LRANet: Towards Accurate and Efficient Scene Text Detection with Low-Rank Approximation Network

Yuchen Su\textsuperscript{1,2}, Zhineng Chen\textsuperscript{1,*}, Zhiwen Shao\textsuperscript{3}, Yuning Du\textsuperscript{2}, Zhilong Ji\textsuperscript{4}, Jinfeng Bai\textsuperscript{4}, Yong Zhou\textsuperscript{3}, Yu-Gang Jiang\textsuperscript{1}

\textsuperscript{1} Shanghai Collaborative Innovation Center of Intelligent Visual Computing, School of Computer Science, Fudan University
\textsuperscript{2} Baidu Inc.
\textsuperscript{3} China University of Mining and Technology
\textsuperscript{4} Tomorrow Advancing Life

ycsu23@m.fudan.edu.cn, \{zhinchen, ygj\}@fudan.edu.cn, duyuning@biadu.com, jizhilong@tal.com, jfbai.bit@gmail.com, \{zhiwen, yzhou\}@cumt.edu.cn

Abstract

Recently, regression-based methods, which predict parameterized text shapes for text localization, have gained popularity in scene text detection. However, the existing parameterized text shape methods still have limitations in modeling arbitrary-shaped texts due to ignoring the utilization of text-specific shape information. Moreover, the time consumption of the entire pipeline has been largely overlooked, leading to a suboptimal overall inference speed. To address these issues, we first propose a novel parameterized text shape method based on low-rank approximation. Unlike other shape representation methods that employ data-irrelevant parameterization, our approach utilizes singular value decomposition and reconstructs the text shape using a few eigenvectors learned from labeled text contours. By exploring the shape correlation among different text contours, our method achieves consistency, compactness, simplicity, and robustness in shape representation. Next, we propose a dual assignment scheme for speed acceleration. It adopts a sparse assignment branch to accelerate the inference speed, and meanwhile, provides ample supervised signals for training through a dense assignment branch. Building upon these designs, we implement an accurate and efficient arbitrary-shaped text detector named LRANet. Extensive experiments are conducted on several challenging benchmarks, demonstrating the superior accuracy and efficiency of LRANet compared to state-of-the-art methods. Code is available at: https://github.com/ychensu/LRANet.git

Introduction

Scene text detection has garnered significant attention in computer vision due to its fundamental role as the initial step for end-to-end text recognition (Sheng, Chen, and Xu 2019; Du et al. 2022; Zheng et al. 2023a). Driven by the rapid development of deep learning, scene text detection methods have made significant progress (Song et al. 2022; Wang et al. 2022b; Fang et al. 2022; Song et al. 2022; Chen et al. 2022; Song et al. 2023; Qin et al. 2023; Zheng et al. 2023b; Du et al. 2023; Wang et al. 2023; Shu et al. 2023).

Existing scene text detection can be roughly divided into segmentation-based methods (Wang et al. 2019b; Zhou et al. 2020; Zhu and Du 2021; Liao et al. 2022) and regression-based methods (Sheng et al. 2019; Hou et al. 2020; Wang et al. 2020b; Zhang et al. 2022b; Su et al. 2022). To explore an arbitrary-shaped scene text detector, segmentation-based methods model text instances with pixel-level classification masks that allow naturally fitting of arbitrary shapes, but they mainly focus on local textual cues rather than the integrated geometric layout of texts, leading to a lack of global perception. Complex post-processing is often required to merge the pixel-level results to text regions.

Recently, regression-based methods, which predict parameterized text shapes for text localization, have sparked a new wave of research in arbitrary-shaped text detection due to their enhanced ability to consider the integrated geometric layout of texts. However, there are still two main problems that remain to be explored.

One problem is that the existing parameterized text shape methods are still challenging in modeling arbitrary-shaped texts. For horizontal and multi-oriented text, regressing quadrilaterals is sufficient to represent the text shape. However, sophisticated representation is required for representing arbitrary-shaped text. Some methods adopt contour points or parametric curves to fit the text shape, but they either lack imposing geometric constraints or fail to consider...
the distinct characteristics of text shapes. As a result, these methods may not faithfully represent the text boundaries, as shown in Fig. 2(a), (b) and (c). Scene text exhibits a wide range of shape diversity and aspect ratios. However, current parameterized text shape methods solely model text shapes individually using data-irrelevant decomposition, ignoring the structural relationships among different text shapes and the utilization of text-specific shape information. This makes it challenging to consistently and robustly represent various text shapes using only a few parameters.

The other problem is that regression-based methods often overlook the overall speed of the entire pipeline. Specifically, current regression-based methods can be divided into dense assignment approaches (Liu et al. 2020; Su et al. 2022; Wang et al. 2022a) and one-to-one assignment approaches (Raisi, Younes, and Zelek 2022; Ye et al. 2023; Shao et al. 2023) according to the allocation of positive samples. However, dense assignment approaches require non-maximum suppression (NMS) to filter a large number of redundant predictions, and this process is time-consuming, especially for arbitrary-shaped redundant predictions. Although one-to-one assignment approaches adopt the set prediction mechanism from DETR (Carion et al. 2020) to address this issue, they lack sufficient supervised signals and explicit position prior information. As a result, it usually stacks multiple decoders for iterative text contour optimization, leading to a complex pipeline.

To tackle the above-mentioned problems, firstly, inspired by the recent instance segmentation work (Park, Jin, and Kim 2022), we propose a low-rank approximation (LRA) method to better represent arbitrary-shaped texts. Unlike previous parameterized text shape methods that solely consider the individual text shape information, our LRA learns to represent text contours by exploring the shape correlation among different text instances. In detail, we first construct a text contour matrix, which contains all text contours in the training set. Then, we apply singular value decomposition (SVD) to extract the first $M$ left singular vectors of the contour matrix, referred to as eigenanchors. Based on the best $M$-approximation property of SVD, each text contour can be approximated as a linear combination of the eigenanchors, as illustrated in Fig. 3. LRA endows the generated shape representation with the following characteristics: compactness (fewer fitting parameters), consistency (fitting various text shapes tightly), simplicity (fitting with only linear operations). Moreover, LRA is robust, as the contour is approximated by a linear combination of the eigenanchors, and fitted curves using different numbers of eigenanchors are shown from left to right.

Next, we propose a dual assignment scheme for positive samples, aiming to accurately learn and efficiently infer the coefficients of eigenanchors. Specifically, we introduce a sparse assignment branch to reduce redundant predictions and thus decrease the processing time of NMS during inference, while employing a dense assignment branch to provide ample supervised signals.

Building upon these designs, we develop an accurate and efficient scene text detector named LRANet, whose performance advantages are shown in Fig. 1. The main contributions of this paper are summarized as follows:

- We propose a novel parameterized text shape method named LRA. It represents text shapes accurately by leveraging a linear combination of eigenvectors learned from arbitrary-shaped text contours.
- We introduce a dual assignment scheme by complementarily combining sparse and dense assignments. It improves the model’s inference efficiency while ensuring its accuracy.
- We propose a light-weighted regression-based text detector named LRANet, which can accurately and efficiently detect arbitrary-shaped texts. Extensive experiments show that LRANet achieves state-of-the-art performance.
Related Work

Segmentation-Based Methods

Segmentation-based methods regard the text detection as a segmentation problem, firstly modeling text instances with pixel-level classification masks, and then formed them into text boundaries through specific heuristic operations. PSENet (Wang et al. 2019a) predicts text instances with different scale kernels and applies a progressive scale expansion strategy to gradually expand the predefined text kernels. DB (Liao et al. 2020) and DB++ (Liao et al. 2022) introduce a differentiable binarization module that assigns a higher threshold to text boundaries, enabling the distinction between adjacent text instances. TextPMs (Zhang et al. 2022a) proposes an iterative model to predict a set of probability maps, and then utilizes region growth algorithms to group these probability maps into text instances.

Unfortunately, these segmentation-based methods model text instances from a local perspective, lacking global perception of texts, resulting in sensitivity to text-like background noise and computationally intensive post-processing.

Regression-Based Methods

Regression-based methods are mainly inspired by the advancement in object detection, where text shapes are represented as vectors through parameterization methods for regression. These methods primarily focus on irregular texts due to their complex representation. Earlier approaches directly regress contour points as text boundary, but they fail to utilize prior information about the continuity of text boundary. Therefore, later approaches employ parameterized curves or parameterized masks to represent the text boundary. For example, TextRay (Wang et al. 2020a) adopts Chebyshev polynomials under the polar system to approximate the text boundary. ABCNet (Liu et al. 2020) adopts Bernstein polynomial to convert the long sides of texts into Bezier curves. FCENet (Zhu et al. 2021) converts text contours to Fourier signature vectors by Fourier contour embedding. TextDCT (Su et al. 2022) transforms text instance masks to the Frequency domain by discrete cosine transform (DCT), and then extracts the low-frequency components to represent the text instance masks. TPSNet (Wang et al. 2022a) utilizes thin plate splines (TPS) to parameterize text contours into TPS fiducial points.

However, these parameterized methods have not sufficiently taken into account text-specific shape information, leading to limitations in modeling arbitrary-shaped texts. For example, chebyshev polynomials and DCT representation struggle to tightly fit the text shape with a compact vector. Fourier representation may lose some corner pixels, as shown in Fig. 2. Although TPS and Bezier can fit long curved texts well through fiducial points, slight perturbations in these points would significantly affect the fitted curve. Meanwhile, using a sparse number of fiducial points may not be sufficient to accurately represent an arbitrary-shaped text. Different from these efforts, we propose LRA that represents the fitted curve from an eigenvector-reconstruction perspective, allowing for effective exploitation of text-specific shape information.

Methodology

Low-Rank Approximation Representation

Low-rank approximation (LRA) is a widely-applied data dimensional reduction method. In this paper, we use LRA to compactly represent text contours. Unlike previous parameterized text shape methods that use curve fitting (Liu et al. 2020; Zhu et al. 2021; Wang et al. 2022a) or mask compression (Su et al. 2022), LRA is a data-driven approach that represents text boundaries in a low-dimensional space by exploiting the distribution of labeled text contours.

Scene text shapes are typically well-structured, mainly characterized by large aspect ratios and right-angle corners. As a result, there is significant correlation among these text shapes. By exploiting this correlation based on labeled data, we design a novel parameterized text shape method. Specifically, the ground-truth text boundary typically consists of multiple vertices, we first flatten them into a column vector \( p = [x_1, y_1, \ldots, x_N, y_N]^T \in \mathbb{R}^{2N \times 1} \), where \( N \) is the number of vertices. Then, we construct a text contour matrix \( A = [p_1, p_2, \ldots, p_L] \in \mathbb{R}^{2N \times L} \) from a corpus containing \( L \) labeled text instances. Thus, we apply singular value decomposition (SVD) to extract the structural relationship among the text contours in \( A \) to

\[
A = U \Sigma V^T.
\]

Here, \( U = [u_1, u_2, \ldots, u_{2N}] \in \mathbb{R}^{2N \times 2N} \) and \( V = [v_1, v_2, \ldots, v_L] \in \mathbb{R}^{L \times L} \) are orthogonal matrices, and \( \Sigma \in \mathbb{R}^{2N \times L} \) is a diagonal matrix composed of singular values \( \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0 \), where \( r \) denotes the rank of \( A \). Therefore, we can decompose \( A \) as

\[
A = \sigma_1 u_1 v_1^T + \cdots + \sigma_r u_r v_r^T,
\]

where \( u_i \) and \( v_i \) are left singular and right singular vectors corresponding to \( \sigma_i \), respectively.

After SVD, each singular vector represents an orthogonal direction in the latent space. By retaining only the first \( M \) singular values and their corresponding singular vectors, and truncating the low-rank components obtained from the decomposition in Eq. (2), matrix \( A \) can be approximated by

\[
A_M = \sigma_1 u_1 v_1^T + \cdots + \sigma_M u_M v_M^T = U_M \Sigma_M V_M^T
\]

where \( \hat{p}_i \) denotes the approximation of \( p_i \), and \( A_M \) is the best \( M \)-rank approximation of \( A \) as it minimizes the Frobenius norm \( ||A - A_M||_F \) (Blum, Hopcroft, and Kannan 2015).

Then, we define \( \Sigma_M V_M \) as matrix \( C_M = [c_1, c_2, \ldots, c_L] \in \mathbb{R}^{M \times L} \), allowing the matrix \( A_M \) to be represented as

\[
A_M = [U_M c_1, \ldots, U_M c_L] = [\hat{p}_1, \ldots, \hat{p}_L].
\]

In Eq. (4), each approximated text contour \( \hat{p}_i \) can be expressed as a linear combination of the first \( M \) left singular vectors \( u_1, u_2, \ldots, u_M \). In other words,

\[
\hat{p}_i = U_M c_i = [u_1, \ldots, u_M] c_i.
\]

We call these \( u_1, \ldots, u_M \) as eigenaunchors, as they are eigenvectors of matrix \( A A^T \) and can be viewed as pre-defined arbitrary-shaped anchors, as shown in Fig. 3. Worth
noting, eigenanchors are similar to the principal components in principal components analysis (PCA). However, PCA eliminates the mean value during matrix approximating, which may lead to information loss in our case.

We refer to the space spanned by the eigenanchors as eigenanchor space. For any 2N-dimensional text contour \( p \), we can project it onto the eigenanchor space to obtain an \( M \)-dimensional representation by

\[
c = U_M^T p,
\]

where \( c \) is a coefficient vector. Also, the approximation of \( p \) can be reconstructed by Eq. (5). Note that the number of contour vertices in \( p \) may not equal to \( N \). In that case, we resample \( N \) vertices from \( p \) using cubic spline interpolation.

**Dual Assignment Scheme**

We propose a dual assignment scheme for positive samples. It adopts a dense assignment branch to provide ample supervised signals for training, and meanwhile, accelerates the inference speed through a sparse assignment branch, enjoying the merits of both branches.

Specifically, following previous regression-based text detection network (Zhu et al. 2021; Wang et al. 2022a), our dense assignment scheme treats the text region as the positive sample region. In our sparse assignment scheme, we construct a prediction-aware matrix \( S \) for selecting \( K \) positive samples for each text instance. The matrix element is defined as

\[
s_{ij} = \begin{cases} 
FL'(b_i) + \lambda \sum_{n=0}^{N-1} \| \tilde{p}_{i}^{(n)} - \tilde{p}_{j}^{(n)} \|, & i \in TR \\
\infty, & i \notin TR
\end{cases},
\]

Here, \( s_{i,j} \) denotes the matching cost between the \( i \)-th point and the \( j \)-th ground-truth text instance, \( b_i \) is the predicted classification score of the \( i \)-th point. \( FL' \) is defined as the difference between the positive and negative terms: \( FL' = -\alpha(1 - x)^\gamma \log(x) + (1 - \alpha)x^\gamma \log(1 - x) \), which is derived from the focal loss (Lin et al. 2017b). We set \( \alpha \) to 0.25 and \( \gamma \) to 2.0. The second term is the L1 distance between the \( i \)-th predicted contour and \( j \)-th ground-truth contour; and \( \lambda \) controls the importance degree of classification and regression. The third item aims to limit the sparse sampling to only the text region for better joint optimization. Note that both the cost of classification and regression are considered in Eq. (7).

Afterwards, we regard the sparse positive sampling as a bipartite matching problem and use the Hungarian algorithm to solve the matrix \( S \) in ascending order, to find the optimal matching point for each text instance. To explore the optimal number of positive sample allocations for each text instance, we replicate it \( K - 1 \) times when constructing the matrix \( S \), and thus assign \( K \) positive samples to each instance.

In summary, the dense assignment branch assists in training the sparse assignment branch through providing ample supervised signals. During inference, we utilize sparse positive samples predicted by the sparse assignment scheme to accelerate post-processing.

**LRANet**

**Network Architecture** As shown in Fig. 4, we adopt a compact one-stage fully-convolutional architecture for effective and efficient detection. The architecture mainly comprises three parts: a feature extraction module, a detection head, and an LRA decoder for inference. Specifically, in the feature extraction module, we utilize ResNet50 (He et al.
In LRANet, the overall loss is formulated as

\[ \mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}}, \]

where \( \mathcal{L}_{\text{cls}} \) and \( \mathcal{L}_{\text{reg}} \) are the losses for the classification and regression branches, respectively.

The classification loss consists of the text region loss \( \mathcal{L}_{\text{tr}} \) and the sparse sampling region loss \( \mathcal{L}_{\text{ssr}} \):

\[ \mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{tr}} + \mathcal{L}_{\text{ssr}}, \]

where \( \mathcal{L}_{\text{tr}} \) and \( \mathcal{L}_{\text{ssr}} \) are the cross entropy loss and focal loss, respectively.

The regression loss is defined as

\[ \mathcal{L}_{\text{reg}} = \mathbb{1} da \sum_{i}^{N_{da}} l_1(\hat{p}_i, p_i), \]

where \( da \) is the positive sample region in our dual assignment scheme, \( \mathbb{1} \) is a spatial indicator, outputting 1 when the point is within \( da \) and 0 otherwise, and \( l_1 \) is smooth-L1 loss.

Experiments

Datasets

**CTW1500** is a challenging dataset for curve text detection. It consists of 1,000 training images and 500 test images. The text instances are annotated by 14-polygon.

**Total-Text** (Ch’ng and Chan 2017) consists of 1,255 training images and 300 test images, including horizontal, curve and multi-oriented texts.

**MSRA-TD500** (Yao et al. 2012) is a multi-language dataset. It consists of 3,000 training images and 200 test images. The texts instances are annotated with quadrilaterals.

**SynthText-150K** (Liu et al. 2020) is a synthetic dataset for arbitrary-shaped scene text, containing nearly 150k images that include curved and multi-oriented texts.

Implementation Details

We set the text scale ranges of P3, P4 and P5 in FPN to [0.3, 0.65], [0.25, 0.65], [0.55, 1] of the image size, respectively. The dimension \( M \) of eigenanchor space is set to 14. The sampling number \( K \) in the sparse assignment scheme is 3. The loss weight \( \lambda \) in Eq. (7) is 2. When training from scratch, we adopt stochastic gradient descent (SGD) to optimize our network with the weight decay of 0.0005 and the momentum of 0.9. For each dataset the training is conducted on its own training set. We set the training batch size to 4 and train 500 epochs, with the learning rate initialized at 0.001. For more comprehensive comparisons, we pre-train our model on SynthText-150K for 5 epochs, and then fine-tune 300 epochs for all datasets. The employed data augmentation includes random rotation, random scale, random flip and random crop.

In the testing stage, we set the short sides of the test image to 704, 1000, 800 for CTW1500, Total-Text and MSRA-

<table>
<thead>
<tr>
<th>Data-Driven</th>
<th>R</th>
<th>P</th>
<th>F</th>
<th>IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask</td>
<td>84.5</td>
<td>87.0</td>
<td>85.7</td>
<td>90.8</td>
</tr>
<tr>
<td>Side</td>
<td>84.6</td>
<td>88.6</td>
<td>86.6</td>
<td>97.9</td>
</tr>
<tr>
<td>Contour</td>
<td>84.9</td>
<td>89.1</td>
<td>86.9</td>
<td>98.0</td>
</tr>
</tbody>
</table>

Table 1: Experimental results on CTW1500 for different data representation. IoU refers to the intersection over union between reconstructed text region and ground-truth text region.
We compare dual assignment schemes best representation quality and detection performance. Consistent with this observation, the results in Table 1 indicate that the mask-driven method has poor representation quality, while the contour-driven method achieves the best representation quality more accurately, with the contour-driven method excelling at capturing complex text shape information. Different Data Representation The current parameterized text shape methods can be divided into three types: parameterized text contour (Zhu et al. 2021), parameterized text mask (Su et al. 2022), and parameterized top and bottom sides (Liu et al. 2020). To explore the influence of different data representations for LRA, we construct the matrix $A$ in Eq. (1) based on each of the above three types, and set $M$ to 14. Qualitative results are shown in Fig. 5. The mask-driven method contains redundant information, making it difficult to extract text shapes exactly. In comparison, the side-driven and contour-driven methods can extract text shape information more accurately, with the contour-driven method excelling at capturing complex text shape information. Consistent with this observation, the results in Table 1 indicate that the mask-driven method has poor representation quality, while the contour-driven method achieves the best representation quality and detection performance.

Effectiveness of Dual Assignment Scheme We compare the proposed dual assignment scheme with both sparse and dense assignments. To make a fair comparison with dense assignment, we adopt the dense matching strategy from (Wang et al. 2022a) as our baseline. As shown in Table 2, our dual assignment achieves a faster inference speed (22.1 vs. 14.6) due to the notable reduction in redundant prediction, while yielding a 0.5% improvement in F-measure compared to dense assignment. Moreover, as shown in Fig. 6, compared to using only sparse assignment, our dual assignment achieves faster convergence and higher accuracy due to receiving ample supervised signals. It attains a similar FPS to sparse assignment but a better F-measure, see Table 2.

Different Number of Sparse Sampling We further explore the influence of the sparse sampling number $K$ on model performance. As $K$ increases to 3, the F-measure is improved from 87.3% to 88.1%, and due to the small sparse sampling number, the impact of NMS on inference speed remains limited. However, as $K$ continues to increase, there is no further performance improvement, and the impact of NMS on inference speed is gradually obvious. This observation is consistent with the comparison results against dense assignment in Table 2, indicating that with sufficient supervision, a sparse positive sampling is sufficient.

Dimension of Eigenanchor Space We also conduct experiments to verify the influence of the eigenanchor space dimension, i.e., $M$. The results are listed in Table 3. At $M = 10$, the eigenanchor space can roughly represent the original space. As $M$ increases to 14, the IoU increases from 95.1 to 98.0, and the F-measure gains an 0.8% improvement. With the further increase of $M$, there is no significant improvement in the representation quality. The F-measure is almost unchanged. Therefore, we set $M = 14$ to balance training complexity and representation quality.

Different Input Image Sizes To demonstrate the trade-off between speed and accuracy, we evaluate our model with different length of short sides. The results are shown in Table 5, revealing an ascending F-measure trend as the short side extends, accompanied by a reduction in FPS.

Comparison with State-of-the-art Methods We compare LraNet with previous methods on Total-Text, CTW1500, and MSRA-TD500. The results are given in Table 7. LraNet consistently has top-tier performance across

<table>
<thead>
<tr>
<th>Method</th>
<th>Dim</th>
<th>IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chebyshev (Wang et al. 2020a)</td>
<td>44</td>
<td>83.6</td>
</tr>
<tr>
<td>DCT (Su et al. 2022)</td>
<td>32</td>
<td>88.5</td>
</tr>
<tr>
<td>Fourier (Zhu et al. 2021)</td>
<td>22</td>
<td>91.5</td>
</tr>
<tr>
<td>Bezier (Liu et al. 2020)</td>
<td>16</td>
<td>97.6</td>
</tr>
<tr>
<td>TPS (Wang et al. 2022a)</td>
<td>22</td>
<td>97.9</td>
</tr>
<tr>
<td>LRA</td>
<td>14</td>
<td>98.0</td>
</tr>
</tbody>
</table>

Table 6: Comparison with different parameterized text shape methods on CTW1500.
the three datasets. Compared to the segmentation-based methods, LRANet achieves competitive results even without pre-training, demonstrating its robustness to noise resistance. Compared to other parameterized text shape methods, LRA achieves superior results with fewer parameters, as shown in Table 6. The results reveal the significance of learning text-specific shape information, and our LRA effectively bridges this gap compared to existing solutions. It is observed that, when compared to the regression-based methods, LRANet shows advantages in terms of inference speed, while also maintaining leading performance. On one hand, the speed advantages come from the dual assignment. It provides sufficient supervised signals for model training, endowing the model with the capability of selecting high-quality positive samples at a single glance. On the other hand, the performance advantages come from LRA. For example, although TPSNet introduces a border alignment loss to alleviate the instability in its predicted TPS points, LRANet outperforms it by 1.0% and 1.5% in terms of F-measure on CTW1500 and Total-Text, respectively, even without intricate loss design. This is mainly attributed to the more advanced shape-relevant representation of LRA. Some detection visualizations are shown in Fig. 7. LRANet performs well on long, curve, and perspective texts.

**Conclusion**

In this paper, we present LRANet, an accurate and efficient arbitrary-shaped text detector. LRANet is featured by proposing an eigenvector-based reconstruction to effectively leverage the text-specific shape information. It accurately regresses the text contour with fewer parameters. Meanwhile, a dual assignment scheme that well integrates both dense and sparse assignments is developed for efficient inference. Experiments conducted on public benchmarks basically verify the proposed LRANet, where top-ranked performance and extremely fast inference speed are simultaneously observed. Given its effectiveness and efficiency, we are also interested in extending LRANet to text spotting in future.
Acknowledgements
The work was supported by National Key R&D Program of China (No. 2020AAA0104500) and the National Natural Science Foundation of China (No. 62172103).

References


