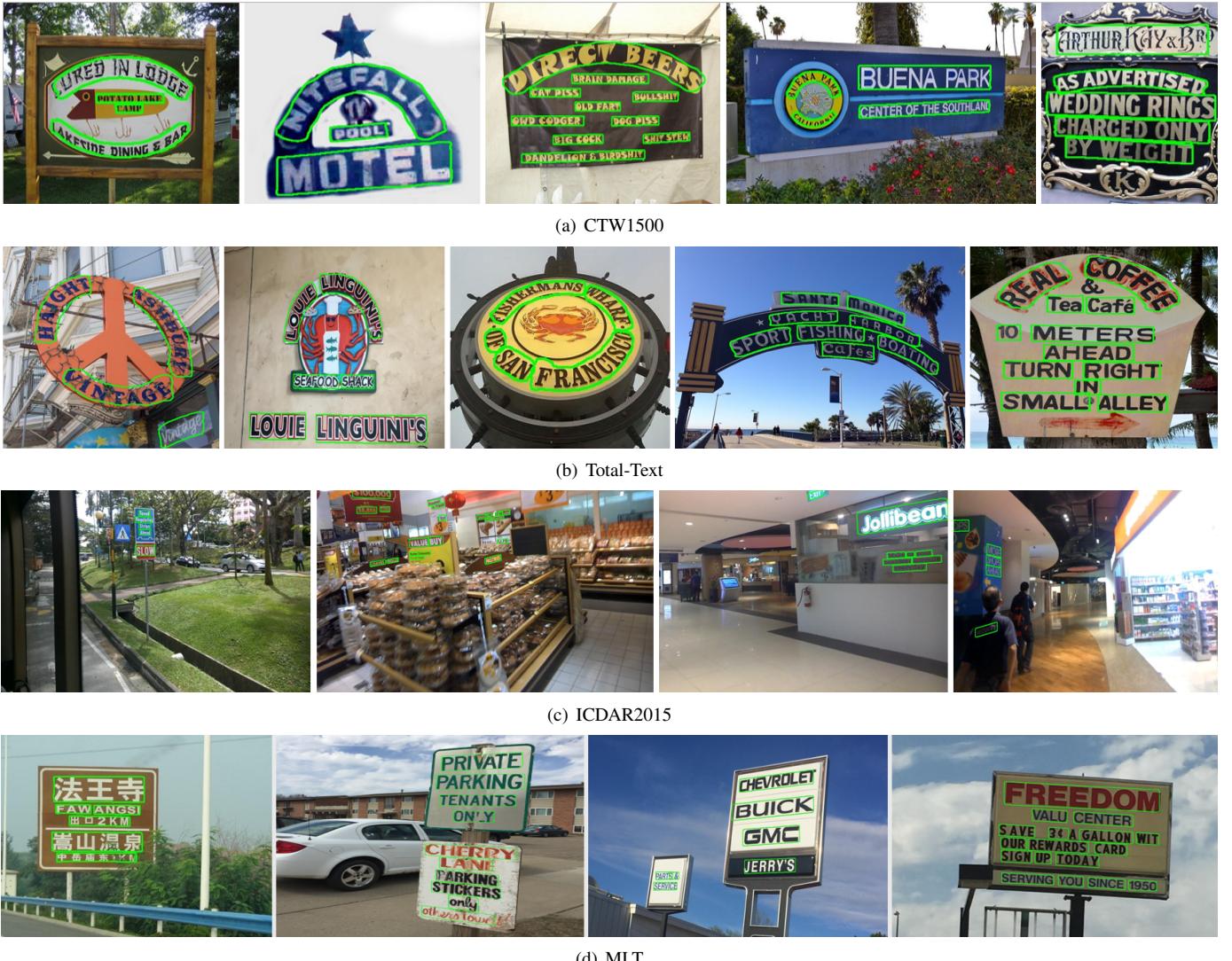


Fig. 8. Example results of our TextDCT on scene text datasets CTW1500, Total-Text, ICDAR2015, and MLT.

with only word-level supervision and shows advantages in F-measure.

Note that TextRay [1] models text contours in polar coordinates and represents text contours by a finite number of contour points, which may have limited ability to model texts with highly curved shapes (see Fig. 1). However, TextDCT can model extremely curved texts well with DCT mask representation and achieve better results. Compared with [6] improving the discrimination of text feature representations through the multi-scale context aware feature aggregation module, TextDCT achieves much better performance (84.9 vs. 81.5 in F-measure) by our single-level framework and DCT mask representation. Besides, compared with the segmentation based methods [28], [29] that include complex post-processing and are susceptible to false positives, our TextDCT uses a simple pipeline and suppresses false positives through S-NMS, outperforming these methods by a large margin.

Some qualitative results on Total-Text dataset are depicted in Fig. 8(b). As a light-weighted single-shot regression-based detector, our TextDCT achieves satisfactory results in various degrees of curvature, aspect ratios and complex backgrounds.

3) Evaluation on Oriented Text Benchmark: Test results on ICDAR2015 following the standard evaluation metric are shown in Table VIII. We can see that our TextDCT achieves 83.7, 86.9, and 85.3 for recall, precision, and F-measure respectively without any extra datasets, and achieves satisfactory results in terms of F-measure (86.8 in F-measure) when pre-trained on SynthText.

As shown in Table VIII, TextDCT surpasses most existing regression based methods [6], [8], [9], [38]. Specifically, compared with EAST [38], although both are single-stage models, our model can achieve spatial-awareness and scale-awareness through FAM, and has better detection capabilities for long texts in ICDAR2015 (86.8 vs. 80.7). R-Net [8] achieves spatial-awareness and scale-awareness by a spatial relationship module (SPM) and a scale relationship module (SRM). However, TextDCT can achieve better performance since more accurate contours can be obtained by mask regression branch and S-NMS eliminates some false positives. Compared with the two-stage anchor-based methods [6], [9], in which the latter contains a more complex network structure and sophis-



Fig. 9. Qualitative comparisons on challenging example images in CTW1500. The top and bottom rows show the results of FCENet* and our TextDCT, respectively.

TABLE IX
QUANTITATIVE COMPARISON OF F-MEASURE RESULTS BETWEEN
FCENET* AND TEXTDCT UNDER DIFFERENT IOU THRESHOLDS ON THE
CHALLENGING SUBSET OF CTW1500.

Method	IOU@0.5	IOU@0.6	IOU@0.7	IOU@0.8
FCENet*	80.9	77.0	69.1	46.8
TextDCT	80.6	77.2	69.6	47.8

ticated post-processing, and is sensitive to the anchor setting with more hyper-parameters, TextDCT achieves much better performance. Some example results on ICDAR2015 are shown in Fig. 8(c).

4) *Evaluation on Multi-Lingual Benchmark:* The comparison results on MLT dataset for evaluating the ability of multi-lingual text detection are presented in Table VIII. Our TextDCT surpasses most existing methods such as LOMO [13], R-Net [8], and TextMountain [63]. Besides, the example results on MLT are shown in Fig. 8(d), which shows that our TextDCT can effectively detect multi-lingual texts.

D. DCT Mask vs. DFT Contour

Recently, there is a work FCENet [43] which adopts DFT to encode text contours as compact vectors. To fairly compare our DCT mask representation with the DFT contour representation, we implement FCENet* by applying the idea of FCENet to our TextDCT framework. The only difference between FCENet* and TextDCT is that FCENet* uses a classification branch and a contour regression branch in the single-level head, in which the box regression branch is discarded.

Considering extremely long texts and extremely curved texts are the main challenges in scene text detection, we build a challenging subset with totally 347 samples of CTW1500, in which the text whose text mask area is less than half of the text box area or the longest edge of the text box is longer than three-quarters of the longest edge of the image are selected. Qualitative and quantitative comparisons were shown in Fig. 9 and Table. IX, respectively. Compared to our TextDCT, FCENet* tends to miss some corner pixels of long texts. Although the F-measure of FCENet* is 0.3 higher

TABLE X
RUNNING TIME OF OUR TEXTDCT.

Dataset	Input size	Network Inference	Post-processing
CTW1500	1,056 × 1,440	0.022s	0.036s
Total-Text	1,024 × 2,000	0.028s	0.038s
ICDAR-2015	1,500 × 2,600	0.054s	0.077s
MLT	1,600 × 2,000	0.034s	0.052s

than TextDCT under the evaluation protocol IOU@0.5, the F-measure of FCENet* drops more when the IOU threshold is increased from 0.5 to 0.8. The F-measure of TextDCT is 1.0 higher than that of FCENet* under the evaluation protocol IOU@0.8, mainly because TextDCT fits text contours more accurately than FCENet*.

E. Running Time Analysis

In the inference stage, the running time of our TextDCT includes network inference time and post-processing time, where post-processing mainly lies in S-NMS and IDCT. We set different input sizes for different datasets to achieve optimal model performance. The size of input images and the number of text instances has a great impact on the model running time, as shown in Table X. We compute the average running time of each image in each dataset on an NVIDIA Tesla V100 GPU. The results reported in Table X show that our TextDCT is able to detect arbitrary-shaped scene texts quickly.

F. Limitations

According to the above experimental results, our TextDCT can perform well in most challenging scenarios, accurately detecting highly curved texts. However, there are a few failure cases, as shown in Fig. 10. If the predicted text kernels stick to each other, it will cause omission because S-NMS only selects the point with the highest classification score for each kernel. Besides, some very text-like contexts are not filtered by our model, which should be alleviated by hard example mining.



Fig. 10. Failure cases. Red contours are ground-truths while green contours are predicted results.

V. CONCLUSION

In this paper, we have proposed a novel arbitrary-shaped scene text detector named TextDCT. The geometric encodings can be effectively learned through FAM and TKS, and false positives can be effectively removed by S-NMS. To obtain high-quality and low-complexity text instances, we have applied DCT to encode text instances as compact vectors. Extensive experiments have shown that the proposed single-level prediction framework can effectively detect arbitrary-shaped scene texts.

In the future work, the idea of mitigating the visual-semantic gap in [67] can be explored to suppress the false positives like the failure cases in Fig. 10. For example, we can extend TextDCT to an end-to-end scene text spotting framework, in which the text recognition module is trained to recognize texts as well as distinguish texts and text-like backgrounds. In this way, the detected text-like backgrounds can be removed based on the outputs of the text recognition module. Besides, our proposed idea of employing DCT to encode text masks as compact vectors is also promising to be applied for other object detection tasks.

REFERENCES

- [1] F. Wang, Y. Chen, F. Wu, and X. Li, "Textray: Contour-based geometric modeling for arbitrary-shaped scene text detection," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 111–119.
- [2] C. Yi and Y. Tian, "Scene text recognition in mobile applications by character descriptor and structure configuration," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 2972–2982, 2014.
- [3] B. Xiong and K. Grauman, "Text detection in stores using a repetition prior," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–9.
- [4] Z. Hong, Y. Petillot, D. Lane, Y. Miao, and S. Wang, "Textplace: Visual place recognition and topological localization through reading scene texts," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 2861–2870.
- [5] M. Liao, B. Shi, and X. Bai, "Textboxes++: A single-shot oriented scene text detector," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3676–3690, 2018.
- [6] P. Dai, H. Zhang, and X. Cao, "Deep multi-scale context aware feature aggregation for curved scene text detection," *IEEE Trans. Multimedia*, vol. 22, no. 8, pp. 1969–1984, 2019.
- [7] S. Zhang, Y. Liu, L. Jin, Z. Wei, and C. Shen, "Opmp: An omnidirectional pyramid mask proposal network for arbitrary-shape scene text detection," *IEEE Trans. Multimedia*, vol. 23, pp. 454–467, 2020.
- [8] Y. Wang, H. Xie, Z. Zha, Y. Tian, Z. Fu, and Y. Zhang, "R-net: A relationship network for efficient and accurate scene text detection," *IEEE Trans. Multimedia*, vol. 23, pp. 1316–1329, 2020.
- [9] P. Dai, Y. Li, H. Zhang, J. Li, and X. Cao, "Accurate scene text detection via scale-aware data augmentation and shape similarity constraint," *IEEE Trans. Multimedia*, pp. 1883 – 1895, 2021.
- [10] M. Xing, H. Xie, Q. Tan, S. Fang, Y. Wang, Z.-J. Zha, and Y. Zhang, "Boundary-aware arbitrary-shaped scene text detector with learnable embedding network," *IEEE Trans. Multimedia*, vol. 24, pp. 3129 – 3143, 2021.
- [11] P. Dai and X. Cao, "Comprehensive studies for arbitrary-shape scene text detection," *arXiv preprint arXiv:2107.11800*, 2021.
- [12] C. Xue, S. Lu, and W. Zhang, "Msr: Multi-scale shape regression for scene text detection," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 989–995.
- [13] C. Zhang, B. Liang, Z. Huang, M. En, and X. Ding, "Look more than once: An accurate detector for text of arbitrary shapes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10 552–10 561.
- [14] S. X. Zhang, X. Zhu, J. B. Hou, C. Liu, and X. C. Yin, "Deep relational reasoning graph network for arbitrary shape text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9696–9705.
- [15] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, and S. Shao, "Shape robust text detection with progressive scale expansion network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9336–9345.
- [16] Z. Tian, M. Shu, P. Lyu, R. Li, C. Zhou, X. Shen, and J. Jia, "Learning shape-aware embedding for scene text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4234–4243.
- [17] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-time scene text detection with differentiable binarization," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11 474–11 481.
- [18] J. Tang, Z. Yang, Y. Wang, Q. Zheng, and X. Bai, "Detecting dense and arbitrary-shaped scene text by instance-aware component grouping," *Pattern Recognit.*, vol. 96, p. 106954, 2019.
- [19] E. Xie, Y. Zang, S. Shao, G. Yu, C. Yao, and G. Li, "Scene text detection with supervised pyramid context network," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 9038–9045.
- [20] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.
- [21] X. Shen, J. Yang, C. Wei, B. Deng, J. Huang, X.-S. Hua, X. Cheng, and K. Liang, "Dct-mask: Discrete cosine transform mask representation for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8720–8729.
- [22] B. Wu, Y. Yang, D. Yang, J. Wu, H. Huang, L. Wang, J. Liu, and Y. Xu, "Progressive hard-case mining across pyramid levels in object detection," *arXiv preprint arXiv:2109.07217*, 2021.
- [23] Z. Chen, C. Yang, Q. Li, F. Zhao, Z.-J. Zha, and F. Wu, "Disentangle your dense object detector," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 4939–4948.
- [24] Z. Tian, C. Shen, H. Chen, and T. He, "Fcoss: A simple and strong anchor-free object detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1922 – 1933, 2020.
- [25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
- [26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.
- [27] D. Deng, H. Liu, X. Li, and D. Cai, "Pixelink: Detecting scene text via instance segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2018, p. 6773–6780.
- [28] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, "Textsnake: A flexible representation for detecting text of arbitrary shapes," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 20–36.
- [29] Y. Xu, Y. Wang, W. Zhou, Y. Wang, Z. Yang, and X. Bai, "Textfield: Learning a deep direction field for irregular scene text detection," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5566–5579, 2019.
- [30] Z. Tian, M. Shu, P. Lyu, R. Li, C. Zhou, X. Shen, and J. Jia, "Learning shape-aware embedding for scene text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4234–4243.
- [31] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9365–9374.
- [32] S.-X. Zhang, X. Zhu, C. Yang, H. Wang, and X.-C. Yin, "Adaptive boundary proposal network for arbitrary shape text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1305–1314.
- [33] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 56–72.
- [34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
- [35] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, 2018.

- [36] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4161–4167.
- [37] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
- [38] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "East: an efficient and accurate scene text detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5551–5560.
- [39] M. He, M. Liao, Z. Yang, H. Zhong, J. Tang, W. Cheng, C. Yao, Y. Wang, and X. Bai, "Most: A multi-oriented scene text detector with localization refinement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8813–8822.
- [40] Y. Liu, L. Jin, S. Zhang, C. Luo, and S. Zhang, "Curved scene text detection via transverse and longitudinal sequence connection," *Pattern Recognit.*, vol. 90, pp. 337–345, 2019.
- [41] Y. Liu, H. Chen, C. Shen, T. He, L. Jin, and L. Wang, "Abcnet: Real-time scene text spotting with adaptive bezier-curve network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9809–9818.
- [42] P. Dai, S. Zhang, H. Zhang, and X. Cao, "Progressive contour regression for arbitrary-shape scene text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7393–7402.
- [43] Y. Zhu, J. Chen, L. Liang, Z. Kuang, L. Jin, and W. Zhang, "Fourier contour embedding for arbitrary-shaped text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3123–3131.
- [44] Y. Wang, H. Xie, Z.-J. Zha, M. Xing, Z. Fu, and Y. Zhang, "Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11 753–11 762.
- [45] J. Ye, Z. Chen, J. Liu, and B. T. Du, "Scene text detection with richer fused features," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 7–15.
- [46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [47] M. Haque, "A two-dimensional fast cosine transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 6, pp. 1532–1539, 1985.
- [48] B. R. Vatti, "A generic solution to polygon clipping," *Commun. ACM*, vol. 35, no. 7, pp. 56–63, 1992.
- [49] C. K. Ch'ng and C. S. Chan, "Total-text: A comprehensive dataset for scene text detection and recognition," in *Proc. Int. Conf. Document Anal. Recognit.*, vol. 1, 2017, pp. 935–942.
- [50] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu *et al.*, "Icdar 2015 competition on robust reading," in *Proc. 13th Int. Conf. Document Anal. Recognit.*, 2015, pp. 1156–1160.
- [51] N. Nayef, F. Yin, I. Bizid, H. Choi, Y. Feng, D. Karatzas, Z. Luo, U. Pal, C. Rigaud, J. Chazalon *et al.*, "Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit.*, vol. 1, 2017, pp. 1454–1459.
- [52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
- [53] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2315–2324.
- [54] W. Wang, E. Xie, X. Song, Y. Zang, W. Wang, T. Lu, G. Yu, and C. Shen, "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 8440–8449.
- [55] C. K. Chng, Y. Liu, Y. Sun, C. C. Ng, C. Luo, Z. Ni, C. Fang, S. Zhang, J. Han, E. Ding *et al.*, "Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art," in *Proc. Int. Conf. Document Anal. Recognit.*, 2019, pp. 1571–1576.
- [56] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "Coco-text: Dataset and benchmark for text detection and recognition in natural images," *arXiv preprint arXiv:1601.07140*, 2016.
- [57] N. Nayef, Y. Patel, M. Busta, P. N. Chowdhury, D. Karatzas, W. Khelif, J. Matas, U. Pal, J.-C. Burie, C.-l. Liu *et al.*, "Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019," in *Proc. Int. Conf. Document Anal. Recognit.*, 2019, pp. 1582–1587.
- [58] J. Tang, Z. Yang, Y. Wang, Q. Zheng, Y. Xu, and X. Bai, "Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping," *Pattern Recognit.*, vol. 96, p. 106954, 2019.
- [59] X. Wang, Y. Jiang, Z. Luo, C.-L. Liu, H. Choi, and S. Kim, "Arbitrary shape scene text detection with adaptive text region representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6449–6458.
- [60] Y. Liu, L. Jin, and C. Fang, "Arbitrarily shaped scene text detection with a mask tightness text detector," *IEEE Trans. Image Process.*, vol. 29, pp. 2918–2930, 2019.
- [61] C. Ma, L. Sun, Z. Zhong, and Q. Huo, "Relatext: Exploiting visual relationships for arbitrary-shaped scene text detection with graph convolutional networks," *Pattern Recognit.*, vol. 111, p. 107684, 2021.
- [62] P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, "Multi-oriented scene text detection via corner localization and region segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7553–7563.
- [63] Y. Zhu and J. Du, "Textmountain: Accurate scene text detection via instance segmentation," *Pattern Recognit.*, vol. 110, p. 107336, 2021.
- [64] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "Fots: Fast oriented text spotting with a unified network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5676–5685.
- [65] P. Lyu, M. Liao, C. Yao, W. Wu, and X. Bai, "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 67–83.
- [66] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
- [67] C. Yan, Q. Zheng, X. Chang, M. Luo, C.-H. Yeh, and A. G. Hauptman, "Semantics-preserving graph propagation for zero-shot object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 8163–8176, 2020.

Yuchen Su is currently pursuing the M.S. degree at the School of Computer Science and Technology, China University of Mining and Technology, China, under the supervision of Prof. Yong Zhou, Prof. Fanrong Meng, and Dr. Zhiwen Shao. His current research interests include scene text detection and recognition.



Zhiwen Shao received his B.Eng. degree in Computer Science and Technology from the Northwestern Polytechnical University, China in 2015. He received the Ph.D. degree from the Shanghai Jiao Tong University, China in 2020. He is now a Tenure-Track Associate Professor at the School of Computer Science and Technology, China University of Mining and Technology, China. From 2017 to 2018, he was a joint Ph.D. student at the Multimedia and Interactive Computing Lab, Nanyang Technological University, Singapore. His research interests lie in computer vision and deep learning. He has been serving as a PC member in IJCAI and AAAI.



Yong Zhou received the M.S. and Ph.D. degrees in Control Theory and Control Engineering from the China University of Mining and Technology, China in 2003 and 2006, respectively. He is currently a Professor with the School of Computer Science and Technology, China University of Mining and Technology, China. His research interests include machine learning, intelligence optimization, and data mining.





Fanrong Meng received the Ph.D. degree from the China University of Mining and Technology, China. She is currently a Professor with the School of Computer Science and Technology, China University of Mining and Technology, China. Her research interests include intelligent information processing, database technology, and data mining.



Hancheng Zhu received the B.S. degree from the Changzhou Institute of Technology, Changzhou, China, in 2012, and the M.S. and Ph.D. degrees from the China University of Mining and Technology, Xuzhou, China, in 2015 and 2020, respectively. He is currently a Postdoctoral Fellow at the School of Computer Science and Technology, China University of Mining and Technology, China. His research interests include image aesthetics assessment and affective computing.



Bing Liu received the B.S., M.S., and Ph.D. degrees in 2002, 2005, and 2013, respectively, from the China University of Mining and Technology, Xuzhou, China. He is currently an Associate Professor at the School of Computer Science and Technology, China University of Mining and Technology, China. His current research interests include natural language processing, image understanding, and deep learning.



Rui Yao received the Ph.D. degree in computer science from the Northwestern Polytechnical University, Xi'an, China, in 2013. From 2011 to 2012, he was a Visiting Student with the University of Adelaide, Adelaide, SA, Australia. He is currently with the School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China. His research interests include computer vision and machine learning.