This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3025696, IEEE Transactions on Multimedia

LCSegNet: An Efficient Semantic Segmentation Network for Large-Scale Complex Chinese Character Recognition

Xiangping Wu, Qingcai Chen, *Member, IEEE,* Yulun Xiao, Wei Li, Xin Liu, and Baotian Hu

***Abstract*—Complex scene character recognition is a challenging yet important task in machine learning, especially for languages with large character sets, such as Chinese, which is composed of hieroglyphics with large-scale categories and similar glyphs. Recently, state-of-the-art methods based on semantic segmentation have achieved great success in scene parsing and have been applied in scene text recognition. However, because of limitations in terms of memory and computation, they are only applied in the small category recognition tasks, such as tasks involving English alphabets and digits. In this paper, we propose an efficient semantic segmentation model based on label coding (LC), called LCSegNet, to recognize large-scale Chinese characters. First, to reduce the number of labels, we design a new label coding method based on the Wubi Chinese characters code, called Wubi-CRF. In this method, glyphs and structure information of Chinese characters are encoded into 140-bit labels. Second, we employ an efficient semantic segmentation model for pixel-wise prediction and utilize a conditional random field (CRF) module to learn the constraint rules of Wubi-like coding. Finally, experiments are conducted on three benchmarks: a large Chinese text dataset in the wild (CTW), ICDAR2019-ReCTS, and HIT-OR3C dataset. Results show that the proposed method achieves state-of-the-art performances in both complex scene and handwritten character recognition tasks.**

***Index Terms*—large-scale categories, label coding, semantic segmentation, character recognition, complex scene, handwriting recognition.**

I. Introduction

NATURAL scene character recognition is an active research topic in the field of computer vision because of its wide application, such as in robotic process automation (RPA), human-computer interaction, automatic driving. Large-scale categories and similar glyphs pose great challenges for Chinese character recognition tasks. With the development of deep learning, great success has been achieved in scene text recognition [1]–[3] and handwriting recognition [4]–[7] by using deep networks.

In traditional methods, usually, handcrafted features are used to represent character images [8]–[10]. The histogram of oriented gradient (HOG), co-occurrence HOG (Co-HOG), and convolutional Co-HOG (ConvCo-HOG) methods are employed to extract image features. For instance, Su *et al.* [10] use the HOG method to convert an image into sequential column vectors. With the development of deep learning, the existing recognition methods mostly use the convolutional neural network (CNN) [11] to extract low/mid/high-level features automatically [12]–[14]. These methods convert a character image into one-dimensional features. For instance, Xiao *et al.* [14] design a nine-layer network for offline handwritten Chinese character recognition (HCCR). Image features extracted by a seven-layer CNN are stretched into fixed-length vectors and then input into two fully-connected layers for classification. Shi *et al.* [13] propose a convolutional recurrent neural network (CRNN) for sequence labeling. They divide the extracted image features equally into a sequence of features and use connectionist temporal classification (CTC) [15] to obtain the final combined result. Fig. 1(a), 1(b) show two examples of character recognition and text recognition from a one-dimensional perspective. Although these methods have excellent performance, they directly compress the features of the characters into a one-dimensional form, and thereby the structural information of the characters are lost and background noise is introduced [16].

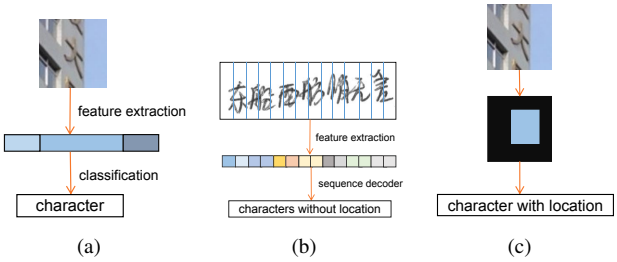


Fig. 1. Illustration of Chinese characters recognition processes. (a) character recognition process from a one-dimensional perspective. (b) text recognition process from sequence prediction perspective. (c) character recognition based on semantic segmentation.

Semantic image segmentation of arbitrary sizes has gained importance in recent times. Semantic segmentation is the recognition of content and location in an image by pixel-wise prediction. This technique shows great potential in object detection and recognition tasks [17]–[20]. In such segmentation, fully convolutional networks (FCN) are used for spatially dense prediction tasks. Recently, state-of-the-art algorithms based on semantic segmentation were applied for scene text recognition (STR). Lyu *et al.* [21] propose an end-to-end

This work is supported by Natural Science Foundation of China (Grant No. 61872113, 61876052), and Strategic Emerging Industry Development Special Funds of Shenzhen (Grant No. XMHT20190108009, JCYJ20190806112210067, JCYJ20170811153836555). The authors are with Shenzhen Chinese Calligraphy Digital Simulation Engineering Laboratory, Harbin Institute of Technology (Shenzhen), Shenzhen University Town, Xili, Shenzhen 518055, China (e-mail: wxpleduole@gmail.com; qingcai.chen@hit.edu.cn; xiaoyulun@stu.hit.edu.cn; weili_hitwh@163.com; hit.liuxin@gmail.com; baotian.nlp@gmail.com).

trainable deep neural network termed Mask TextSpotter for spotting text with arbitrary shapes. Liao *et al.* [16] propose the "character attention fully convolutional network" to recognize scene text and predict character positions. Despite their success, the existing methods for STR are only designed for English and for digital recognition with a few categories. Since the amount of memory and computation needs increase rapidly with an increase in the number of categories, using these methods for large-category tasks, e.g., the recognition of Chinese characters, still remains a great challenge.

This is the challenge that the present study aimed to address. An effective semantic segmentation method based on label coding, called LCSegNet, is proposed to recognize large-scale Chinese characters by pixel-wise predictions from a two-dimensional perspective, as shown in Fig. 1(c). The four main contributions of this work are as follows.

1) A novel label coding method based on the Wubi input method and a conditional random field (CRF) is proposed. This method is called Wubi-CRF, and it encodes the glyphs and structural information of Chinese characters into 140-bit labels. Wubi-CRF can be applied for both Chinese and non-Chinese character recognition.

2) A lightweight semantic segmentation model is integrated to solve complex scenes and handwritten character recognition problems; this is possibly the first semantic segmentation method for Chinese character recognition.

3) The number of parameters and amount of computation required for the proposed LCSegNet is independent of the number of Chinese character categories. Compared to the traditional FCN-8s model, LCSegNet reduces the computational cost to 1/527 and compresses the number of parameters to 1/120 when classifying 3650 categories.

4) Experiments are conducted on three public datasets, and state-of-the-art performances have been achieved on two of three. The results show that the proposed method performs better than the comparing methods in dealing with complex scene character recognition and similar handwritten Chinese character recognition.

The rest of this paper is organized as follows. Section II briefly reviews related work. Section III details the proposed LCSegNet and Wubi-CRF method. Section IV describes the experimental settings, results, and analyses. Section V makes the conclusion.

II. Related Work

A. Methods for Scene Character Recognition

Because of the characteristics of complex background, diverse characters, uneven illumination, and low resolution, scene character recognition (SCR) tasks are not addressed well by the general optical character recognition technology. To tackle the issues in SCR, early methods heavily relied on handcrafted features. Zhang *et al.* [22] employ the histogram of oriented gradient (HOG) to extract image features. Tian *et al.* [23] propose two new feature descriptors based on the HOG method—the Co-HOG and ConvCo-HOG—to extract scene image features of different languages. Su *et al.* [24] employ the HOG to convert a word image into a sequential feature and feed them into LSTM and CTC to handle word recognition without character segmentation. Despite their performances, these methods have the limitation that they are usually designed for specific scenarios.

Recently, deep learning methods have been introduced for SCR tasks, and outstanding performances have been achieved [25]–[27]. In these tasks, the CNN is usually employed to extract different level image features automatically. For example, Wang *et al.* [28] use the pre-trained CNN to extract the convolution activations and learn spatially embedded discriminative part detectors for SCR. Zhang *et al.* [29] design a consecutive convolutional activations (CCA) method to integrate low-level and high-level patterns into the final feature vector. Generally, these methods transform two-dimensional images into one-dimensional feature vectors. Thereby, they introduce background noise and are unable to locate the specific position of characters [16].

Chinese characters have a large number of categories and complex structures. Hence, some scholars design generative tasks [30]–[32] to obtain the discriminative features. For example, Wang *et al.* [33] design a novel method based on generative adversarial networks (GANs) [34] to learn canonical forms of glyphs in several standard font styles. Their framework consists of a feature extraction network, character classification network, glyph generating network and glyph discrimination network. Lin *et al.* [35] propose a new architecture called the multitask coupled generative adversarial network (MtC-GAN) to generate realistic data to improve the accuracy of SCR. However, both [33] and [35] had to use the font renderer to pre-generate standard printed characters.

B. Methods for Semantic Segmentation

A segmentation-recognition framework is a basic approach in the text line recognition field. Many efficient methods are based on the over-segmentation framework. For instance, Liu's team [36], [37] employ the connected component-based method to over segment the text line image into a sequence of primitive segments and then combine them as candidate character patterns, which are classified to form character candidate lattices. Finally, they use a word-level language model or semi-CRF module to search the segmentation recognition path. Peng *et al.* [38] employ FCN to process images with diverse sizes and add a detection branch to obtain character segmentation. To avoid the segmentation procedure, some works focus on segmentation-free methods. Xie *et al.* [39] use FCN to learn images features and use residual recurrent networks with CTC to do the sequential transcription. Shi *et al.* [40] use spatial transformer network (STN) to rectify irregular text and utilize an attentional sequence-to-sequence model for end-to-end training. These methods are usually based on the LSTM-CTC framework for sequence prediction.

Unlike the traditional concept of segmentation, which segment text lines into characters, semantic segmentation is to classify each pixel of the image. Semantic segmentation [41]–[43] is the typical computer vision task of identifying the object category, location, and shape in the given image. Most algorithms based on semantic segmentation are extended

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3025696, IEEE Transactions on Multimedia

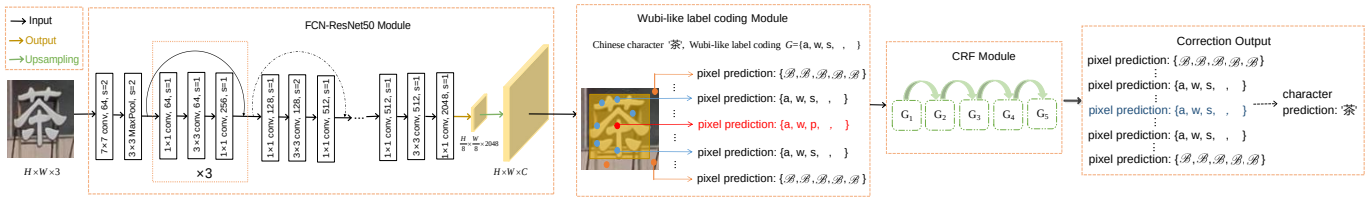


Fig. 2. Architecture of the LCSegNet method. The backbone network is based on the ResNet50 / Res2Net50 structure. First, the input image is reduced to 1/8 of the original image size after the ResNet50 / Res2Net50, and then upsampled to the original image size by bilinear interpolation. Second, the outputs of the FCN-ResNet50 module are pixel-wise classified into the Wubi-like coding labels. Finally, the CRF module is used to learn the constraint rules of the Wubi-like coding label for each pixel. In the test phase, the final output is obtained by using the Viterbi algorithm to search the optimal sequence for each pixel through the learned transition matrix.

from the architecture of a framework of FCN [20], such as DeepLab [44], U-Net [45], PSPNet [46], etc. Lyu *et al.* [21] and Liao *et al.* [16] successfully apply the semantic segmentation model to perform the STR task. However, they are able to recognize only small categories of characters, such as English characters and digits. Liao *et al.* [47] also integrate a spatial attention module (SAM) to improve the Mask TextSpotter [21] model and solve the problem of multi-language end-to-end recognition. However, when recognizing large categories, the character segmentation branch is disabled, and only SAM is used for sequence predictions, without pixel-level classification of Chinese characters [47]. In this paper, our work is based on a classic semantic segmentation model FCN [20]. There are some studies [38], [39], [48], [49] using FCN for Chinese character recognition. However, these methods only utilize FCN to extract features and do not perform deconvolution to the original image size for pixel-wise classification. They are not semantic segmentation-based models.

C. Methods for Label Coding

The computational complexity and model size increase with the number of categories. In the past few decades, many label coding methods [50]–[53] have been proposed to handle large-scale category classification. For example, Zhang *et al.* [53] design a label mapping method to convert the large-scale category classification problem into medium category classification sub-problems and trains a base learner for each sub-problem. For character recognition, a popular method based on character shape coding is employed [54]–[58]. For instance, Lu *et al.* [54] use three topological character shape features to convert each word image into a sequence of character codes. The shape features contain character ascenders/descenders, character holes, and character water reservoirs. Based on [54], Bai *et al.* [55] propose seven features to represent word images for keyword spotting in document images. Zhang *et al.* [57], [58] propose a radical analysis network (RAN) to decompose a Chinese character into radicals and several substructures, which are represented by the IDS sequence. They finally predict the IDS sequence and search the most likely sequence in the IDS dictionary to obtain the output character label.

The study focuses on Wubi coding, which is also a shape-based encoding method. In previous studies [59]–[61], attempts have been made to use the Wubi input method to obtain character embedding for natural language processing (NLP) tasks. For instance, Yang *et al.* [61] employ the five-stroke input method to decompose a character into different parts. They use one-hot embedding to represent each part and feed them into CNN or LSTM layer to obtain stroke-level embeddings. Then, the character embeddings are concatenated with stroke embeddings to form the final character representations for Chinese named entity recognition. Unlike these previous works wherein the Wubi coding method was used to represent the character embeddings that are input for NLP tasks, in this study, Wubi-like coding is used to encode the label system that is the output for image recognition tasks. In addition, in many works [37], [61], CRF is employed to obtain the optimal sequence. For instance, the CRF in [61] is used to consider the relationship between adjacent entity tags. Zhou *et al.* [37] propose a semi-Markov CRF (semi-CRF) method to model the relationship between characters through the language model for text line recognition. Different from these studies, in the present study, CRF is used to model the association between groups in Wubi-like label coding for character recognition.

III. Methodology

The framework of the proposed LCSegNet is shown in Fig. 2, which mainly consists of three modules, i.e., FCN-ResNet50, Wubi-like label coding, and CRF. Among them, the FCN-ResNet50 module is used to extract image features for upsampling to the input image size for pixel-level classification. The Wubi-like label coding module is utilized to encode the glyphs and structure information of Chinese characters into 140-bits labels. Further, the CRF module is employed to learn the transformation matrix between groups to avoid invalid label coding.

A. FCN-ResNet50 Module

We employ FCN [20] as our segmentation architecture for dense prediction of arbitrary size images. Unlike the baseline model FCN-8s [20], we utilize a pre-trained network ResNet50 [62] or Res2Net50 [63] instead of VGG-16 as the backbone network and discard the skip connections of the reuse features.

The ResNet50 module consists of five stages, the last four of which are stacked by multiple bottleneck building blocks. For an input image of shape $H \times W \times 3$, let H and W are the height and width of the input image, respectively. The output of stage-0 is the feature maps of size $\frac{H}{4} \times \frac{W}{4}$. Downsampling

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3025696, IEEE Transactions on Multimedia

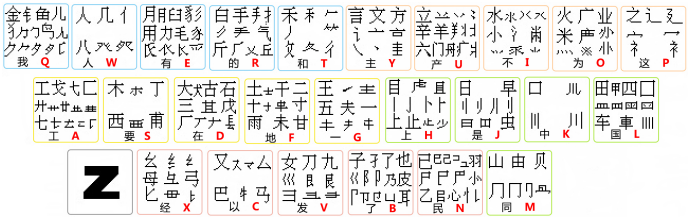


Fig. 3. Keyboard layout of the 98's version of Wubi radical table. Each color region denotes one type of stroke. The blue, orange, yellow, green, and pink regions indicate downwards right-to-left, dot or downwards left-to-right, horizontal, vertical, and hook strokes, respectively. Further, "z" is used as a wildcard.

is performed by stage-2 with a stride of 2. Following [64] and [65], we adopt dilated convolution instead of the 2×2 stride in stage-3 and stage-4 to enlarge the receptive field of networks without loss of resolution or coverage. After using the ResNet50 module, we can obtain the heatMap $\boldsymbol{F_{last}}$ of size $\frac{H}{8} \times \frac{W}{8}$. Let $\boldsymbol{F_{out}}$ be the final output of the FCN-ResNet50 module, then $\boldsymbol{F_{out}} = up\_8(\boldsymbol{F_{last}})$, where the up_8 represents 8 times upsampling. In this study, bilinear interpolation upsampling is used instead of the learnable deconvolution of the baseline model FCN-8s, thereby greatly reducing the amount of calculation and number of parameters. Finally, the output $\boldsymbol{F_{out}}$ of the same size as the original input image for pixel-wise classification is obtained.

In order to further improve the performance, we also try to use Res2Net50 [63] instead of ResNet50 [62] as the backbone network to extract multi-scale features at a granular level. The Res2Net50 module also contains five stages. In the stage-0, a 3×3 max-pool with the stride of 2 is performed after three 3×3 convolutions with the stride of 2, 1, 1 and the channels of 32, 32, 64. The parameters of the latter 4 stages are similar to ResNet50, except that the bottle2neck [63] is used instead of the bottleneck. In each bottle2neck, the width of filters is 26 and the number of scale is 4. After using the Res2Net50 module, the size of the output is reduced to 1/8 of the original image size. Bilinear interpolation is also used to upsample the output.

B. Wubi-like Label Coding Module

As described above, the output $\boldsymbol{F_{out}}$ of the FCN-ResNet50 module has the dimensions of $H \times W \times C$, where C denotes the number of classes of semantic segmentation. We assume $\boldsymbol{Y}$ is the ground truth label map $\boldsymbol{Y} \in \{0, 1, ..., C\}^{H \times W}$. In the original FCN, $\boldsymbol{Y}$ is encoded with one-hot coding, then $\boldsymbol{Y} \in \{0, 1\}^{H \times W \times C}$. For Chinese character recognition, C is the number of categories of characters. In level-1 and level-2 of GB2312-80, there are 3755 and 3008 character classes, respectively. Because of the large-scale nature of the categories, the number of parameters and computational complexity are extremely high.

In this section, we describe our novel coding method, namely Wubi-CRF, based on the Wubi encoding of Chinese characters[1]. Wubi is used as an encoding-based Chinese input method. Unlike the Chinese pinyin-based methods, the Wubi

[1]http://en.wikipedia.org/wiki/Wubi_method

Char	Parse	Code	Char	Parse	Code	Char	Parse	Code	Char	Parse	Code
晴	晴晴晴	jge	琴	琴琴琴	ggw	一	一	g	矗	矗矗矗矗	fhfh
请	请请请	yge	瑟	瑟瑟瑟	ggn	人	人	w	整	整整整整	gkih
情	情情情	nge	琶	琵琶琶	ggx	太	太太	dy	觀	觀觀觀觀	akgq
清	清清清	ige	琶	琶琶琶	ggc	开	开开	ga	耀	耀耀耀耀	iqny

Fig. 4. Examples of Wubi coding of Chinese characters. The first two columns represent the coding of similar characters. The third and fourth columns represent the coding of simple and complex characters, respectively.

method is based on the graphical structure of characters, and it literally means "five strokes". In fact, the Wubi method uses the so-called "radical" rather than a stroke as the minimal input unit. Each radical can represent one or multiple strokes. To ensure that extremely complex characters do not require too long coding, the Wubi method divides radicals into five types, which corresponding to the 26 ASCII characters a to z. The corresponding characters of the 98's version Wubi radical table are shown in Fig. 3.

The main motivation behind selecting the Wubi coding is inspired by its two important properties: 1) According to the Wubi method, Chinese characters are coded according to the structure of their strokes. 2) Each character can be represented by at the most five alphabet letters. These characteristics meet the requirements of retaining spatial structure information in semantic segmentation. Fig. 4 shows the Wubi coding of some Chinese characters. The first two columns show that similar characters have similar coding. From the last two columns, the simple characters have few coding letters, while complex characters have longer codes. To unify the representation, we use a Wubi-like encoding method, i.e., 5 letters are used to represent each Chinese character. If the original Wubi code of a character has less than 5 letters, the empty symbol $\varnothing$ is padded as a suffix. For convenience, we use groups to represent the character coding. Each Chinese character has 5 groups, represented by $\boldsymbol{G}$, and $\boldsymbol{G} = \{G_1, G_2, ..., G_n\}, n = 5$. Let $\boldsymbol{C}$ be the Wubi-like coding label of a pixel in character image; $\boldsymbol{C_i}$ is the one-hot representation of $G_i, i = [1, 5]$; then, the i-th group $\boldsymbol{C_i}$ contains 26 bits, representing key letter values from a to z. In order to identify the background category in semantic segmentation, we further expand the 26 bits of each group to 28 bits. Fig. 5 shows our Wubi-like coding. The 28 bits in each group are represented in a one-hot form. Finally, the index with a value of 1 in each group is extracted to form the final coding of the pixel in character image. In brief, a Wubi-like coding of a Chinese character consists of 140 bits, which can represent more than 27,000 Chinese characters.

In order to encode non-Chinese characters, the proposed Wubi-like coding method is further adapted to make it compatible with the multi-language situation. In this case, we still use 5 groups to represent non-Chinese characters. Similar to the background category extension, we expand each group from the previous 28 bits to $(28 + M)$ bits. M denotes the number of bits that need to be extended. For example, If $M = 3$, then $M^5 = 243$ non-Chinese characters can be encoded. Specifically, the M bits are added at the end of each group to represent non-Chinese characters. Then the j^{th} bit of the i^{th}

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3025696, IEEE Transactions on Multimedia

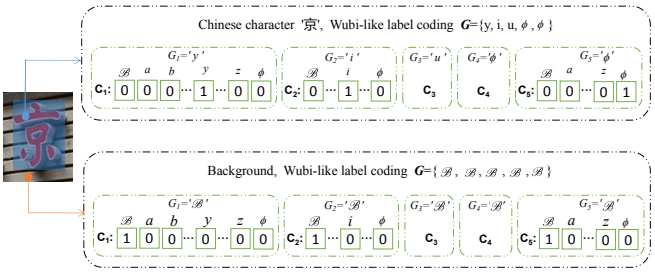


Fig. 5. The illustration of the Wubi-like method. Each character has five groups of codes, each of which consists of 28 bits. The 1st bit of each group identifies the background label, the 2nd to 27-th represents the Wubi code of the corresponding Chinese character that the pixel is belonged to, and the last bit denotes the empty label.

group in $\boldsymbol{C}$ can be rewritten as:

$$C_{ij} = \begin{cases} \mathscr{B}, \ j = 1 \\ a \sim z, \ 2 \leq j \leq 27 \\ \varnothing, \ j = 28 \\ \mathbb{E}, \ 29 \leq j < 29 + M \end{cases} \tag{1}$$

where $\mathscr{B}$ and $\varnothing$ represent the "background" and "empty" labels respectively, and $1 \leq i \leq 5$. $\mathbb{E}$ denotes the extended bits of non-Chinese characters coding. In this study, we set $M = 3$ to extend the coding of 26 uppercase letters, 26 lowercase letters, and 10 numeric characters. According to the proposed coding design, each group has 31 bits, and the label code of each character is 155 bits, which can represent more than 27,000 Chinese characters and 243 non-Chinese characters.

C. CRF Module

The general semantic segmentation models only predict the label for pixels without considering the smoothness and consistency of the label assignments [66]. Inspired by Chen *et al.* [44], in this study, we use CRF to model the association between groups in Wubi-like label coding. The CRF module is used to learn a transition matrix and avoid invalid Wubi-like coding. Let a label sequence of the pixel be a random field, i.e., $\boldsymbol{y} = \{y_1, ..., y_{|\boldsymbol{G}|}\}$, $y_i \in \{1, 2, ..., B\}$, B is the number of bits in each group. Let $\boldsymbol{P}$ to be the matrix of scores output by the Wubi-like label coding module; P_{ij} corresponds to the score of the j^{th} bit of the i^{th} group. Given an observed sequence $\boldsymbol{x} = \{x_1, ..., x_{|\boldsymbol{G}|}\}$ that is output by the Wubi-like label coding module, we can get the pixel label of the input image by maximizing the conditional probability $P(\boldsymbol{y}|\boldsymbol{x})$:

$$\begin{aligned} P(\boldsymbol{y}|\boldsymbol{x}) &= \frac{1}{Z(\boldsymbol{x})} \exp(s(\boldsymbol{x}, \boldsymbol{y})) \\ s(\boldsymbol{x}, \boldsymbol{y}) &= \sum_{i=1}^{|\boldsymbol{G}|} P_{ij} + \sum_{i=0}^{|\boldsymbol{G}|} A_{y_i, y_{i+1}} \end{aligned} \tag{2}$$

where $\boldsymbol{A}$ is a square matrix of transition scores of size $B+2$, which is learned from the CRF module. $A_{y_i, y_{i+1}}$ denotes the score of assigning labels y_i and y_{i+1} to the i^{th}, $(i+1)^{th}$ group simultaneously. $y_0, y_{|\boldsymbol{G}|+1}$ are the initial and end labels of a sequence. $Z(\boldsymbol{x})$ is a normalization term. In the test stage, according to the trained transition matrix, we use the Viterbi algorithm to search for the sequence of the maximum-likelihood.

D. Loss Function

In the proposed model, the loss function for one-hot coding can not be directly applied. We define a new loss function for the proposed model. Let the training set $\boldsymbol{T} = \left\{I^{(d)}, \boldsymbol{C}^{(d)}\right\}$, here $d = 1, 2, ..., |\boldsymbol{T}|$, $I^{(d)}$ is the d^{th} input image and $\boldsymbol{C}^{(d)}$ is the corresponding ground truth. In order to better learn the coding labels, we divided the 140-bits labels into 5 groups according to the characteristics of the Wubi-like coding, and each group is optimized using cross-entropy. Then, the loss function of LCSegNet without CRF module can be written as follows:

$$L_{entropy} = -\frac{1}{|\boldsymbol{T}|} \frac{1}{N} \frac{1}{|\boldsymbol{G}|} \sum_{d=1}^{|\boldsymbol{T}|} \sum_{k=1}^{N} \sum_{i=1}^{|\boldsymbol{G}|} \sum_{j=1}^{B} {C_{ij}}^{(d)(k)} \ln({p_{ij}}^{(d)(k)}) \tag{3}$$

where $\boldsymbol{G}$ is the Wubi-like coding label represented by five codes, and $|\boldsymbol{G}|$ is the number of groups in the Wubi-like coding label. $N = H \times W$ is the number of pixels of the input image. Further, ${p_{ij}}^{(d)(k)}$ is the probability of assigning labels C_{ij} to the j^{th} bit of the i^{th} group. Here, k and d denote the index of the pixel and input image, respectively.

Furthermore, to ensure that the five groups for a character code are valid according to the whole Wubi-like codebook, we employ the CRF module to automatically learn the constraints of the label sequence. According to Eq. (2), the final loss function of LCSegNet with CRF module can be rewritten as follows:

$$L_{crf} = -\frac{1}{|\boldsymbol{T}|} \frac{1}{N} \sum_{d=1}^{|\boldsymbol{T}|} \sum_{k=1}^{N} log(P(\boldsymbol{y}|\boldsymbol{x})^{(d)(k)}) \tag{4}$$

IV. Experiments

The proposed LCSegNet is evaluated on three benchmarks, which include both scene and handwritten character recognition datasets: Chinese Text in the Wild (CTW) [67] dataset, ICDAR2019-ReCTS [68], and HIT-OR3C dataset [69]. We use character recognition accuracy as a performance evaluation measure. The accuracy of the test set is defined as $acc = m_c/m$. Where m_c is the number of correctly recognized images, and m is the total number of test images. We evaluate the effectiveness of the proposed model by comparing it with state-of-the-art methods and different coding methods. Unless otherwise specified, the LCSegNet used in the following ablation experiments and experimental analysis uses ResNet50 as the backbone network.

A. Dataset

Since the main goal of the proposed model is to address the issue of character recognitions with large-scale categories and complex scenes, the main dataset which we conduct

our experiments on is CTW. To verify the effectiveness of the proposed LCSegNet in real scenes, we evaluate it on the ICDAR2019-ReCTS Dataset and compare it with those winning methods. An additional handwritten dataset is used to further verify its capability of classifying large-scale categories and its compatibility with non-Chinese characters.

1) CTW: CTW is a large dataset of Chinese text in street view images with about 1 million annotated Chinese characters. The dataset contains occluded, background complexity, distorted, 3D raised, wordart, and handwritten characters. The diversity, complexity and large categories make this dataset very challenging. According to the standard partitioned by [67], the training data contains 812,872 Chinese characters of 3650 categories and the test data consists of 103,519 Chinese characters.

2) ICDAR2019-ReCTS: ICDAR2019-ReCTS is a competition dataset for ICDAR 2019 robust reading challenge on reading Chinese text signboard. The dataset comes from the Chinese signboards in street view, which have different backgrounds, fonts, and layouts. This study focuses on task 1 for character recognition. There are 439946 character images for training and 29335 character images for testing. ICDAR2019-ReCTS contains simplified Chinese characters, traditional Chinese characters, letters, digits, punctuation symbols, etc., with a total of 4103 categories. In addition, we use images with 3D raised and word art attributes from the CTW dataset, and synthetic characters as the additional data.

3) HIT-OR3C: HIT-OR3C is an opening handwriting recognition corpus for Chinese characters. It consists of 5 subsets, namely, GB1, GB2, letter, digit, and document. The first 4 corpora contain 6825 categories generated by 122 writers. The document subset is produced by 20 persons and contains 2442 categories. In this study, we chose the previous four subsets of 100 writers as the training set (i.e., 6825 categories). The data of the remaining 22 writers' data and the document subset are used as test data.

B. Implementation Details

1) Data Preprocessing: For the CTW dataset, when using ResNet50 as the backbone network, we keep the aspect ratio of the original image and zoom the short side to 60. Random padding between 2 and 10 pixels is applied to the four sides of the scaled image. The size of the final image is 64×64 by random crop. In the test stage, four pixels are padded on each side of the image without random cropping. When using Res2Net50 as the backbone network, we zoom the short side to 92 and apply random padding the four sides. The size of the final image is 96×96 by random crop. For the ICDAR2019-ReCTS and HIT-OR3C datasets, the original image is normalized to 96×96 pixels.

2) Parameter Setting: All experiments on LCSegNet are optimized by the stochastic gradient descent (SGD) with a momentum set to 0.99. No data augmentation and no multi-scale training technology is applied in the model training step. In the test phase, we vote on the predicted results for all the pixels of an image, and the category with the largest number of votes is the final predicted character. Other parameters for each dataset are as follows.

CTW dataset: training epochs = 50, initial learning rate $lr(no\_CRF)$ =1.0e-4, $lr(CRF)$ =1.0e-6, constant learning rate, $|\boldsymbol{G}| = 5$, $M = 0$, $B = 28$.

ICDAR2019-ReCTS dataset: training epochs = 50, initial learning rate $lr(no\_CRF)$ =1.0e-3, $lr(CRF)$ =1.0e-4, learning rate is divided by 10 per 10 iterations, $|\boldsymbol{G}| = 5$, $M = 3$, $B = 31$.

HIT-OR3C dataset: training epochs = 35, initial learning rate $lr(no\_CRF)$ =1.0e-4, $lr(CRF)$ =1.0e-6, constant learning rate, $|\boldsymbol{G}| = 5$, $M = 3$, $B = 31$.

Especially, the training data of the CTW dataset is further divided into the actual training set and the validation set according to the ratio of 9:1 for parameter selecting.

C. Comparison Methods

1) Scene Character Recognition: On the CTW dataset, the state-of-the-art CNNs are used for comparison, which include **AlexNet** [70], **OverFeat** [71], **ResNet50** [62], **ResNet152** [62] and **Google Inception** [72]. **SMN** [73] is style-melt net that consists of a style net and content net for learning the style and content representations, respectively. **RAN** [58] is a radical analysis network, which use the DenseNet [74] as an encoder and the attentional RNN with gated recurrent units (GRU) as decoder. **Faster R-CNN** [75] is a object detection technology based on region proposal network (RPN). **ResNet50 (Wubi-CRF)** replaces one-hot with Wubi-CRF in ResNet50 model. **MtC-GAN** [35] with different backbones use multitask coupled GAN to generate synthetic data. For the segmentation-based methods, the mainstream **FCN-8s** [20] and **FCN-ResNet50** are chosen for comparison.

On the ICDAR2019-ReCTS dataset, **BASELINE-v1** uses traditional image classification methods and its ensemble. **Amap_CVLab** proposes the method that adds res-block [62] and se-block [76]. **TPS-ResNet-v1** consists of ResNet, BiLSTM, and spatial transformer network (STN) modules. **SANHL_v4** is an ensemble method that integrates 6 ResNet-34 models. **Tencent-DPPR** integrates five types of deep models based on CTC nets and multi-head attention nets. **ResNet_HUSTer** uses ResNet50 as the backbone. **ReCTS_Task1** is based on ResNet-152. **Task1-re5** combines the results of CNN and RNN models and uses a dictionary for post-processing.

2) Handwritten Character Recognition: On the HIT-OR3C dataset of 6825 categories, the comparison methods are as follows. **KMean+MQDF** [77] extracts the gradient direction features of characters, and then uses the K-mean method for coarse classification and the modified quadratic discriminant function (MQDF) method for fine classification. **HCCR-AlexNet** [78] consists of five convolutional layers and three fully connected layers. **SqueezeNet_BN** [79] is a lightweight network for image classification in which the fire module is introduced to build CNN architectures. The batch normalization operation is also performed to accelerate network convergence. **RAN** [58] and **JSRAN** [80] are based on a encoder-decoder framework, which decomposes Chinese characters into radicals and structures, and uses IDS sequences to represent them. They only recognize Chinese characters and

are not compatible with other languages. **GhostNet-1.0** [81] is composed of Ghost modules that use a series of linear transformations to generate more feature maps. **HCTR-SRM** [38] uses segmentation and recognition module for handwritten Chinese text line recognition. It consists of location, detection, and classification branches.

3) Coding Methods: To verify the efficacy of Wubi-CRF coding, we conduct experiments to compare different label coding methods. Except for loss function, the other experimental settings of the comparative experiments are remained the same.

a) Random binary coding: Random binary coding is based on a balanced binary tree, which randomly assigns characters to leaf nodes. For each node, the left branch is labeled as 0, and the right branch is labeled as 1. The path from the root node to the leaf node is the coding of the character. One bit is added in front of the path coding to represent the background category. Our previous experiments have shown that the training using multilabel margin loss, binary cross-entropy or Hamming loss [82] can not converge well. Hence, we use cosine similarity as the loss function.

b) Binary coding based on the confusion matrix: Binary coding based on the confusion matrix is based on a balanced binary tree. In this process, characters are assigned to leaf nodes corresponding to the confusion matrix. First, the confusion matrix of characters is constructed according to the recognition result of the validation set through a generic model ResNet50. Second, a balanced binary tree is constructed based on the confusion matrix, and all characters are allocated on leaf nodes. One bit is also added at the start node of the path coding to represent the background category. We construct a 12-level binary tree for the CTW dataset. Here, the cosine similarity is used as the loss function too.

c) One-hot coding: One-hot coding is a commonly used form of most models. Each bit corresponds to a character label. Therefore, the length of the label sequence is the same as the number of character categories. However, as number of the categories increases, both the number of parameters and the computational complexity will increase. In this study, we use cross-entropy to optimize the objective function for one-hot coding.

D. Experimental Results and Analyses on the CTW Dataset

The proposed LCSegNet is evaluated by comparing it with state-of-the-art methods and the different coding methods on the CTW dataset. In addition, other details of the proposed model are also given in this section.

1) Results Comparison with the State-of-the-art Methods: The LCSegNet is compared with the state-of-the-art systems in terms of the performance on the CTW dataset. The results are listed in Table I. All the test results on the CTW are obtained through the official online evaluation server[2]. The results show that when only the original training set is used, the proposed LCSegNet-Res2Net50 achieves the best result (87.74%) in recognition accuracy (higher than that of CNN-based methods by 7.24%). Furthermore, a comparison of the proposed method with the MtC-GAN [35] method that adds synthetic data showed an absolute performance gain of 4.24% in terms of recognition accuracy, ie, 25.70% of the decrease in the error. LCSegNet-Res2Net50 surpasses the style-melt nets SMN [73] by a large margin (8.14%). Compared with the radical-based network RAN [58], LCSegNet-Res2Net50 also brings an improvement of 2.18% over it. We also compare LCSegNet with object detection technology Faster R-CNN [75]. The result is obtained by the MMDetection [83] framework. For better detection, we expanded the image size of the CTW dataset to 512×640. Faster R-CNN [75] reached the performance of 80.54%, which is 7.20% lower than LCSegNet-Res2Net50 in accuracy. Because of the high computational cost and large scale of parameters, the FCN-8s [20] model encounters the problem of memory overflow on an NVIDIA V100 32G GPU. Compared with the FCN-ResNet50 model, under the same backbone, LCSegNet-ResNet50 achieves a performance improvement of 1.36%. In addition, we also conducted experiments to replace one-hot with Wubi-CRF in the general classification model. Comparing ResNet50 with ResNet50 (Wubi-CRF), we can find that Wubi-CRF brings more performance gain than one-hot. It shows the generality of the Wubi-CRF module, which can be integrated into semantic segmentation-based methods and non-semantic segmentation-based methods. The performance gain demonstrates the superiority of the proposed LCSegNet in dealing with complexities (such as occlusion, rotation, distortion, and complex background) in scene character recognition.

2) Results Comparison of Different Label Coding Methods: To verify the efficacy of coding methods, we compare the proposed Wubi-CRF coding with the three aforementioned coding methods. The results are listed in Table II. The Wubi-CRF approach surpasses the other coding methods in complex

TABLE I
COMPARISON TO EXISTING METHODS IN CHARACTER ACCURACY (%) ON THE OFFICIAL TEST SET OF THE CTW DATASET. "ORIGINAL" REPRESENTS OFFICIAL TRAINING DATA OF THE CTW DATASET. "SYNTHETIC" REPRESENTS THE SIMULATION DATA GENERATED BY THE GAN MODEL.

Method	Train set	Accuracy
†AlexNet	original	73.0
†OverFeat	original	76.0
†ResNet50	original	78.2
†ResNet152	original	79.0
†Google Inception	original	80.5
MtC-GAN-ResNet18 [35]	original+ synthetic	80.7
MtC-GAN-ResNet34 [35]	original+ synthetic	82.2
MtC-GAN-VGG16 [35]	original+ synthetic	83.5
SMN [73]	original	79.6
Faster R-CNN [75]	original	80.54
ResNet50 (Wubi-CRF)	original	81.97
RAN [58]	original	85.56
FCN-8s [20]	original	/
FCN-ResNet50	original	85.35
LCSegNet-ResNet50 (Ours)	original	**86.71**
LCSegNet-Res2Net50 (Ours)	original	**87.74**

† The results listed from the 1st to 5th rows are reported in [67].
/ Out of CUDA memory (32G).

[2]https://competitions.codalab.org/competitions/18634#learn_the_details-evaluation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3025696, IEEE Transactions on Multimedia

TABLE II
PERFORMANCES OF DIFFERENT CODING METHODS BY CHARACTER ACCURACY (%) ON THE CTW DATASET.

Methods	**Accuracy**
random binary coding	76.30
confusion matrix coding	76.85
one-hot coding	85.35
Wubi-CRF coding	**86.71**

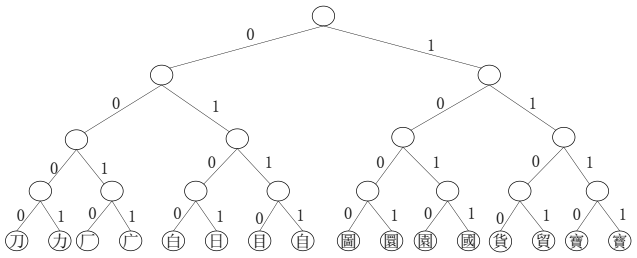


Fig. 6. Illustration of the binary tree of 16 characters that are easily confused on the CTW validation set. The left branch of each node is coded as 0, and the right branch is coded as 1. The more easily confused the characters are, the longer similar paths they have.

scene character recognition tasks. The main limitation of random binary coding is that it lacks the structural information of characters. To understand the results for confusion-matrix-based coding, as shown in Fig. 6, similar characters with high error rate on the validation set of CTW are selected as examples to construct the binary tree. From Fig. 6, we can see that, the more easily confused the characters are, the longer are their similar paths. In this method, the glyph information of characters is taken into account to some extent in label coding. However, it does not involves the specific structural information of characters. Hence, the coding method based on the confusion matrix only performs slightly better than random binary coding. One-hot coding also does not contain any character structure information, and with the increase in the number of categories, its computational cost and parameter scale rise sharply. Compared to the above coding methods, Wubi-CRF takes smaller stroke structures as units, and different parts have different coding. Hence, Wubi-CRF can better integrate the local structural information of characters. Therefore, the Wubi-CRF method not only improves the performance, but also reduces the computational cost and parameter scale.

3) Attributes and Character Sizes Analyses: The top-1 accuracy of the proposed LCSegNet for character images with different attributes and sizes is shown in Fig. 7. The 6 attributes include: occluded, complex background, distorted, 3D raised, word art, and handwritten. Fig. 7 shows that, the recognition performance is sorted by character size: recognition performance of large characters $>$ that of medium characters $>$ that of small characters. The main difficulties are posed by the ultra-low resolution and contrast of the small images. Besides, handwritten characters are more difficult to be recognized than those with other attributes characters. This is because the number of training samples for handwritten attributes is too small.

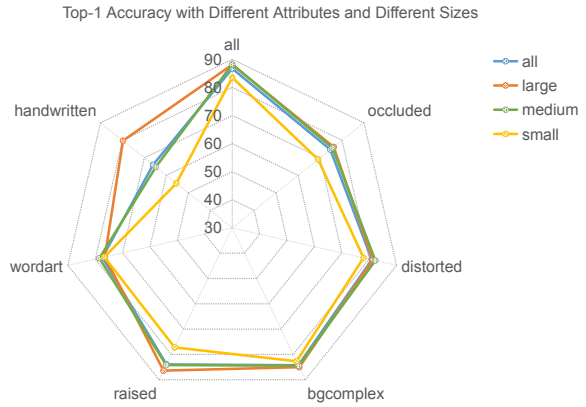
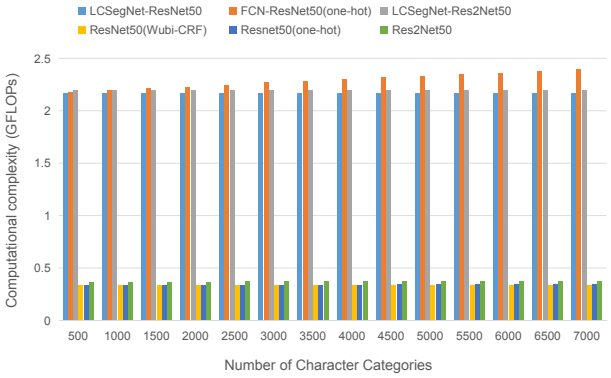
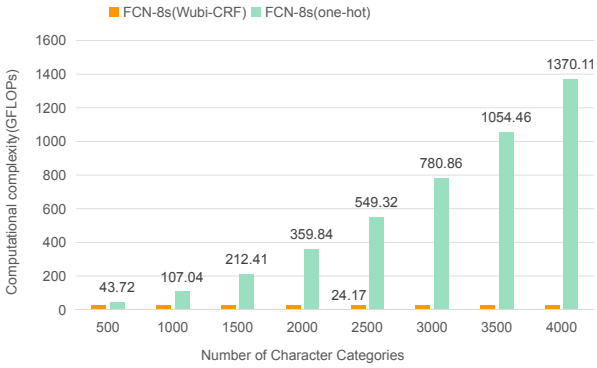


Fig. 7. The top-1 accuracy (%) of LCSegNet on the CTW dataset. Results for different attributes and sizes are shown. "all" means to include all attributes or character sizes. Small, medium and large indicate character size <16, $\in [16, 32)$ and ≥ 32, respectively.



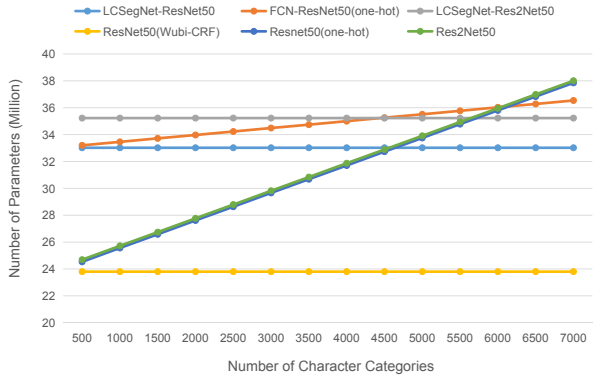
(a) Comparison of computational complexity between LCSegNet and other methods.



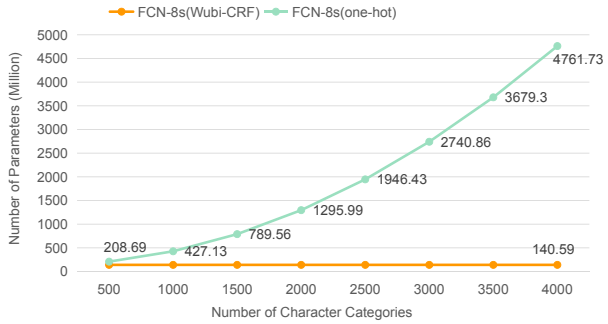
(b) Comparison of the computational complexity of Wubi-CRF coding and one-hot coding in the FCN-8s model.

Fig. 8. Increasing the computational complexity with increasing character categories for different methods. "FLOPs" stands for floating-point operations, which is a measure of the complexity of a model. 1 GFLOPs = 10^9 FLOPs.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3025696, IEEE Transactions on Multimedia



(a) Comparison of parameters between LCSegNet and other methods.



(b) Comparison of parameters between Wubi-CRF coding and one-hot coding in the FCN-8s model.

Fig. 9. The number of parameters corresponding to different character categories.

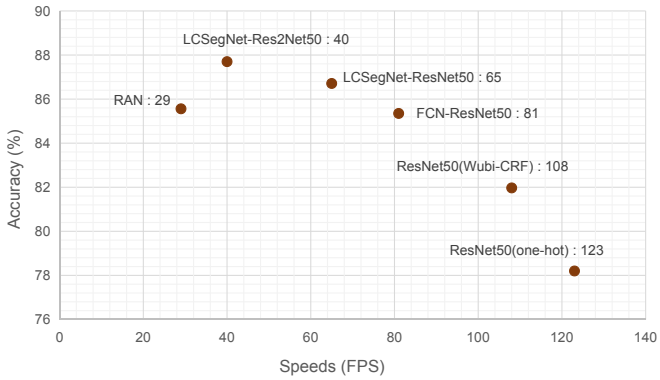


Fig. 10. The top-1 accuracy (%) vs running speeds (FPS) comparison on the CTW dataset. Results show that LCSegNet achieves state-of-the-art performance with the acceptable cost of inference time. "FPS" stands for frames per second.

TABLE III
COMPARISONS OF CHARACTER ACCURACY (%) WITH DIFFERENT UPSAMPLING METHODS AND CRF MODULE IN LCSEGNET ON THE CTW DATASET.

Module	Bilinear	Deconvolution
Without CRF	81.73	79.00
With CRF	**86.71**	85.41

4) Computational Complexity, Parameters and Speeds Analyses: To demonstrate the effectiveness of the proposed model, we compare the computational complexity, parameter scales, and running speeds with those of segmentation-based methods and methods without segmentation. For an input image of size $3 \times 64 \times 64$, the results are illustrated in Fig. 8 and Fig. 9, respectively. These two figures show that the computational cost and the number of parameters for both FCN-8s and FCN-ResNet50 increase sharply with the increase in the number of character categories, while those of the proposed LCSegNet remain constant. Compared to the FCN-ResNet50 model, Fig. 8(a) and Fig. 9(a) show that, the proposed method can reduce the computational cost by 5.24% and the number of parameters by 5.17%, while still achieving an increase of 1.36% in accuracy for the CTW dataset with 3650 character categories. Fig. 8(b) and Fig. 9(b) show that compared to one-hot coding, the proposed Wubi-CRF coding reduces the computational cost to 1/47 and compresses the parameter scale to 1/28 of that of the FCN-8s [20] model. Moreover, a comparison of LCSegNet and FCN-8s (one-hot) [20] in Fig. 8 and Fig. 9 shows that the proposed LCSegNet only uses the 0.19% computations and 0.83% parameters of the FCN-8s (one-hot) model to achieve state-of-the-art performance. For non-semantic segmentation methods, though their computational complexities are lower than those of methods based on semantic segmentation, the number of parameters gradually exceeds that of semantic segmentation-based methods with the increase in the number of character categories. In addition, the Wubi-CRF coding method can effectively reduce the number of parameters and the computational complexity for methods with or without semantic segmentation.

For a fair comparison, the average inference speed for all methods is measured on a single Tesla V100 GPU with batch size as 1. We use frames per second (FPS) as the evaluation metric. The results are shown in Fig. 10. Compared with the method based on semantic segmentation, the running speed of LCSegNet is 65 FPS, which is a bit slower than FCN-ResNet50 (one-hot), but much faster than FCN-8s (5FPS). Compared with methods without segmentation, both the speed and accuracy of LCSegNet surpass those of RAN [58]. In addition, the Wubi-CRF coding method can achieve a 3.77% performance gain on ResNet50 with a speed loss of only 15 PFS. The results show that LCSegNet uses a compromised speed to achieve the best accuracy.

5) Ablation Study: Ablation experiments are conducted to study the efficacy of each LCSegNet module in terms of recognition accuracy. The results of an ablation study of the CRF module and upsampling methods for the CTW dataset are list in Table III. The kernel size and stride of deconvolution are set to 16×16 and 8, respectively. The results show that when the CRF module is used, a performance gain of 4.98% is achieved. In addition, from the data in Table III, the performance of using deconvolution for upsampling is lower than bilinear interpolation. Therefore, bilinear interpolation is used for other related experiments.

6) Case Illustration: As shown in Fig. 11, the proposed model is robust to word art, complex background, occluded, distorted, 3D raised, handwritten, low-contrast and low-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3025696, IEEE Transactions on Multimedia

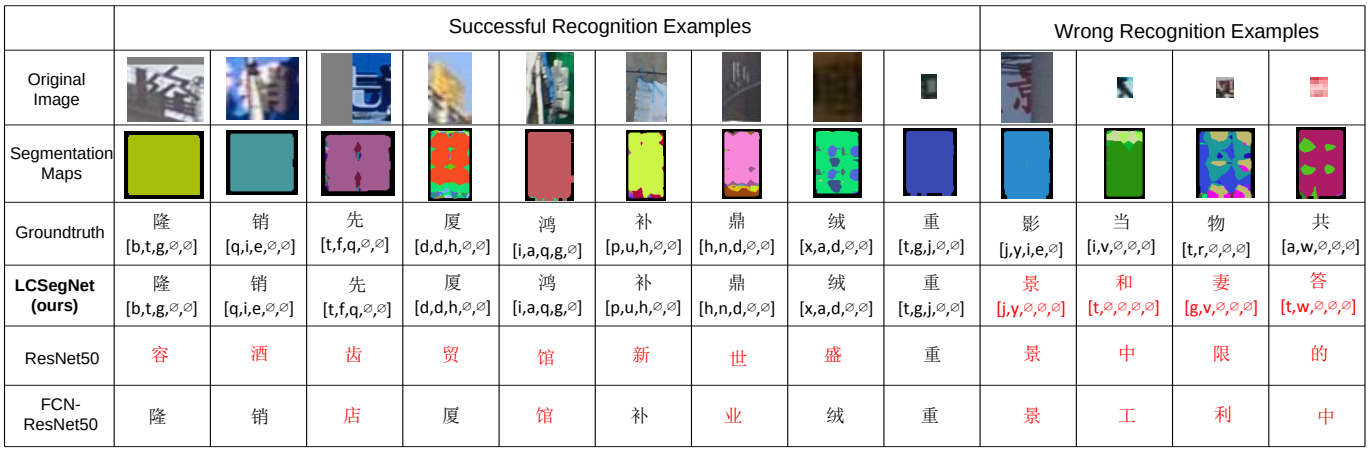


Fig. 11. Examples of recognition results for the CTW dataset. The left part shows some cases of successful recognition, and some examples of error in recognition are shown on the right. Red characters indicate recognition errors.

TABLE IV
COMPARISON OF WINNING METHODS IN CHARACTER ACCURACY (%) ON THE ICDAR2019-RECTS DATASET. "ADD DATA" DENOTES ADD SYNTHETIC DATA OR REAL EXTERNAL DATASET. "ENSEMBLE" DENOTES THE INTEGRATION OF MULTIPLE MODELS.

Ranking	Team	Add data	Ensemble	Accuracy
1	BASELINE-v1	**Yes**	**Yes**	97.37
2	Amap_CVLab	Yes	Unclear	97.28
	LCSegNet-data (ours)	Yes	No	96.30
3	TPS-ResNet-v1	Yes	No	96.12
4	SANHL_v4	No	Yes	95.94
	LCSegNet (ours)	**No**	**No**	95.25
5	Tencent-DPPR	Yes	Yes	95.12
6	ResNet_HUSTer	Yes	No	94.73
7	ReCTS_Task1	Unclear	Unclear	93.90
8	Task1-re5	Yes	Yes	93.89

resolution characters. Hence, it is preferable over the compared methods. The right-most columns in Fig. 11 show some samples of incorrect recognition. The first type of recognition error is caused by ambiguity after serious occlusion of characters. In the shown sample, the left parts of the characters "影" and "景" are the same. If the radical on the right part of "影" is occluded, it will be difficult to determine which character this image actually should be. Another important cause for the recognition errors is that the images with the extremely low resolution are seriously occluded.

E. Experimental Results and Analyses of the ICDAR2019-ReCTS Dataset

To further verify the effectiveness of the proposed LCSegNet in real scenes, we evaluate it on the ICDAR2019-ReCTS Dataset and compare it with those winning methods that participated in the challenge of reading Chinese character on the signboard.

1) Comparison Results with State-of-the-art Methods: The comparison results on the ICDAR2019-ReCTS dataset are listed in Table IV. When the proposed LCSegNet model with Res2Net50 backbone is trained with the original training set, it achieves the recognition accuracy of 95.25%. The single model without any external data ranks fifth in all contestants. Most top-ranked methods are trained with a large amount of external data or provide their ensemble versions. When the external data is used to train LCSegNet, higher accuracy of 96.30% is reached, which ranks third in all the competition methods. Compared with the methods that add large amounts of data, such as **TPS-ResNet-v1** with the synthetic dataset and real datasets (ArT, LSVT, ReCTS, and RCTW), and **Tencent-DPPR** with fifty million synthetic images and open-source datasets (LSVT, ReCTS, COCO-Text, RCTW, and ICPR-2018-MTWI), our model only adds a small amount of data to surpass them in accuracy. Our external data consists of three hundred thousand synthetic images and two hundred thousand images with 3D raised and word art attributes from the CTW dataset. In addition, the top-ranked methods provide their ensemble versions while LCSegNet is a single model. Adding more large-scale external data or adopting ensemble models may bring greater improvement for LCSegNet. Therefore, the results on the ICDAR2019-ReCTS dataset further prove the effectiveness of LCSegNet for complex scene recognition tasks and the scalability of non-Chinese character recognition (such as punctuation symbols).

2) Attention Visualization: Because the semantic segmentation-based model makes pixel-wise predictions and a semantic segmentation map is a form of visualization. In order to observe whether the predicted label coding focuses on the corresponding parts of the input character during the test phase, we use ResNet-50 model integrated with the Wubi-CRF coding method to conduct experiments and analyses on the ICDAR2019-ReCTS dataset. The examples of heatMap are shown in Fig. 12. Each heatMap corresponds to the label code under it. From the code "c" in the characters "双" and "妈", it can be seen that even though a single code may include diversity shape strokes, it will not disturb the stroke classification in the overall character recognition. For the same parts of different characters, the focused area corresponds to the code, such as parts "古" and "月" in "湖" and "胡". The results show that the predicted label coding can roughly focus on the corresponding parts of the input

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3025696, IEEE Transactions on Multimedia

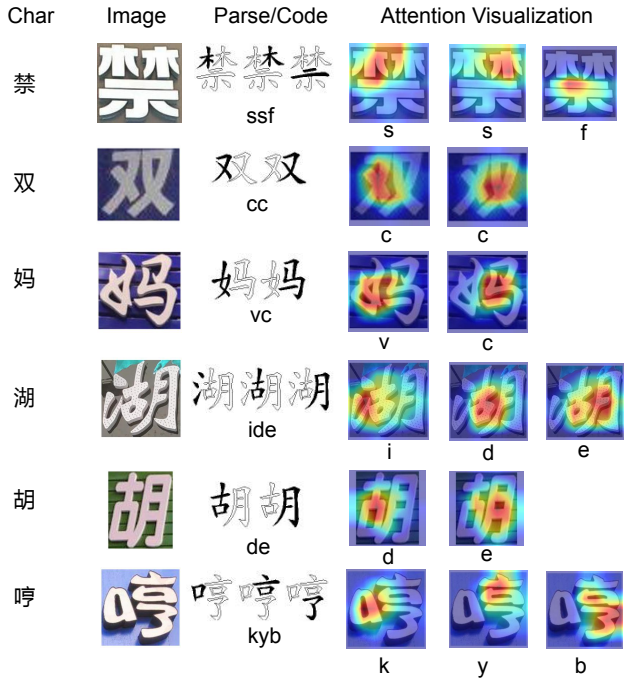


Fig. 12. Examples of the attention visualization on the ICDAR2019-ReCTS dataset. It shows that the predicted label codes can roughly focus on the corresponding parts of the input character image during the test phase.

TABLE V
COMPARISON OF EXISTING METHODS IN CHARACTER ACCURACY (%) ON THE HIT-OR3C DATASET. "GB1-W", "GB2-W", "D&L-W" AND "ALL-W" REPRESENT GB1 SUBSETS, GB2 SUBSETS, DIGIT&LETTER SUBSETS AND ALL SUBSETS OF 22 WRITER DATASETS, RESPECTIVELY. "DOC" INDICATES THE DOCUMENT SUBSET OF THE HIT-OR3C DATASET.

Method	GB1-W	GB2-W	D&L-W	ALL-W	Doc
KMean+MQDF [77]	77.71	85.67	-	-	71.58
HCCR-AlexNet [78]	97.85	98.67	84.16	98.09	93.57
ResNet50 [62]	96.75	97.94	81.23	97.14	91.54
SqueezeNet_BN [79]	97.24	98.25	78.08	97.51	92.30
HCTR-SRM [38]	95.26	96.65	-	-	86.40*
RAN [58]	97.99	98.72	-	-	90.81*
JSRAN [80]	98.01	98.69	-	-	90.80*
GhostNet-1.0 [81]	98.30	98.92	**84.16**	98.44	94.50
FCN-8s [20]	/	/	/	/	/
LCSegNet-ResNet50	**98.35**	**99.09**	83.14	**98.54**	**94.28**
LCSegNet-Res2Net50	**98.74**	**99.33**	81.67	**98.84**	**94.80**

/ Out of cuda memory (32G).
* Treat digits and letters as recognition errors.

TABLE VI
COMPARISONS OF CHARACTER ACCURACY (%) BETWEEN WITH AND WITHOUT THE CRF MODULE IN LCSEGNET ON THE HIT-OR3C DATASET.

Module	22-Writers	Documents
Without CRF	98.39	93.96
With CRF	**98.54**	**94.28**

character image. This is because LCSegNet uses the label coding to implicitly guide the attention to different parts of the character.

F. Experimental Results and Analyses of the Handwritten Dataset

To verify the capability of classifying large-scale categories and non-Chinese characters, the LCSegNet is also evaluated for a handwritten dataset: HIT-OR3C [69].

1) Comparison Results with State-of-the-art Methods: As seen from the data in Table V, on the HIT-OR3C dataset with 6825 categories, the proposed LCSegNet achieves state-of-the-art recognition accuracy on the 22 writers' dataset as well as the document dataset. Compared with JSRAN [80], the recognition error rate of LCSegNet-Res2Net50 in the GB1 subset is relatively reduced by 36.68%, and the error rate in the GB2 subset is relatively reduced by 48.85%. Compared with the latest network GhostNet-1.0 [81], which is based on the Ghost bottlenecks, LCSegNet also surpasses it on all subsets. We also reproduce the HCTR-SRM [38] method, which is based on segmentation and recognition modules. Because the HCTR-SRM [38] method is designed for text line recognition, we add a pooling operation in the final stage of each branch of HCTR-SRM to perform single character recognition. HCTR-SRM obtains an accuracy of 95.26% on the GB1 subset and 96.65% on the GB2 subset. LCSegNet also outperforms it on all subsets. Thus, the proposed LCSegNet is proven to be suitable for classifying super-large categories. Because RAN and JSRAN only recognize Chinese characters, on the doc subset, the digits and letters are regarded as the recognition errors. In the same evaluation metric, LCSegNet-Res2Net50 obtains an accuracy rate of 91.09% on the doc subset, still exceeding JSRAN and RAN. In addition, LCSegNet-ResNet50 achieves 83.14% accuracy in the digit and letter subsets of 22 writer datasets, which surpasses the ResNet50 and SqueezeNet_BN. It shows that LCSegNet can be extended to non-Chinese character recognition tasks.

2) Ablation Study: We also conduct ablation experiments on the HIT-OR3C dataset. Table VI shows that the LCSegNet with the CRF module performs better than the model without the CRF on the handwritten datasets, with a slight increase of 0.15% in recognition accuracy. Comparing the data in Table III and VI show that the CRF module is more effective for the complex scene task than for the handwritten task. The main reason is that for the base model, the more difficult a task is, the more are the error codes that the output may generate.

3) Case Illustration: Fig. 13 shows examples of correctly recognized characters. We can see that the proposed model is robust to the recognition of similar glyphs, which is not the case for popular deep-learning methods such as ResNet50, HCCR-AlexNet and SqueezeNe_BN. The possible reason is that the proposed network integrates the structural information of characters, and classifies each pixel of characters from a two-dimensional perspective. This allows it to distinguish similar characters better. Fig. 13 also shows examples of errors in recognizing characters that are scribbled and altered.

V. CONCLUSION

In this paper, we propose an effective semantic segmentation method based on label coding to recognize large-scale Chinese characters by pixel-wise prediction. We term this method

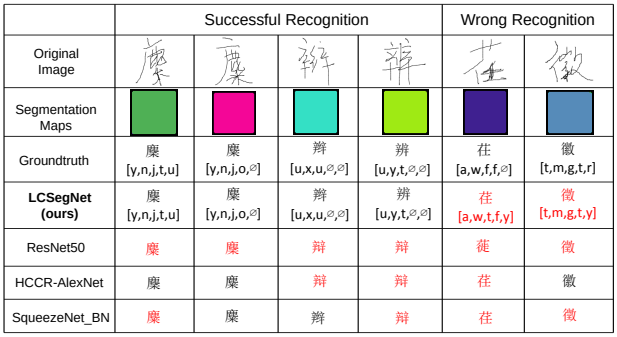


Fig. 13. Examples of the recognition results for the HIT-OR3C dataset. The left part shows correctly recognized examples. The right part shows wrongly recognized examples. Red characters indicate recognition errors.

LCSegNet. We design a new coding method called Wubi-CRF to encode the glyphs and structural information of Chinese characters into 140-bits labels, which deeply alleviates the limitations of computation and memory requirements for the classification of large-scale classes. Wubi-CRF coding method can be extended to non-Chinese characters and is compatible with semantic segmentation and non-semantic segmentation networks. Empirical evaluations demonstrate that the proposed LCSegNet achieves state-of-the-art performances for both complex scene character and handwritten character recognition tasks. It is robust to complex scenes, such as occluded, distorted, and handwritten text. In the future, we will focus on improving the proposed model to solve other challenges in text recognition.

REFERENCES

[1] B. Shi, C. Yao, M. Liao, M. Yang, P. Xu, L. Cui, S. Belongie, S. Lu, and X. Bai, "Icdar2017 competition on reading chinese text in the wild (rctw-17)," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 1429–1434.

[2] S. Karaoglu, R. Tao, T. Gevers, and A. W. Smeulders, "Words matter: Scene text for image classification and retrieval," *IEEE transactions on multimedia*, vol. 19, no. 5, pp. 1063–1076, 2016.

[3] F. Zhan and S. Lu, "Esir: End-to-end scene text recognition via iterative image rectification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2059–2068.

[4] X. Zhang, Y. Bengio, and C. Liu, "Online and offline handwritten chinese character recognition: A comprehensive study and new benchmark," *Pattern Recognition*, vol. 61, pp. 348–360, 2017.

[5] X. Wu, Q. Chen, J. You, and Y. Xiao, "Unconstrained off-line handwritten word recognition by position embedding integrated resnets model," *IEEE Signal Processing Letters*, 2019.

[6] J. Zhang, J. Du, and L. Dai, "Track, attend, and parse (tap): An end-to-end framework for online handwritten mathematical expression recognition," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 221–233, 2018.

[7] B. Su, X. Zhang, S. Lu, and C. L. Tan, "Segmented handwritten text recognition with recurrent neural network classifiers," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2015, pp. 386–390.

[8] S. Tian, S. Lu, B. Su, and C. L. Tan, "Scene text recognition using co-occurrence of histogram of oriented gradients," in *2013 12th International Conference on Document Analysis and Recognition*. IEEE, 2013, pp. 912–916.

[9] B. Su, S. Lu, S. Tian, J. H. Lim, and C. L. Tan, "Character recognition in natural scenes using convolutional co-occurrence hog," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 2926–2931.

[10] B. Su and S. Lu, "Accurate scene text recognition based on recurrent neural network," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 35–48.

[11] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[12] F. Sheng, C. Zhai, Z. Chen, and B. Xu, "End-to-end chinese image text recognition with attention model," in *International Conference on Neural Information Processing*. Springer, 2017, pp. 180–189.

[13] 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, 2016.

[14] X. Xiao, L. Jin, Y. Yang, W. Yang, J. Sun, and T. Chang, "Building fast and compact convolutional neural networks for offline handwritten chinese character recognition," *Pattern Recognition*, vol. 72, pp. 72–81, 2017.

[15] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[16] M. Liao, J. Zhang, Z. Wan, F. Xie, J. Liang, P. Lyu, C. Yao, and X. Bai, "Scene text recognition from two-dimensional perspective," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8714–8721.

[17] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7151–7160.

[18] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid, "Exploring context with deep structured models for semantic segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1352–1366, 2018.

[19] J. Fu, J. Liu, Y. Wang, J. Zhou, C. Wang, and H. Lu, "Stacked deconvolutional network for semantic segmentation," *IEEE Transactions on Image Processing*, 2019.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[21] 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)*, 2018, pp. 67–83.

[22] Z. Zhang, Y. Xu, and C.-L. Liu, "Natural scene character recognition using robust pca and sparse representation," in *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*. IEEE, 2016, pp. 340–345.

[23] S. Tian, U. Bhattacharya, S. Lu, B. Su, Q. Wang, X. Wei, Y. Lu, and C. L. Tan, "Multilingual scene character recognition with co-occurrence of histogram of oriented gradients," *Pattern Recognition*, vol. 51, pp. 125–134, 2016.

[24] B. Su and S. Lu, "Accurate recognition of words in scenes without character segmentation using recurrent neural network," *Pattern Recognition*, vol. 63, pp. 397–405, 2017.

[25] Y. Wang, C. Shi, C. Wang, B. Xiao, and C. Qi, "Multi-order co-occurrence activations encoded with fisher vector for scene character recognition," *Pattern Recognition Letters*, vol. 97, pp. 69–76, 2017.

[26] J. Bai, Z. Chen, B. Feng, and B. Xu, "Image character recognition using deep convolutional neural network learned from different languages," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 2560–2564.

[27] Y. Tang and X. Wu, "Scene text detection using superpixel-based stroke feature transform and deep learning based region classification," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2276–2288, 2018.

[28] Y. Wang, C. Shi, B. Xiao, and C. Wang, "Learning spatially embedded discriminative part detectors for scene character recognition," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 363–368.

[29] Z. Zhang, H. Wang, S. Liu, and B. Xiao, "Consecutive convolutional activations for scene character recognition," *IEEE Access*, vol. 6, pp. 35 734–35 742, 2018.

[30] A. K. Bhunia, A. K. Bhunia, P. Banerjee, A. Konwer, A. Bhowmick, P. P. Roy, and U. Pal, "Word level font-to-font image translation using convolutional recurrent generative adversarial networks," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3645–3650.

[31] A. K. Bhunia, A. Das, A. K. Bhunia, P. S. R. Kishore, and P. P. Roy, "Handwriting recognition in low-resource scripts using adversarial learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4767–4776.

[32] S. Wu, W. Zhai, and Y. Cao, "Pixtextgan: structure aware text image synthesis for license plate recognition," *IET Image Processing*, 2019.

[33] Y. Wang, Z. Lian, Y. Tang, and J. Xiao, "Boosting scene character recognition by learning canonical forms of glyphs," *International Journal on Document Analysis and Recognition (IJDAR)*, pp. 1–11, 2019.

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[35] Q. Lin, L. Liang, Y. Huang, and L. Jin, "Learning to generate realistic scene chinese character images by multitask coupled gan," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2018, pp. 41–51.

[36] Q.-F. Wang, F. Yin, and C.-L. Liu, "Handwritten chinese text recognition by integrating multiple contexts," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 8, pp. 1469–1481, 2011.

[37] X.-D. Zhou, D.-H. Wang, F. Tian, C.-L. Liu, and M. Nakagawa, "Handwritten chinese/japanese text recognition using semi-markov conditional random fields," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 10, pp. 2413–2426, 2013.

[38] D. Peng, L. Jin, Y. Wu, Z. Wang, and M. Cai, "A fast and accurate fully convolutional network for end-to-end handwritten chinese text segmentation and recognition," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 25–30.

[39] Z. Xie, Z. Sun, L. Jin, H. Ni, and T. Lyons, "Learning spatial-semantic context with fully convolutional recurrent network for online handwritten chinese text recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 8, pp. 1903–1917, 2017.

[40] 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, 2018.

[41] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1520–1528.

[42] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[43] C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A. L. Yuille, and L. Fei-Fei, "Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 82–92.

[44] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[46] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[47] M. Liao, P. Lyu, M. He, C. Yao, W. Wu, and X. Bai, "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," *CoRR*, vol. abs/1908.08207, 2019. [Online]. Available: http://arxiv.org/abs/1908.08207

[48] T. Wang and C. Liu, "Deepad: A deep learning based approach to stroke-level abnormality detection in handwritten chinese character recognition," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 1302–1307.

[49] T.-Q. Wang and C.-L. Liu, "Fully convolutional network based skeletonization for handwritten chinese characters," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[50] T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *Journal of artificial intelligence research*, vol. 2, pp. 263–286, 1994.

[51] N. Garcia-Pedrajas and D. Ortiz-Boyer, "Improving multiclass pattern recognition by the combination of two strategies," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 6, pp. 1001–1006, 2006.

[52] A. Rocha and S. K. Goldenstein, "Multiclass from binary: Expanding one-versus-all, one-versus-one and ecoc-based approaches," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 2, pp. 289–302, 2013.

[53] Q. Zhang, K. Lee, H. Bao, Y. You, W. Li, and D. Guo, "Large scale classification in deep neural network with label mapping," in *2018 IEEE International Conference on Data Mining Workshops, ICDM Workshops, Singapore, Singapore, November 17-20, 2018*, 2018, pp. 1134–1143.

[54] S. Lu, L. Li, and C. L. Tan, "Document image retrieval through word shape coding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1913–1918, 2008.

[55] S. Bai, L. Li, and C. L. Tan, "Keyword spotting in document images through word shape coding," in *2009 10th International Conference on Document Analysis and Recognition*. IEEE, 2009, pp. 331–335.

[56] S. Lu and C. L. Tan, "Retrieval of machine-printed latin documents through word shape coding," *Pattern Recognition*, vol. 41, no. 5, pp. 1799–1809, 2008.

[57] J. Zhang, Y. Zhu, J. Du, and L. Dai, "Radical analysis network for zero-shot learning in printed chinese character recognition," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[58] J. Zhang, J. Du, and L. Dai, "Radical analysis network for learning hierarchies of chinese characters," *Pattern Recognition*, p. 107305, 2020.

[59] X. Shi, J. Zhai, X. Yang, Z. Xie, and C. Liu, "Radical embedding: Delving deeper to chinese radicals," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2015, pp. 594–598.

[60] M. X. Tan, Y. Hu, N. I. Nikolov, and R. H. Hahnloser, "wubi2en: Character-level chinese-english translation through ascii encoding," *arXiv preprint arXiv:1805.03330*, 2018.

[61] F. Yang, J. Zhang, G. Liu, J. Zhou, C. Zhou, and H. Sun, "Five-stroke based cnn-birnn-crf network for chinese named entity recognition," in *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 2018, pp. 184–195.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[63] S. Gao, M. Cheng, K. Zhao, X. Zhang, M. Yang, and P. H. S. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[64] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[65] H. Zhang, H. Zhang, C. Wang, and J. Xie, "Co-occurrent features in semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 548–557.

[66] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1529–1537.

[67] T. Yuan, Z. Zhu, K. Xu, C. Li, T. Mu, and S. Hu, "A large chinese text dataset in the wild," *Journal of Computer Science and Technology*, vol. 34, no. 3, pp. 509–521, 2019.

[68] R. Zhang, Y. Zhou, Q. Jiang, Q. Song, N. Li, K. Zhou, L. Wang, D. Wang, M. Liao, M. Yang *et al.*, "Icdar 2019 robust reading challenge on reading chinese text on signboard," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 1577–1581.

[69] S. Zhou, Q. Chen, and X. Wang, "Hit-or3c: an opening recognition corpus for chinese characters," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*. ACM, 2010, pp. 223–230.

[70] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[71] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[72] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[73] W. Tang, Y. Jiang, N. Gao, J. Xiang, J. Shen, X. Li, and Y. Su, "Reducing style overfitting for character recognition via parallel neural networks with style to content connection," in *2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*. IEEE, 2019, pp. 784–791.

[74] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
[75] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.
[76] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
[77] S. Zhou, Q. Chen, X. Wang, X. Guo, and H. Li, "An empirical evaluation on hit-or3c database," in *2011 International Conference on Document Analysis and Recognition*. IEEE, 2011, pp. 1150–1154.
[78] Z. Zhong, L. Jin, and Z. Xie, "High performance offline handwritten chinese character recognition using googlenet and directional feature maps," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2015, pp. 846–850.
[79] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size," *CoRR*, vol. abs/1602.07360, 2016.
[80] C. Wu, Z. Wang, J. Du, J. Zhang, and J. Wang, "Joint spatial and radical analysis network for distorted chinese character recognition," in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 5, 2019, pp. 122–127.
[81] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1580–1589.
[82] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," in *Data mining and knowledge discovery handbook*. Springer, 2009, pp. 667–685.
[83] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "Mmdetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

Xiangping Wu received the M.S. degrees in computer science from Harbin Institute of Technology (Shenzhen), Shenzhen, China, in 2015. She is currently working toward the Ph.D. degree in the Shenzhen Chinese Calligraphy Digital Simulation Engineering Laboratory, Harbin Institute of Technology (Shenzhen), China. Her research interests include computer vision, machine learning, and pattern recognition.

Qingcai Chen received the M.S. and Ph.D. degrees from Harbin Institute of Technology, Harbin, China, in 1998 and 2003, respectively. He is currently a professor and the director of the Center for Intelligent Computing Research with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), China. He is also a PI of the Peng Cheng Laboratory, Shenzhen, China. His research interests include natural language processing, artificial intelligence, machine learning, financial and medical information processing.

Yulun Xiao received the B.S. degree in software engineering from the University of South China, Hengyang, China, in 2017. He is currently working toward the master degree in the Harbin Institute of Technology, Shenzhen, School of Computer Science and Technology. His research interests include text recognition and data mining.

Wei Li received the B.S. degree in software engineering from the Harbin Institute of Technology (Weihai), Shandong, China. He is currently pursuing the master degree in computer science and technology at the Harbin Institute of Technology (Shenzhen), Shenzhen, China. His current research interests are mainly including handwriting recognition, computer vision, and machine learning.

Xin Liu received the M.S. degree in computer science from Harbin Institute of Technology Shenzhen Graduate School, China, in Sept. 2015. He is currently pursuing the Ph.D. degree in computer science and technology at Harbin Institute of Technology (Shenzhen), China. His current research interests include deep learning, question answering, semantic representation learning, and handwritten recognition.

Baotian Hu received the M.S. and Ph.D. degrees in computer science from the Shenzhen Graduate School of Harbin Institute of Technology, Shenzhen, China, in 2012 and 2016, respectively. He is currently an Assistant Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen. His current research interests include deep learning and its application on natural language processing and image recognition.