Chinese Text Recognition with A Pre-Trained CLIP-Like Model Through Image-IDS Aligning
–Supplementary Material–

1. Choices of Hyperparameters

In this section, we present the experimental results of determining the appropriate hyperparameters for the proposed CCR-CLIP model. We conduct experiments on the printed artistic character dataset [3] for character zero-shot settings and the scene character dataset CTW [10] for non-zero-shot settings to choose λ, and on the handwriting dataset of the CTR benchmark [4] to determine β.

Choice of λ. We use two contrastive losses ($\mathcal{L}_T$ and $\mathcal{L}_I$) in the training stage of the proposed CCR-CLIP model, and λ is the hyperparameter that balances these two loss functions. Table 1 shows the experimental results for different values of λ ranging from 0 to 5. Based on our experimental results, we find that setting λ to 1 achieves the best performance. Furthermore, when λ is set to 0, which is the ablation study on λ, the performance of the CCR-CLIP model is clearly improved with $\lambda = 1$, validating the effectiveness of $\mathcal{L}_I$. Therefore, we set λ to 1 in pre-training experiments.

Choice of β. To prevent overfitting on seen characters, we introduce a regularization item in $\mathcal{L}_{tr}$. We conduct experiments on different values of β ranging from 0 to 1 and find that the proposed method achieves the highest performance when β is set to 0.001 on the CTR benchmark. Specifically, when β is set to 0, 0.001, 0.01, 0.1, and 1, the proposed method achieves 59.54%, 60.30%, 59.53%, 59.07%, and 58.76%, respectively. Therefore, we set β to 0.001 in all experiments on the CTR benchmark.

λ	m for Character Zero-Shot Setting					CTW
	500	1000	1500	2000	2755	
0	23.84%	48.13%	65.13%	72.33%	80.48%	83.29%
0.5	24.49%	48.20%	65.23%	73.55%	**81.90%**	84.86%
1	**25.00%**	**49.89%**	**65.25%**	**74.26%**	81.51%	**85.78%**
2	21.90%	48.62%	64.96%	72.60%	81.18%	83.12%
5	21.42%	46.85%	61.71%	71.60%	79.22%	83.06%

Table 1. Choice of λ.

2. Details of CTR Benchmark

The CTR benchmark comprises four distinct types of scenarios, namely, scene, web, document, and handwriting. Since the samples of these datasets are collected from various publicly available competitions, projects, and papers, some of the samples may contain non-Chinese characters. Therefore, in this paper, we filtered out such samples as our focus is on Chinese text recognition. Table 2 provides the statistical results of the four filtered datasets. It is worth noting that each of the four datasets includes some zero-shot characters, which pose a significant challenge for existing methods.

3. Examples of Adopted Datasets

In this paper, we evaluate the proposed method in Chinese character recognition and Chinese text recognition tasks, where four datasets (*i.e.*, HWDB1.0-1.1 [7], ICDAR2013 [9], CTW [10], and CTR benchmark [4]) are adopted. Some examples of these datasets are shown in Figure 1.

Dataset	Training	Validation	Test	Alphabet Size	ZS Characters
Scene	369085	45876	46062	5326	103
Web	52103	6585	6454	3843	81
Document	158317	20025	19905	4301	51
Handwriting	34830	8876	11018	5051	227

Table 2. The statistical results of four datasets. “ZS Characters” represents the number of zero-shot characters in the test dataset.

Figure 1. Examples of the adopted datasets.

4. More Experimental Results

In the Chinese character recognition task, we conduct additional zero-shot experiments to evaluate the effectiveness of the proposed CCR-CLIP model. We follow [3] to construct corresponding datasets for character zero-shot and radical zero-shot settings. For character zero-shot settings, we collect samples with labels falling in the first m classes as the training set and the last k classes as the test set. For the handwritten character dataset HWDB, m ranges in $\{500, 1000, 1500, 2000, 2755\}$ and k is set to 1000; for the scene character dataset CTW, m ranges in $\{500, 1000, 1500, 2000, 3150\}$ and k is set to 500. For radical zero-shot settings, we first calculate the frequency of each radical in the lexicon. Then the samples of characters that have one or more radicals appearing less than n times are collected as the test set, otherwise, collected as the training set, where n ranges in $\{10, 20, 30, 40, 50\}$ in radical zero-shot settings. It is important to note that even though radicals in the test set may be few-shot, we still use the term “radical zero-shot setting” in accordance with previous work [3].

The experimental results presented in Table 3 demonstrate that the proposed CCR-CLIP model outperforms the compared methods by a clear margin in both character zero-shot and radical zero-shot settings. This improvement can be attributed to the architecture of aligning IDSs and character images, which enables the model to better capture the discriminative features of characters. Furthermore, the introduction of contrastive loss $\mathcal{L}_I$ between the input images of the same character helps the feature extractor to focus on the texture of characters rather than complex backgrounds, resulting in further performance improvement. Compared with those methods that introduce template character images during training, the proposed CCR-CLIP model can still achieve the best performance (shown in Table 4).

5. Visualizations of Recognition Results and Failure Cases

In this section, we visualize some recognition results of the proposed method including results of CCR and CTR. Compared with decompose-based methods [3, 8], the proposed CCR-CLIP model is more robust to the characters with scribbled

HWDB	m for Character Zero-Shot Setting					n for Radical Zero-Shot Setting				
	500	1000	1500	2000	2755	50	40	30	20	10
DenseRAN [8]	1.70%	8.44%	14.71%	19.51%	30.68%	0.21%	0.29%	0.25%	0.42%	0.69%
HDE [2]	4.90%	12.77%	19.25%	25.13%	33.49%	3.26%	4.29%	6.33%	7.64%	9.33%
Chen et al. [3]	5.60%	13.85%	22.88%	25.73%	37.91%	5.28%	6.87%	9.02%	14.67%	15.83%
Ours	**21.79%**	**42.99%**	**55.86%**	**62.99%**	**72.98%**	**11.15%**	**13.85%**	**16.01%**	**16.76%**	**15.96%**
CTW	m for Character Zero-Shot Setting					n for Radical Zero-Shot Setting				
	500	1000	1500	2000	3150	50	40	30	20	10
DenseRAN [8]	0.15%	0.54%	1.60%	1.95%	5.39%	0%	0%	0%	0%	0.04%
HDE [2]	0.82%	2.11%	3.11%	6.96%	7.75%	0.18%	0.27%	0.61%	0.63%	0.90%
Chen et al. [3]	1.54%	2.54%	4.32%	6.82%	8.61%	0.66%	0.75%	0.81%	0.94%	2.25%
Ours	**3.55%**	**7.70%**	**9.48%**	**17.15%**	**24.91%**	**0.95%**	**1.77%**	**2.36%**	**2.59%**	**4.21%**

Table 3. The experimental results in the character zero-shot settings (left) and radical zero-shot settings (right). m represents that samples of the first m classes are used for training in the character zero-shot settings; n represents that samples with one or more radicals appearing less than n time are collected for testing in the radical zero-shot settings. These experiments do not involve additional template character images during training.

	m for Character Zero-Shot Setting (**HWDB**)					m for Character Zero-Shot Setting (**CTW**)				
	500	1000	1500	2000	2755	500	1000	1500	2000	3150
DMN [5]	66.33%	79.09%	84.14%	86.79%	88.98%	0.47%	1.20%	0.93%	1.60%	3.12%
CMPL [1]	72.49%	80.57%	84.40%	86.47%	89.29%	-	-	-	-	-
CCD [6]	90.93%	94.10%	94.58%	95.55%	-	58.22%	68.56%	74.45%	77.18%	-
Ours	**93.80%**	**94.97%**	**95.35%**	**95.71%**	**95.73%**	**62.13%**	**70.16%**	**75.88%**	**78.85%**	**80.03%**

Table 4. Comparison with previous methods in the case of using template character images during training.

strokes and complex backgrounds in the non-zero-shot setting, which benefits from the utilization of loss $\mathcal{L}_I$ between character images with the same label (shown in Figure 2). Additionally, we evaluate the proposed method on the CTR task and demonstrate its superior performance in recognizing zero-shot and few-shot Chinese characters, as shown in Figure 3.

As mentioned in the main text, the proposed method includes a pre-processing step where text images are rotated by 90 degrees anticlockwise if they are in a vertical orientation. Visualizations of failure cases shown in Figure 4 demonstrate that features of the same character in different orientations may cause confusion in the proposed model because it relies on canonical representation matching.

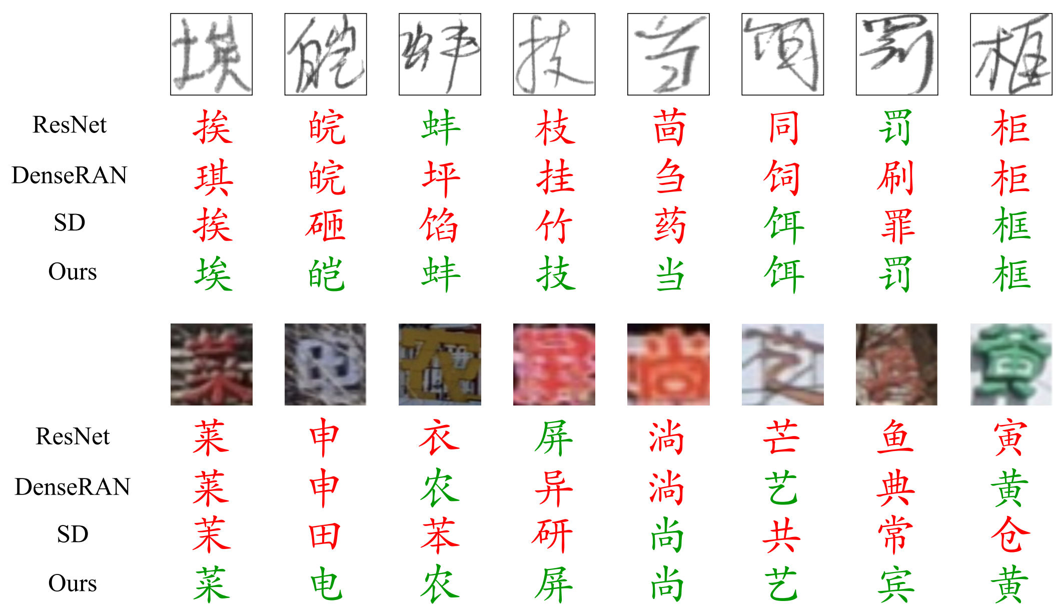

Figure 2. Recognition results of CCR.

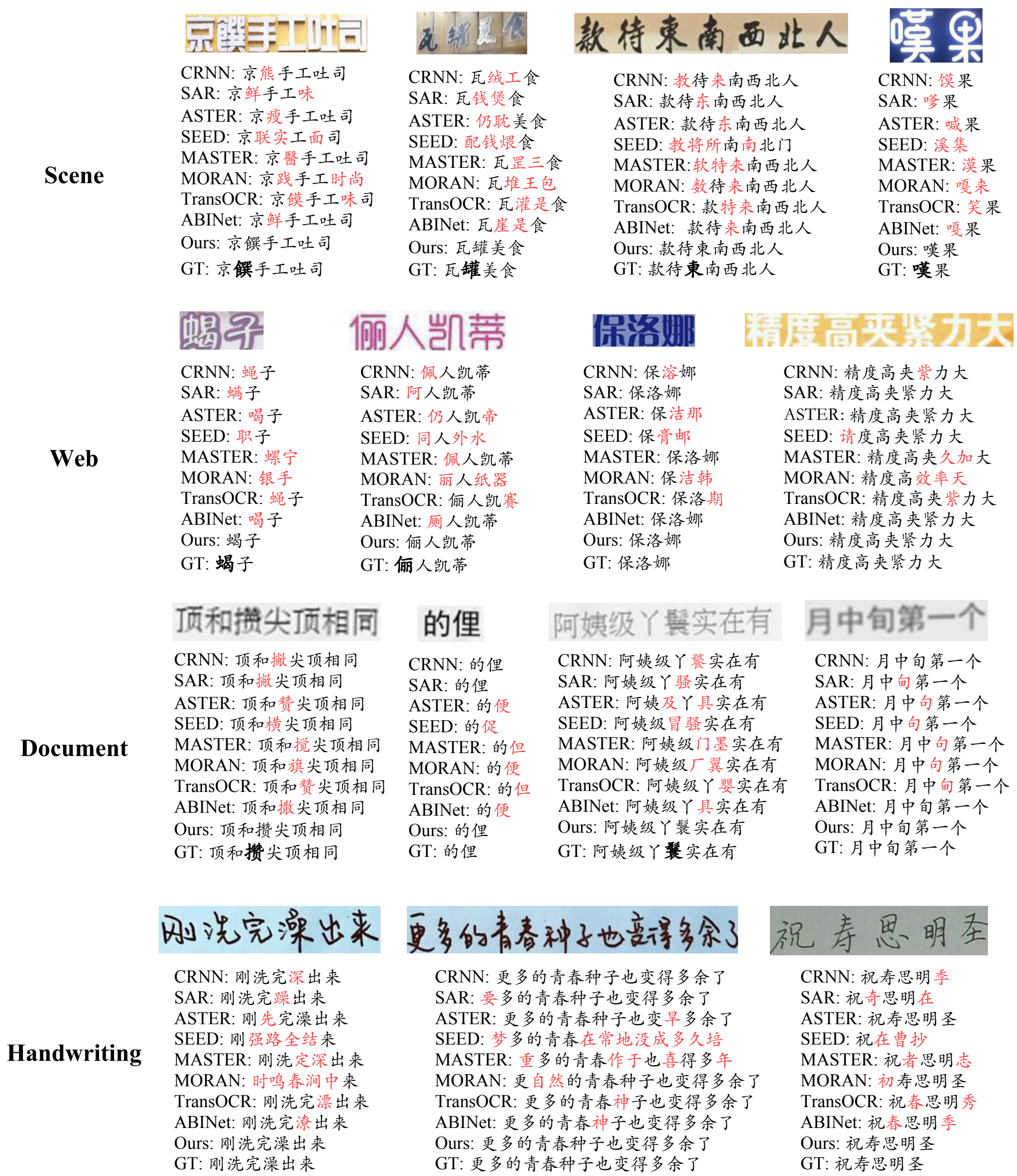

Figure 3. Recognition results of CTR. Red characters indicate wrongly predicted results, while bold characters represent zero-shot and few-shot ones in the training dataset.

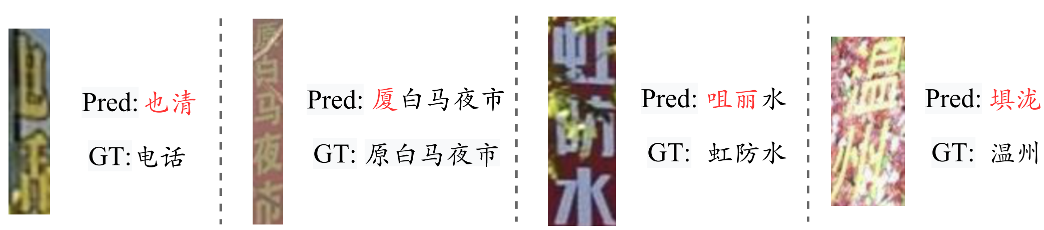

Figure 4. Visualizations of failure cases.

References

[1] Xiang Ao, Xu-Yao Zhang, and Cheng-Lin Liu. Cross-modal prototype learning for zero-shot handwritten character recognition. *Pattern Recognition*, 131:108859, 2022.

[2] Zhong Cao, Jiang Lu, Sen Cui, and Changshui Zhang. Zero-shot handwritten chinese character recognition with hierarchical decomposition embedding. *Pattern Recognition*, 107:107488, 2020.

[3] Jingye Chen, Bin Li, and Xiangyang Xue. Zero-shot chinese character recognition with stroke-level decomposition. *IJCAI*, 2021.

[4] Jingye Chen, Haiyang Yu, Jianqi Ma, Mengnan Guan, Xixi Xu, Xiaocong Wang, Shaobo Qu, Bin Li, and Xiangyang Xue. Benchmarking chinese text recognition: Datasets, baselines, and an empirical study. *arXiv preprint arXiv:2112.15093*, 2021.

[5] Zhiyuan Li, Qi Wu, Yi Xiao, Min Jin, and Huaxiang Lu. Deep matching network for handwritten chinese character recognition. *Pattern Recognition*, 107:107471, 2020.

[6] Chang Liu, Chun Yang, and Xu-Cheng Yin. Open-set text recognition via character-context decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4523–4532, 2022.

[7] Cheng-Lin Liu, Fei Yin, Da-Han Wang, and Qiu-Feng Wang. Online and offline handwritten chinese character recognition: benchmarking on new databases. *Pattern Recognition*, 46(1):155–162, 2013.

[8] Wenchao Wang, Jianshu Zhang, Jun Du, Zi-Rui Wang, and Yixing Zhu. Denseran for offline handwritten chinese character recognition. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 104–109. IEEE, 2018.

[9] Fei Yin, Qiu-Feng Wang, Xu-Yao Zhang, and Cheng-Lin Liu. Icdar 2013 chinese handwriting recognition competition. In *2013 12th international conference on document analysis and recognition*, pages 1464–1470. IEEE, 2013.

[10] Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, and Shi-Min Hu. A large chinese text dataset in the wild. *Journal of Computer Science and Technology*, 34(3):509–521, 2019.