Hindawi
Mobile Information Systems
Volume 2022, Article ID 6378011, 13 pages
https://doi.org/10.1155/2022/6378011

Research Article

RWYI: Reading What You Are Interested in with a Learning-Based Text Interactive System

Zhenghong Yu,[1] **Hao Wang**,[2,3] **Hang Yang**,[2,3] **and Huabing Zhou**[2,3]

[1]*School of Robotics, Guangdong Polytechnic of Science and Technology, Zhuhai, China*
[2]*College of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, China*
[3]*Hubei Key Laboratory of Intelligent Robot (Wuhan Institute of Technology), Wuhan, China*

Correspondence should be addressed to Hao Wang; wangh@stu.wit.edu.cn

Received 10 June 2022; Revised 12 July 2022; Accepted 19 July 2022; Published 21 August 2022

Academic Editor: Praveen Kumar Reddy Maddikunta

As computer vision and human-computer interaction technology mature, vision-based auxiliary text reading has become the mainstream method to optimize the learning and reading experience. Most of the existing auxiliary text reading methods use scene text recognition combined with human gesture recognition to complete the task in multiple stages. However, these methods cannot accurately and effectively extract the textual information that readers are interested in complex and varied reading scenarios. To improve the text reading experience, we propose a human-centered fast auxiliary text reading method. It utilizes a hand-text hybrid object detection (HTD) model to instantly locate text of interest to readers, a font-consistent prior text image superresolution network (FCSRN) to recover low-resolution text images to enhance the accuracy of text recognition, and a convolutional recurrent neural network (CRNN) text recognition operator to obtain the content of the text, that is, interesting to readers. To verify the effectiveness of the proposed method, we tested the performance of the text localization module on a homemade HTD dataset and the performance of the FCSRN on the public text image superresolution dataset called TextZoom. Quantitative experiments on the overall performance of the fast auxiliary reading system, called reading what you are interested in (RWYI), were designed. The experiments indicate that the proposed method can meet the needs of human-computer interactive auxiliary reading in text reading scenarios and optimize the reading experience.

1. Introduction

The popularization of multimedia vision sensors has led to the development of various human-centered computer vision technologies, which have been gradually integrated into and changed our lives. Vision-based human-computer interaction tasks are mostly used in text reading comprehension [1], gesture interaction [2], human action recognition [3, 4], face detection [5, 6], and other fields. However, unfamiliar or forgotten words make the reading and learning experience negative for both children and adults.

To optimize the reading experience, applications that assist readers in reading are slowly becoming available to the public. The early reading aids based on optical character encoding can only work on printed books containing two-dimensional optical encoding, which directly lead to the failure of its popularization [7]. With the maturity of computer vision and human-computer interaction scientific research technology, vision-based auxiliary text reading has become the mainstream method to optimize the learning and reading experience. In general, vision-based auxiliary text reading consists of five modules: scene image input, image preprocessing, text-of-interest localization, text-of-interest recognition, and feedback output. To accurately locate the text of interest, usually, from the perspective of object detection, the finger distribution is first detected, and then the text of interest is located. Multistep detection is utilized, and engineering techniques are added to achieve a compromised result. The image quality of the positioning area will also interfere with the effect of content recognition in the scene text image [8]. This method consisting of

multiple stages of text detection and object detection needs to design an accurate position alignment strategy to achieve accurate text region localization, but a fixed alignment strategy cannot fully meet the complex and changeable text reading scenarios. Because the auxiliary text reading task has strict requirements on the processing speed and accuracy and the text reading scene itself is considerably complex, and includes the processing of low-quality images and the detection of small objects; these problems are all important to the auxiliary text reading task. The research presents great challenges.

In this study, a fast auxiliary text reading method is proposed to improve the efficiency of visual auxiliary text reading tasks from two perspectives: the rapid localization of interesting text and the effective enhancement of text image quality. On the one hand, in the stage of locating the text of interest, a one-stage target detection algorithm [9] is used to directly locate the opponent-text hybrid object, and the traditional multimodal combination of hand key point or finger distribution detection and local search text is not utilized. The proposed method significantly reduces the processing time. On the other hand, the low-quality text image obtained by locating the text first uses the text superresolution technique [10] to improve the quality of the image and then uses the obtained high-resolution image for text recognition [11]. Instead of using low-resolution images for text recognition directly, the accuracy of text recognition can be improved.

The main contributions of this paper are as follows:

(i) We propose a systematic method to quickly read an auxiliary text and can more quickly and accurately identify the text content of interest to readers.

(ii) A new method for locating interesting text using hand-text hybrid object detection (HTD) can efficiently locate the text of interest to readers.

(iii) An HTD dataset containing 12400 hand-text hybrid object images and annotations are used for training an HTD model.

(iv) A new superresolution network architecture for text images is proposed to improve the quality of text images so as to improve the accuracy of text recognition.

The remainder of the paper is organized as follows. In Section 2, we survey the recent works regarding vision-based human-computer interaction and auxiliary text reading tasks. Each component in the proposed system is described in Section 3. In Section 4, we report and discuss our experimental results that leads to the conclusions in Section 5.

2. Related Work

2.1. Vision-Based Human-Computer Interaction. Human-machine interaction (HMI [12]) or human-computer interaction (HCI) is the convergence of computer science, behavioral science, artificial intelligence, design, and other applied disciplines and involves the in-depth study of the scientific implications and practices of the interface between humans and computers. There are two lines of related research. At a superficial level, related research includes the research and design of new technologies to make computers more convenient tools for human life. At a deeper level, related research includes the study of intelligent technologies that use the natural interaction between humans and computers, thereby enabling computers to become more harmonious human partners. With the rapid development of fields such as artificial intelligence and deep learning, human-computer interaction technology has made great progress. Now, an increasing number of human-oriented human-computer interaction applications are appearing in our lives, which has promoted the formation of smart cities. Based on the design concept of HCI, it has become a research hotspot to allow machines to have perception capabilities such as vision [1] and hearing [13], to complete specific tasks. Table 1 shows the existing human-computer interaction technologies. In particular, in the scenario of human-computer nonverbal communication, vision-based human-computer interaction tasks require the establishment of communication channels that infer intentions from human behaviors, including facial expressions, human poses, and gestures [2]. Notably, the current implementation of these vision-based human-computer interaction tasks usually follows the process of image preprocessing, detection, and recognition, but the details of specific tasks are also different, and they depend on the data sets produced under specific functions to varying degrees. In fact, in the field of human-computer interaction, only visual or auditory-based interaction methods cannot fully meet the needs of human beings to disseminate and obtain information. Therefore, the multimodality of interactive information between humans and machines will be the trend of future research.

2.2. Auxiliary Text Reading. The earliest machine-auxiliary reading method used a reading pen to select a predetermined part of a supporting publication so that the optical tip of the reading pen recognizes the two-dimensional code printed in the publication, and then, the matching voice package could be played through the body circuit [7]. The technical areas that deep learning-based auxiliary text reading may cover are shown in Table 2. With the proposal of earlier two-stage object detection methods [22–24] and the birth of feature extraction backbone networks, such as feature pyramid networks (FPNs) [37] and PAN [38], later one-stage object detection methods [9, 25–28] have also been sequentially proposed and have provided a variety of possibilities for the realization of the localization of text of interest to a reader in the auxiliary text reading task. Even so, the recognized problem of small object detection in object detection tasks has not been effectively solved. In addition, due to the particularity of text objects in scene text detection and recognition tasks, based on the object detection method, a series of scene text detection methods [29–31] and text recognition methods [11, 32–36] are proposed. Although these methods cannot directly solve the problem of efficiently and accurately extracting textual information of

Table 1: Related to existing human-computer interaction technologies.

HCI technologies	Category	Ref.	Purpose
Vision-based HCI	Gesture interaction	[2]	Look at human actions and analyze human intentions.
	Human action recognition	[3, 4]	
	Face detection	[5, 6]	
Hearing-based HCI	Speech recognition	[13]	Listen to human language and analyze human intent

Table 2: Technical fields that may be covered by the auxiliary text reading task.

Technical fields	Features	Category	Ref.
Image preprocessing	Enhance images, improve image quality, and help downstream tasks achieve good results	Image warping	[14]
		Image fusion	[15, 16]
		Image superresolution	[10, 17–21]
Object detection	Locate target areas of text that are of interest to readers	Two-stage object detection	[22–24]
		One-stage object detection	[9, 25–28]
Scene text detection	Text region localization for complex scene images in real life	Text detection	[29–31]
Text recognition	Get detected text content of interest	Text recognition	[11, 32–36]

interest to readers, they provide the possibility for the auxiliary text reading task to achieve effective localization of the text of interest to the reader, which also enables the auxiliary text reading task to progress steadily. Lighten AI's vision-based artificial intelligence method realized finger-point reading. It can locate the text of interest by combining multiple object detection models and local search approaches and then uses the text recognition operator to identify the located text. Finally, the built-in voice package enables the model to read aloud and explain the meaning [8]. However, this is very demanding on the image acquisition equipment, as only this method can obtain extremely high-quality scene images to achieve good results. It is easy to overlook that image quality improvements [14–16] can also have beneficial effects on vision-based assisted text reading tasks. In particular, image superresolution techniques [17–21] can restore low-resolution images to high-resolution images to enhance image data quality and improve the effect of downstream tasks. TSRN [10] used the first real-world text-image superresolution dataset, TextZoom, and a baseline for text-image superresolution, which enables the reconstruction of low-quality text images, was proposed. Obviously, the text-image superresolution method, as an intermediate task of text detection and recognition, should be lightweight. When building a text-image superresolution network, it is necessary to balance the image quality improvement effect and the actual resource overhead.

3. Methodology

From a human-centered point of view and to improve a reader's text reading experience, we propose a fast auxiliary text reading method, which aims to obtain the text content of interest to a reader from the images acquired by a visual sensor in the text reading scene. Figure 1 illustrates the overall framework. In the first step, we use a single-stage hybrid object detection method to locate the target text area pointed at by a finger of the input image and obtain the text image with a lower quality than the original input image. In the second step, we perform superresolution processing on the detected low-quality text image to obtain the enhanced high-resolution text image. In the third step, we use high-quality text images for text recognition and use conventional text recognition operators to recognize the content of the text of interest pointed at by a finger. Therefore, the recognized text content can be recited and interpreted using devices such as speech, translation and word-finding devices, and used as feedback, which also takes full advantage of the interaction of visual and auditory information between humans and machines to help readers learn and understand the current text vocabulary.

3.1. Hybrid Object Detection. In the auxiliary reading task, we let the reader's behavior have a positive effect on the task; here, the reader is not just the receiver of the auxiliary reading task but can be thought of as the leader in the process of human-computer interaction. Using a priori knowledge that readers have a high probability of pointing at unfamiliar text with their fingers in auxiliary reading scenarios, we define the text pointed at by the reader's finger and the reader's finger as a hybrid object category, called "hand-text." The definition of this mixed class weakens the difficult problem of small object detection. We aim to make the machine learn the contextual feature information contained in the behavior of a finger pointing at the text in an image and not only consider the mixed features of the two types of objects of the fingertip and the text area but also consider a wider range of gesture features. Furthermore, we take the feature of this mixed object of "hand text" as the basis for locating the text region of interest (ROI) from the image. Therefore, our proposed hybrid object detection method still belongs to the category of object classification and localization tasks but has a different starting point from traditional object detection methods that only focus on the

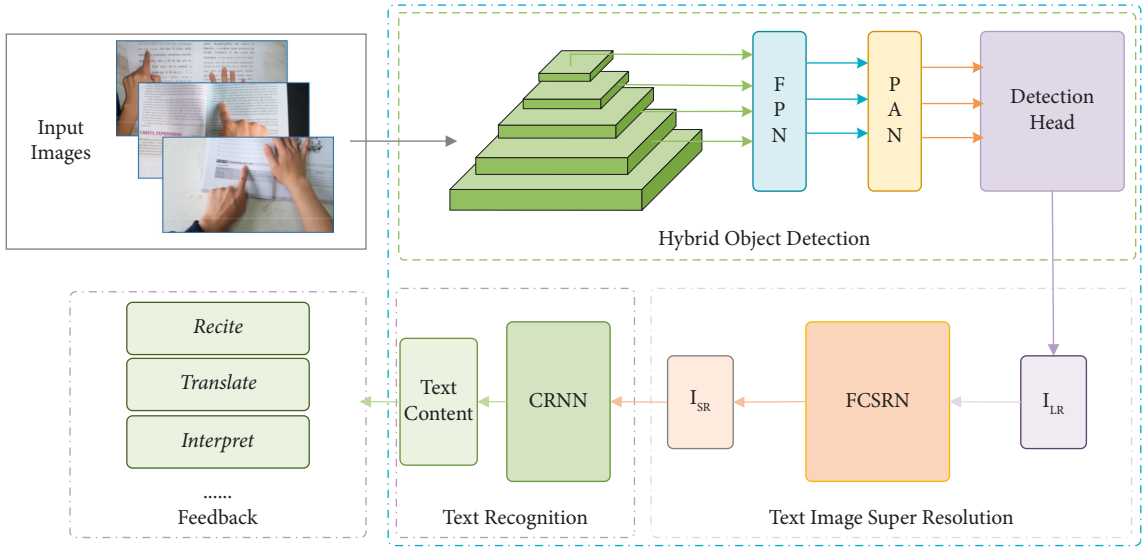

Figure 1: Overall framework of the fast auxiliary text reading method, called reading what you are interested in (RWYI).

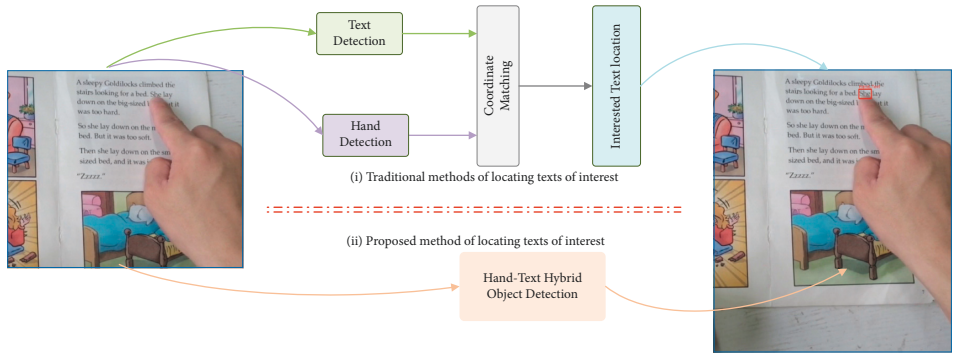

Figure 2: Comparison of text-of-interest localization methods.

feature information of a single object category. The difference between the proposed HTD method and the traditional localization methods for locating the text of interest is shown in Figure 2, and the proposed method does not have multiple stages of alignment tasks such as text detection and object detection, which simplifies the idea of locating textual information of interest to readers.

Since the proposed method is applied to the object localization task, it is natural to follow the principle of object localization task implementation. We utilize a convolutional neural network to extract the complex features of the hybrid object "hand text" in the image and use the localization of this hybrid object as the output of the prediction head of the hybrid object detection task. The key point of this task is whether the machine can learn the shallow texture features, shape features, and deep context features of the hybrid object category "hand text" from images in a reading scene through a single-stage or end-to-end object detection method to correctly generate positioning predictions. Thus far, we can see that our proposed fast auxiliary text reading method uses hybrid object detection to directly locate the text ROI instead of using multiple object detection steps to narrow the detection range to obtain text regions.

3.1.1. Hand-Text Hybrid Object Detection Dataset. To better achieve the task of HTD in the fast auxiliary text reading method, we prepared a HTD dataset from a text reading scene. It contains nearly 4,000 "hand-text" objects that have been marked and are considered "background" when a finger is not pointing at the text or when the text is not being pointed at by a finger, and only instances when a finger is pointing to the exact text area are marked as "foreground." The fingers pointing to the text in these images are those of different readers, and the pointed text is obtained from books with different fonts and font sizes. The lighting and background of the reading scenes are varied to fully ensure the diversity of the dataset. The sizes of the images in the HTD dataset are not the same to ensure that the machine text model learns the "hand-text" features in the reading scenes.

Data augmentation is very common in object detection tasks. The ultimate purpose is to enable the object detection model to learn more generalized expression capabilities with more diverse data and to accurately classify and locate objects in more complex environments. To make our proposed HTD method has a better detection effect, we use Gaussian perturbations, brightness changes, small-angle rotations, scaling, up and down

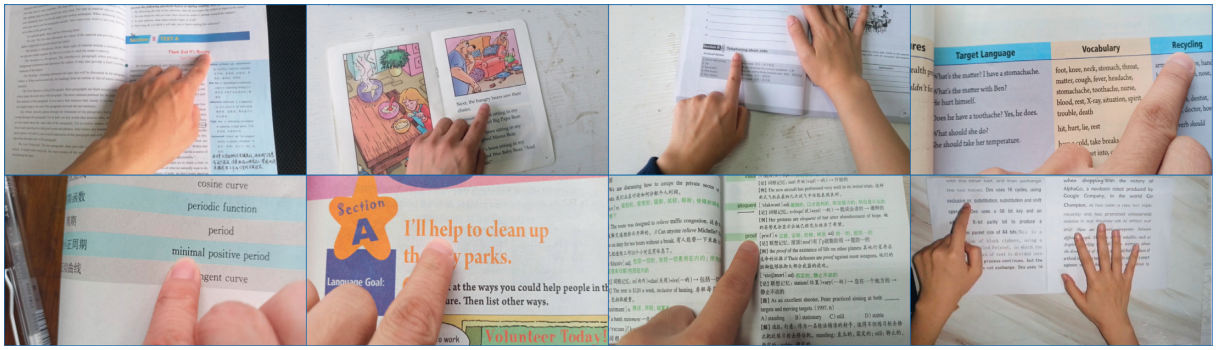

Figure 3: Samples from the prepared hand-written mixed object detection dataset.

flips, and so on, to augment the HTD dataset to train the hybrid object detection network. The image data augmentation method processes the original images and labels, resulting in an HTD dataset with four times the number of original images. Figure 3 shows some examples of HTD images.

3.1.2. Hybrid Object Detection Network. YOLO is one of the most representative models for object detection tasks. After continuous iterative development in recent years, a good balance between detection accuracy and real-time performance has been achieved by the YOLO-V5 [9] version. We use the latest YOLO-V5 as the basic model for hybrid object detection to discriminate and locate finger-pointing actions in text reading scenarios. We utilize YOLO-V5 for the detection of "hand-text" hybrid object categories in text reading scene images. In essence, "hand-text" hybrid object detection is an object localization task with a single predicted category in a special scene. Therefore, YOLO-V5 can obtain the ROI of the text pointed at by a finger with relatively high accuracy and a relatively small number of iterations with reasonable computational complexity.

The YOLO-V5 model can predict both the object class probability and its bounding box in an end-to-end manner. Therefore, we use many visual sensors in the text reading scene to obtain a video stream of a reader while reading, use the image frames in the video stream as the input of the hybrid object detection model, and use the YOLO-V5 model to locate "hand-text" hybrid objects and obtain the bounding box information. Since the YOLO-V5 model contains multiple artificially set anchor boxes, the nonmaximum suppression (NMS) strategy is essential. Based on the particularity of the "hand-text" hybrid object detection task, we can directly perform weighted NMS on the predicted bounding box without judging whether the class labels of the initially predicted bounding boxes are the same when implementing the weighted NMS strategy. Compared with traditional NMS, weighted NMS uses the process of bounding box culling, and those boxes whose intersection over union (IOU) is greater than the threshold and that are in the same category as the current bounding box are not directly culled but are based on the confidence of the network prediction and are weighted to obtain a new bounding box, which is used as the final predicted bounding box. Then, those boxes that are not the most suitable are eliminated. The formulation of weighted NMS is as follows:

$$\begin{aligned} \widehat{M} &= \frac{\sum_i \omega_i B_i}{\sum_i \omega_i}, \\ B_i &\in \{B | \mathrm{IOU}(M, B) \geq \mathrm{thres}\} \cup M, \\ \omega_i &= s_i \mathrm{IOU}(M, B_i), \end{aligned} \tag{1}$$

where B_i represents the initial prediction box generated by the model, s_i is the prediction confidence of the "hand-text" hybrid object category of the predicted bounding box B_i, M represents the bounding box to be calculated currently, $\widehat{M}$ is the weighted bounding box, and thres is the artificially specified confidence threshold.

Through the YOLO-V5 end-to-end object positioning model, we directly obtain the bounding box coordinates of the text area pointed at by a finger and naturally use the text area image as the output of the hybrid object detection task, and it is also the input of the next text-image superresolution processing task.

3.2. Text-Image Superresolution. In a text reading scene, the image of the text area of interest obtained by the common visual sensor after HTD is often not of high quality, and the effect of text recognition will decline, which will make it difficult for the auxiliary text reading task to achieve the expected effect. We attempt to reconstruct text images using text-image superresolution processing techniques. Generally, text-image superresolution processing is usually an upstream task of image text recognition and is used to improve the resolution of low-resolution text images and restore text-image details. We try to use text-image superresolution processing technology to reconstruct text images to prevent using low-quality images for text recognition so as to improve the accuracy of text recognition.

We exploit the prior that the text in the text ROI to the reader maintains the same font style in auxiliary text reading scenarios and embed this prior knowledge on the basis of

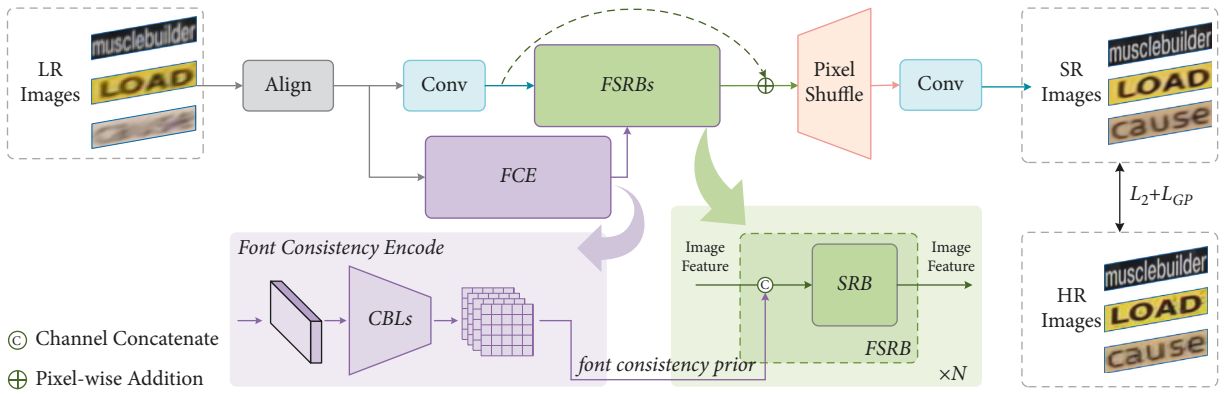

Figure 4: The overall architecture of our proposed FCSRN architecture.

previous text-image superresolution methods. We try to let the machine help restore low-quality images to high-quality images by learning the font style consistency of text images. Therefore, based on TSRN [10], we introduce a feed-forward convolution encoder (FCE) for extracting font-consistency prior information of text images, which is extracted into font-consistency prior information and embedded into multiple different text superresolution modules to improve the ability of the text-image superresolution network to construct high-quality text images. Figure 4 shows our proposed font-consistency priors for the text-image superresolution network (FCSRN) architecture.

Specifically, we take a low-resolution image $I_{\mathrm{LR}} \in \mathbb{R}^{\mathrm{h}\times\mathrm{w}\times 3}$ as input and first use the spatial transformation network (STN) [39] to align low-resolution input images of different sizes. This process can be expressed as follows:

$$\widehat{I}_{\mathrm{LR}} = \mathrm{STN}\left(I_{\mathrm{LR}}\right), \tag{2}$$

where $\mathrm{STN}(\cdot)$ represents the dimension alignment and $\widehat{I}_{\mathrm{LR}}$ represents the aligned low-resolution image. We perform two-way branch processing on the aligned low-resolution text images. On the one hand, the FCE branch is used to extract font-consistency prior features to obtain F_f, and on the other hand, the 9×9 convolution operation is used to extract shallow image features to obtain F_c. Then, the extracted font-consistency prior features are embedded into the image features at different network depths extracted by the backbone through a concise and efficient method of concatenating on the channel, and they are used as the input of the sequential residual block (SRB). These processes can be formulated as follows:

$$\begin{aligned} F_f &= \mathrm{FCE}\left(\widehat{I}_{\mathrm{LR}}\right), \\ F_c &= \mathrm{Conv}\left(\widehat{I}_{\mathrm{LR}}\right), \\ F_{sr}^{i} &= \mathrm{FSRB}^{i}\begin{cases} \mathrm{SRB}^{i}\left(\mathrm{Concat}\left(F_c, F_f\right)\right), & i = 1, \\ \mathrm{SRB}^{i}\left(\mathrm{Concat}\left(F_{sr}^{i-1}, F_f\right)\right), & 1 \in [2, N], \end{cases} \end{aligned} \tag{3}$$

where FSRB^{i} is the i th text-image superresolution module embedded with font-consistency priors, F_{sr}^{i} is the text-image superresolution feature obtained by the i th FSRB, and N is the number of FSRB stacks. It is worth mentioning that we implement FCE using a stack of CBL modules consisting of convolution, batch normalization, and Leaky-ReLU activation functions. To ensure the role of the deeper network, we follow the residual network structure between the stacked $N = 5$ SRBs. Finally, we use the pixel shuffle module to increase the resolution by a factor of 2 and use the convolution operation again to output the final superresolution text image.

For the loss calculation part of model training, we follow the method in the baseline task. Generally, the quality of text-image superresolution can be measured not only by image quality metrics but also by the downstream task of text recognition. We evaluate the performance of text-image superresolution methods based on font-consistency priors in Section 4.2.

3.3. Text Recognition. In the context of auxiliary text reading, the text images that we use to identify the text of interest to readers have the characteristics of regular text and large image ratios. Therefore, the text recognition operator, we use needs to have a good balance between the recognition accuracy of regular text and the computational complexity. In addition, for our proposed auxiliary text reading method to satisfy different audiences, we need a text recognition operator that can recognize multiple languages.

After weighing various factors, we choose to use a convolutional recurrent neural network (CRNN) [11], which is an early deep learning method for regular text recognition. The network architecture of CRNN is shown in Figure 5, which is composed of the convolution layer, recurrent layer, and transcription layer. In the head of CRNN, the convolution layer automatically extracts a feature sequence from each input image. In the neck, a recursive network is established, which is output by the convolution layer to predict each frame of the feature sequence. The transcription layer is at the tail of the CRNN, which converts the prediction of each frame of the cyclic layer into a tag sequence, and maps out the blank and redundant parts in the sequence. Although CRNN is composed of three parts: head, neck, and tail, it can still use

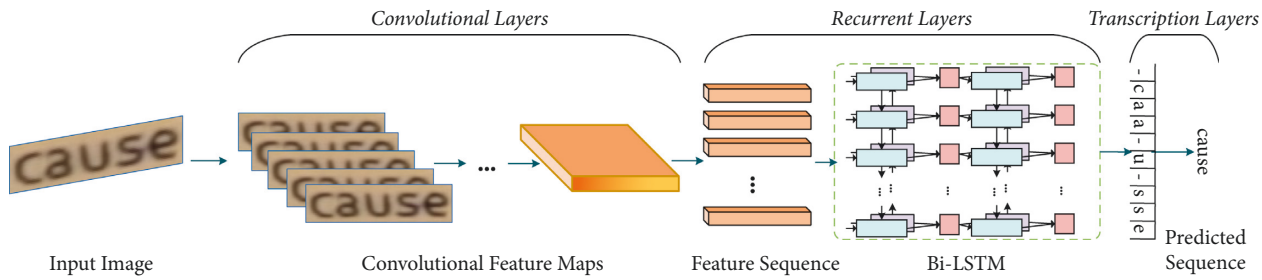

Figure 5: The overall architecture of CRNN.

the same loss function for joint training. CRNN text recognition algorithm introduces a bidirectional long short-term memory (Bi-LSTM) network to enhance context modeling, inputs the output feature sequence to CTC module, and directly decodes the sequence results. This algorithm is widely used in scene text image recognition tasks.

3.4. The Overall Process of the System RWYI. The proposed fast auxiliary text reading system RWYI can achieve accurate localization of the text of interest to readers and improve the accuracy of text recognition by enhancing the quality of text images. A summary of the proposed method is shown in Algorithm 1. Among them, HTD is the hand-text hybrid object detection model, FCSRN is a text-image super-resolution model with font-consistency prior, and CRNN is the proposed text recognition operator. The input $\mathbf{I_S}$ is an image obtained by the visual sensor in the reading scene, and the output $\mathbf{C_I}$ is the content of the text that the reader is interested in. In the first step, we use the single-stage hybrid object detection method HTD to locate the target text area pointed at by the finger on the input image $\mathbf{I_S}$, and the best prediction frame $\mathbf{O_{ht}^*}$ obtained by weighted NMS screening is cropped the text image $\mathbf{I_{ht}^*}$ with lower quality than the original input image. In the second step, we perform superresolution processing on the detected low-quality text images after alignment processing, and obtain an enhanced high-resolution text image $\mathbf{I_{SR}}$. In the third step, we use high-quality text images for text recognition and use the CRNN text recognition operator to recognize the content $\mathbf{C_I}$ of the text of interest pointed at by the finger.

4. Experimental Results

The proposed method uses HTD to directly locate the text of interest to readers and embeds font-consistent priors for text-image superresolution recovery of text images. To verify whether these two modules can play a role in the auxiliary text reading scenario, we design experiments to verify the effectiveness of the two modules in the proposed method. In addition, a comparative experiment is conducted between our proposed fast auxiliary text reading method and the traditional multistage combination method to comprehensively consider the performance of our proposed method. All experiments were performed on a machine with an Ubuntu 18.04 operating system, two NVIDIA RTX3090 24 GB GPUs, 128G DDR4 RAM, and two Intel Xeon Gold 6148 processors, and the system was developed using the Python programming language.

4.1. Evaluation of Text Localization Performance. The purpose of the HTD task is to locate the text of interest pointed at by a reader's hand in an image. As the primary task of auxiliary text reading, the accuracy and speed of its localization should be considered. The traditional localization methods using the combination of multiple object detection tasks can be roughly divided into three types.

Method 1. The global text and hand key points are detected first, and the coordinates of specific key points are matched with the text box to locate the text of interest.

Method 2. The key points of the hand are detected first, and the text of interest is located in the preset frame at the specific key point.

Method 3. First, the finger distribution is detected, and then the text closest to the finger is detected as the text of interest according to the finger distribution.

It should be noted that the detection algorithm of finger distribution and hand key points in the traditional multi-model combination method is the same as that in the proposed target detection algorithm, and the text detection algorithm uses EAST [29]. Using the above three traditional methods and the proposed HTD method, the test is carried out on a homemade private HTD dataset, and the obtained detection accuracy (mAP) and the number of images processed per second (FPS) are taken as performance evaluation metrics for text localization models of interest.

Figure 6 shows the mAP values and speeds (FPS) of four methods for locating the text of interest pointed at by a reader's finger; the HTD private dataset was used for this task. The experimental results show that the hand-text hybrid target detection method proposed by us is better than the traditional method in detection accuracy, especially in the index of localization speed. Our method is much faster than the multiple detection model combination localization method, because our method only needs to detect the image once, and there are no engineering strategies such as target position alignment and fixed area mask. In terms of recognition accuracy, HTD adds finger features for learning, which makes the target location of interested text more accurate.

Model: *HTD, FCSRN, CRNN*
Input: Image or video frame in text reading scenarios $\mathbf{I_S}$.
Output: The content of the text of interest to the reader $\mathbf{C_I}$.
While $(\mathbf{I_S})$**Do**
  Stage 1:
    $\mathbf{I_S} \longrightarrow$ **HTD** Hand-text hybrid object detection
    If there is a hybrid hand-text object **Then**
      $\mathbf{O^n_{ht}} \leftarrow \mathbf{I_S}$ Obtain n hand-text hybrid object prediction boxes $\mathbf{O^n_{ht}}$
      $\mathbf{O^*_{ht}} \leftarrow$ **Weighted NMS** Screening to get the best prediction box $\mathbf{O^*_{ht}}$
      $\mathbf{I^*_{ht}} \leftarrow \mathbf{O^*_{ht}} \cap \mathbf{I_S}$ Low-quality image of the text area $\mathbf{I^*_{ht}}$
    Else
      $\mathbf{I^*_{ht}}$ = None
    End
  Stage 2:
    $\mathbf{I^*_{ht}} \longrightarrow$ **FCSRN** Text image superresolution processing
    If $\mathbf{I^*_{ht}}$! = None **Then**
      $\mathbf{I_{LR}} \leftarrow \mathbf{I^*_{ht}}$ Aligned low-quality text image $\mathbf{I_{LR}}$
      $\mathbf{I_{SR}} \leftarrow \mathbf{I^*_{ht}} \cap \mathbf{P_{FC}}$ High-quality text area image $\mathbf{I_{SR}}$, font-consistency prior $\mathbf{P_{FC}}$
    Else
      $\mathbf{I_{SR}}$ = None
    End
  Stage 3:
    $\mathbf{I_{SR}} \longrightarrow$ **CRNN** Text recognition
    If $\mathbf{I_{SR}}$! = None Then
      $\mathbf{C_I} \leftarrow \mathbf{I_{SR}}$ The content of the text of interest $\mathbf{C_I}$
    Else
      $\mathbf{C_I}$ = None
    End
  If $\mathbf{C_I}$! = None Then
    $\mathbf{I_S} = \mathbf{I_{S+1}}$ Continue to the next image or video frame $\mathbf{I_{S+1}}$
    return $\mathbf{C_I}$
    End
  End

Algorithm 1: The overall process of the system RWYI.

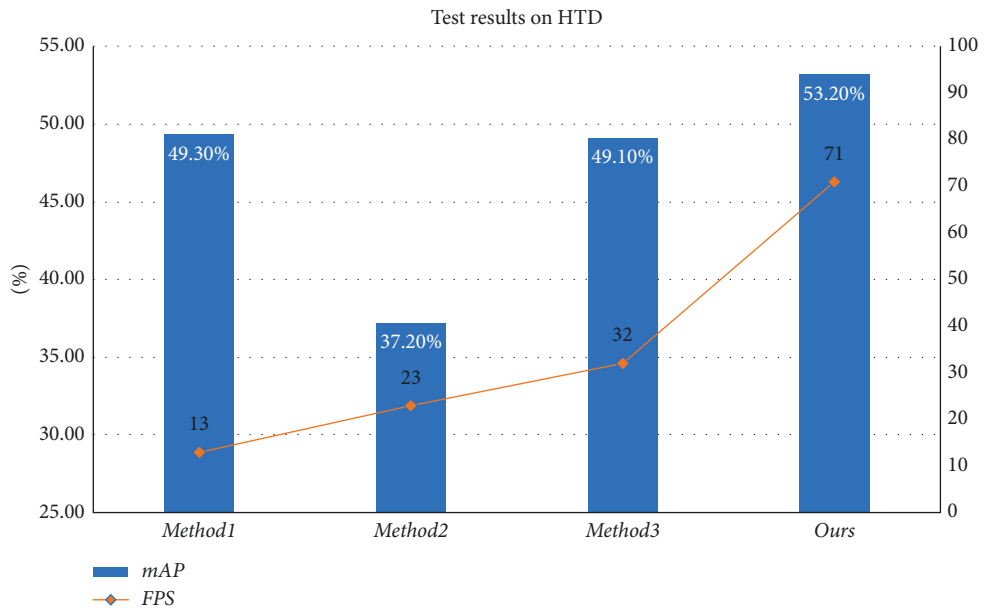

Figure 6: Comparison of the results of various methods of text location of interest on the HTD dataset.

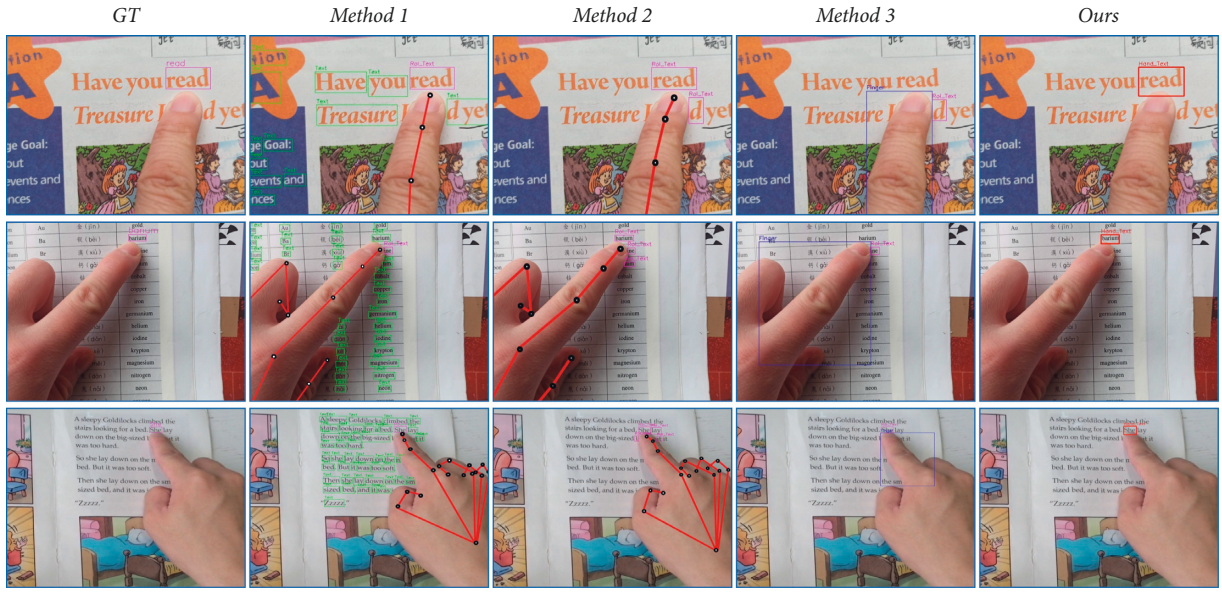

Figure 7: Visualization of the results of multiple interested text localization methods on the HTD dataset.

To more intuitively demonstrate the advantages of HTD and traditional localization methods on the HTD dataset, we visualized some examples of different interesting text localization methods on the HTD test set. As shown in Figure 7, the first column is the text that the actual reader is interested in; the second, third, and fourth columns are the positioning results of traditional methods 1, 2, and 3; and the fifth column is the proposed hand-text hybrid target detection and positioning result of the text. From the visualization results, it can be seen that the proposed hand-text hybrid target detection module can more accurately locate the text of interest to readers.

We also generalize different object detection networks, including YOLO-V4 [27], YOLO-V3 [25], and a single shot detector (SSD) [26], for HTD. Similarly, we evaluate the detection accuracy of these popular object detection networks on the HTD dataset. Table 3 shows the experimental results of our selected YOLO-V5 [9] model and other networks. Although this is not the focus of our experiments, it can be seen from the experimental data that the mAP value of the YOLO-V5 model we selected is higher than those of the other models because of its unique data enhancement method and novel feature extraction network.

Based on the above discussion, it is confirmed that the proposed method based on HTD is effective for locating the text of interest pointed at by a reader's finger in the auxiliary text reading scene and can quickly locate the text of interest from a reading scene image.

4.2. Evaluation of Text-Image Superresolution Performance. TextZoom is the first dataset to focus on real text-image superresolution and contains a total of 21,740 low-resolution-high-resolution text-image pairs and the text content as the text recognition label for each sample. We conduct a performance evaluation of FCSRN on TextZoom. On the one hand, the purpose of FCSRN is to improve the quality of text images and improve the accuracy of text recognition for downstream tasks. Therefore, we use the most general CRNN [11] text recognition operator to compare different text-image superresolution methods with our method. On the other hand, FCSRN is still intended for the image superresolution task, so no matter what the application scenario is, the use of the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) to evaluate the quality of superresolution images is necessary. We compare this model with other existing image superresolution algorithms and present the experimental results of a series of different image superresolution algorithms, including SCRNN [17], VDSR [18], SRResNet [19], RRDB [20], RDN [21], and TSRN [10], on the TextZoom dataset. The results are shown in Table 4. From the experimental data, it can be concluded that our method has achieved advanced performance on the test sets with different difficulty levels from TextZoom using the CRNN [11] text recognition accuracy for performance evaluation. According to the distribution of the different difficulty subsets of the TextZoom test set, the average text recognition accuracy is calculated, and the optimal performance is still achieved. Compared with other image superresolution methods, the image quality is also improved. This may be because the font-consistency prior can effectively promote the machine to extract text features from the same font category images. At the same time, it also shows that the minimalist prior fusion method can completely

Table 3: Comparison of the detection accuracy of different object detection networks on the HTD dataset, and the models involved in the comparison include YOLO-V4 [27], YOLO-V3 [25], SSD [26], and YOLO-V5 [9].

	YOLO-V3 [25]	SSD [26]	YOLO-V4 [27]	YOLO-V5 [9]
mAP	30.6	36.7	51.1	**53.2**

Table 4: Performance of the state-of-the-art SISR algorithms on the three subsets of TextZoom [10].

Method	Accuracy of CRNN [11]				PSNR	SSIM
	Easy (%)	Medium (%)	Hard (%)	Average (%)		
SCRNN [17]	38.7	21.6	20.9	27.7	20.78	0.7227
VDSR [18]	41.2	25.6	23.3	30.7	21.31	0.7331
SRResNet [19]	39.7	27.6	22.7	30.6	21.03	0.7403
RRDB [20]	40.6	22.1	21.9	28.9	19.99	0.7196
RDN [21]	41.6	24.4	23.5	30.5	20.41	0.7312
TSRN [10]	52.5	38.2	31.4	41.4	**21.42**	0.7690
FCSRN (ours)	**56.1**	**43.0**	**32.9**	**45.0**	21.18	**0.7771**

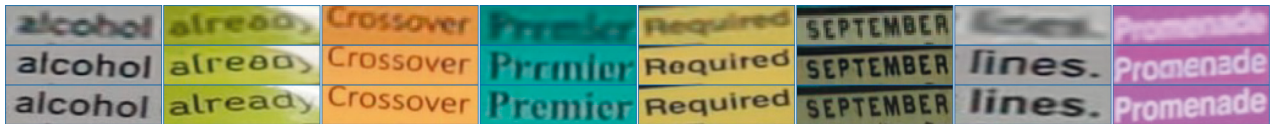

Figure 8: Sample test results of FCSRN on TextZoom.

Table 5: Processing times and the text recognition accuracies of various auxiliary text reading methods on different resolution images from the HTD dataset.

Traditional method 3	HOD	FCSRN	Time (ms)			Acc (%)
			1080 × 720	1920 × 1080	4096 × 2160	
√			75	120	210	38.7
√		√	100	145	245	42.6
	√		**59**	**88**	**146**	40.1
	Ours		84	113	171	**53.1**

transfer the feature information required for the superresolution of text images.

Similarly, Figure 8 shows some test results of our proposed FCSRN model on TextZoom [10], where the first row is a low-resolution image, the second row is our network-obtained superresolution map, and the third row is the high-resolution map. The experiments confirm from many aspects that our proposed font-consistency prior text-image superresolution network can improve the quality of text images in the context of auxiliary text reading.

4.3. Evaluation of the Overall Performance of RWYI. For practicality, in addition to evaluating the performance of individual modules, we should also focus on the overall performance of the system approach. However, for the overall performance evaluation experiment of our proposed fast auxiliary text reading method, we only care about the recognition efficiency of the text that readers are interested in from the reading scene. In response to this problem, we designed a corresponding experiment to explore the overall performance of RWYI.

We additionally make text content labels for the test set in the HTD dataset so that the system can recognize the text of interest pointed at by a reader's finger and use the recognition accuracy (Acc) as one of the system performance evaluation indicators. In addition, using a Logitech C1000 E multiresolution webcam, we adjust the lens resolution to obtain different quality image frames of different reading scenes and input them into the auxiliary text reading system we built. The computing system processes each image, and the time spent (time) is used as another indicator for system performance evaluation. We divide our proposed RWYI into three main modules: text localization, text superresolution, and text recognition. Under the premise of keeping the text recognition operator unchanged, three other auxiliary text reading methods are constructed to evaluate our proposed method. The experimental results are shown in Table 5. Among them, Traditional Method 3 is the third method with better performance among the three traditional multiple object detection model combined text localization methods; that is, to locate the text closest to the finger in a fixed area near the finger, HOD represents the HTD method for localizing text, and FCSRN is a text-image superresolution network embedded with the font-consistency prior.

From the abovementioned qualitative experimental results, it can be concluded that the accuracy rate of RWYI in identifying the text of interest is better than that of the other methods constructed. Although the auxiliary text reading system that does not use FCSRN to restore text images will speed up the processing of each image, the recognition accuracy is 13% lower than that of the proposed RWYI method, which also shows that FCSRN can promote the task of text recognition. Balancing the relationship between image processing speed and text recognition accuracy, the proposed RWYI method should be the

optimal aided text reading method. The efficiency of each method in processing images of different resolutions follows the abovementioned conclusions. This can also indirectly verify that the hybrid object detection approach in the proposed method can improve the speed of the auxiliary text reading task, and the text superresolution processing module can improve the accuracy of the auxiliary text reading task.

5. Conclusion

In this paper, we propose a fast auxiliary text reading method that improves a reader's text reading experience from a human-centered perspective. The method consists of three main tasks: first, the proposed hand-text hybrid object detection (HTD) model is used to quickly locate the text of interest to a reader in the input reading scene image, and then the proposed text-image superresolution model embedded with font-consistency priors are used to restore the low-resolution text image of the location to a high-resolution image to significantly improve the text image quality. Finally, the regular text recognition CRNN algorithm is used to identify and obtain the content of the text that a reader is interested in. To demonstrate the effectiveness of the proposed method, three quantitative comparing experiments were designed. The experimental results show that in the task of text location that readers are interested in, the proposed HTD is better than the existing multiple detection model combination method, and can locate the target region more quickly and accurately. In the text-image superresolution task, the proposed FCSRN can significantly improve the quality of the text images and promote text recognition compared with other image superresolution methods. In summary, we adopt the idea of directly locating the text region of interest to the reader and improving the pixel quality of the text image and propose an efficient and accurate auxiliary text reading method. However, the interesting text location used in our RWYI does not make full use of the regular arrangement of text in reading books but simply locates the interested text region in the image. In the future, we hope to take the document image layout as an auxiliary condition of the hand-text hybrid object model, so as to realize the handwritten interaction in a richer reading scene.

Abbreviations

HTD: Hand-text hybrid object detection
FCSRN: Font-consistent prior text image superresolution network
CRNN: Convolutional recurrent neural network
RWYI: Reading what you are interested in
HMI: Human-machine interaction
HCI: Human-computer interaction
FPNs: Feature pyramid networks
PAN: Path aggregation network
TSRN: Text superresolution network
ROI: Region of interest
YOLO: You only look once
NMS: Nonmaximum suppression
IOU: Intersection over union
FCE: Font-consistency encoder
STN: Spatial transformation network
SRB: Sequential residual block
CBLs: Convolution batch normalization Leaky-ReLU modules
Bi-LSTM: Bidirectional long short-term memory
CTC: Connectionist temporal classification
FSRBs: Font consistent prior text image superresolution blocks
LR: Low-resolution
SR: Superresolution
HR: High-resolution
PSNR: Peak signal-to-noise ratio
SSIM: Structural SIMilarity index
SCRNN: Superresolution convolutional neural network
VDSR: Very deep convolutional network
SRResNet: Superresolution residual network
RRDB: Residual-in-residual dense block
RDN: Residual dense network
SSD: Single shot multibox detector
HOD: Hybrid object detection.

Data Availability

The data that support the findings of this study are available from the corresponding author HW upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

ZY and HZ conceptualized the study. ZY and HW proposed the methodology. HW, HY, and HZ provided software. HY validated the study. HZ investigated the study. ZY and HW wrote the original draft. HW and HY reviewed and edited the manuscript. HW and HZ were responsible for resources. HZ supervised the study. ZY and HZ were responsible for project administration. ZY and HZ were responsible for funding acquisition. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant nos. 62171327, 62171328, 62072350, and 61773295), the open fund projects of Hubei Key Laboratory of Intelligent Robot (Wuhan Institute of Technology) (Grant no. HBIR 202004), the Collaborative Intelligent Robot Production and Education Integrates Innovative Application Platform based on the Industrial Internet (Grant no. 2020CJPT004), the first batch of application basic technology and science research foundation in Hubei Nuclear Power Operation Engineering

Technology Research Center (Grant no. B210610), and the GDPST&DOBOT Collaborative Innovation Center (Grant no. K01057060).

References

[1] A. W. Yu, D. Dohan, M. T. Luong et al., "Qanet: combining local convolution with global self-attention for reading comprehension," in *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, BC, Canada, April 2018, https://arxiv.org/abs/1804.09541.

[2] B. K. Chakraborty, D. Sarma, M. K. Bhuyan, and K. F. MacDorman, "Review of constraints on vision-based gesture recognition for human–computer interaction," *IET Computer Vision*, vol. 12, no. 1, pp. 3–15, 2018.

[3] Q. Yang, T. Lu, and H. Zhou, "A spatio-temporal motion network for action recognition based on spatial attention," *Entropy*, vol. 24, no. 3, 2022.

[4] Z. Chen, L. Wu, H. He, Z. Jiao, and L. Wu, "Vision-based skeleton motion phase to evaluate working behavior: case study of ladder climbing safety," *Human-centric Computing and Information Sciences*, vol. 12, 2022.

[5] C. H. Choi, J. Kim, J. Hyun, Y. Kim, and B. Moon, "Face detection using haar cascade classifiers based on vertical component calibration," *Human-centric Computing and Information Sciences*, vol. 12, 2022.

[6] Z. Mu, L. Jin, J. Yin, and Q. Wang, *Research on a Driver Fatigue Detection Model Based on Image Processing*, 2022.

[7] H. Lin, "How the reading pen works," 2016, http://www.cgan.net/cganself/founder/?p=3489.

[8] B. Technology, "A word stick to solve problems that electronic dictionaries and word memorization apps cannot solve," 2019, https://new.qq.com/omn/20190717/20190717A0JIAE00.html.

[9] G. Jocher, A. Stoken, J. Borovec, A. Chaurasia, T. Xie, and C. Liu, *ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 Models, AWS, Supervise.Ly and YouTube Integrations*GIT HUB, 2021.

[10] W. Wang, E. Xie, X. Liu, W. Wang, D. Liang, and C. Shen, "Scene text image super-resolution in the wild," in *Proceedings of the European Conference on Computer Vision*, pp. 650–666, Springer, Glasgow, UK, August, 2020.

[11] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, 2017.

[12] F. Ren and Y. Bao, "A review on human-computer interaction and intelligent robots," *International Journal of Information Technology and Decision Making*, vol. 19, no. 1, pp. 5–47, 2020.

[13] J. Kim, M. El-Khamy, and J. Lee, "Residual LSTM: design of a deep recurrent architecture for distant speech recognition," 2017, https://arxiv.org/abs/1701.03360.

[14] H. Zhou, Z. Xu, Y. Tian, Z. Yu, Y. Zhang, and J. Ma, "Interpolation-based nonrigid deformation estimation under manifold regularization constraint," *Pattern Recognition*, vol. 128, Article ID 108695, 2022.

[15] K. Li, W. Zhang, X. Tian, J. Ma, H. Zhou, and Z. Wang, "Variation-net: interpretable variation-inspired deep network for pansharpening," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, Shenzhen, China, July 2021.

[16] W. Wu, D. Zhang, J. Hou, Y. Wang, T. Lu, and H. Zhou, "Semantic guided infrared and visible image fusion," *IEICE - Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E104.A, no. 12, pp. 1733–1738, 2021.

[17] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[18] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1646–1654, Las Vegas, NV, USA, June 2016.

[19] C. Ledig, L. Theis, F. Huszár, C. Jose, C. Andrew, and A. Acosta, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, Honolulu, HI, USA, July 2017.

[20] X. Wang, K. Yu, S. Wu et al., "Esrgan: enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, Munich, Germany, September 2018.

[21] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2472–2481, Salt Lake City, UT, USA, June 2018.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[23] Z. Cai and N. Vasconcelos, "Cascade r-cnn: delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6154–6162, Salt Lake City, UT, USA, June 2018.

[24] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, Venice, Italy, October 2017.

[25] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," 2018, http://arxiv.org/abs/1804.02767.

[26] D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: single shot multibox detector," in *Proceedings of the European conference on computer vision*, pp. 21–37, Springer, Amsterdam, The Netherlands, October 2016.

[27] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: optimal speed and accuracy of object detection," 2020, http://arxiv.org/abs/2004.10934.

[28] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of the European conference on computer vision*, pp. 213–229, Springer, Glasgow, UK, August 2020.

[29] X. Zhou, C. Yao, H. Wen et al., "East: an efficient and accurate scene text detector," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 5551–5560, Honolulu, HI, USA, July 2017.

[30] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proceedings of the European conference on computer vision*, pp. 56–72, Springer, Amsterdam, The Netherlands, October, 2016.

[31] Y. Zhu, J. Chen, L. Liang, Z. Kuang, L. Jin, and W. Zhang, "Fourier contour embedding for arbitrary-shaped text detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3123–3131, Nashville, TN, USA, June 2021.

[32] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 67–83, Munich, Germany, September 2018.

[33] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "Aster: an attentional scene text recognizer with flexible rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2035–2048, 2019.

[34] C. Luo, L. Jin, and Z. Sun, "Moran: a multi-object rectified attention network for scene text recognition," *Pattern Recognition*, vol. 90, pp. 109–118, 2019.

[35] M. Liao, G. Pang, J. Huang, T. Hassner, and X. Bai, "Mask textspotter v3: segmentation proposal network for robust scene text spotting," in *Proceedings of the European Conference on Computer Vision*, pp. 706–722, Springer, Glasgow, UK, August 2020.

[36] W. Wang, E. Xie, X. Li et al., "PAN++: towards efficient and accurate End-to-End spotting of arbitrarily-shaped text," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[37] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, Honolulu, HI, USA, July 2017.

[38] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8759–8768, Salt Lake City, UT, USA, June 2018.

[39] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.