Supplementary Material for One-stage Low-resolution Text Recognition with High-resolution Knowledge Transfer

Hang Guo[1], Tao Dai[2], Mingyan Zhu[1], GuangHao Meng[1,3],
Bin Chen[4], Zhi Wang[1], and Shu-Tao Xia[1,3]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]College of Computer Science and Software Engineering, Shenzhen University
[3]Research Center of Artificial Intelligence, Peng Cheng Laboratory
[4]Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China

A More Details on Adapting Different Recognizers to LR

In the main paper, we take three state-of-the-art scene text recognizers, namely ABINet [4], MATRN [7] and PARSeq [1], as examples and adapt them to low-resolution. Since the design of different models varies, we give the specific distillation details of different methods.

- For ABINet and MATRN, since both models use an additional language model for correction, we place the soft logits loss on the distribution of the language model output instead of the original distribution. This design allows the gradient flow to adapt the language model to character confusion at low resolution.
- Since PASeq uses permuted autoregression (one text image corresponds to multiple decoding paths), we propose a permuted distillation strategy on semantic contrastive loss and soft logits loss. Specifically, we take the average of the distillation losses for each permutation iteration as the final loss. This design allows for the implicit integration of linguistic knowledge into the model during distillation.

Despite these minor model-related differences above, the entire distillation framework remains consistent among different methods. In addition, these three picked recognizers (CNN based or ViT based) can cover most scenarios.

B Comparison with Distillation based Baselines

In this section, we design some powerful knowledge distillation based baselines to compare with our method. Since there are few works studying knowledge distillation for LTR tasks, we resort to other related distillation methods and adapt them to LTR. In specific, we refer to the following works. (i) The Deep Feature Distillation (DFD) [11] utilizes KD for low-resolution image classification and we use the loss in DFD and also add the cross-entropy loss. (ii) The Attention Similarity Knowledge Distillation (A-SKD) [8] employs the similarity between teacher and student attention map for low-resolution face recognition and we also add cross-entropy loss to A-SKD. (iii) Bhunia *et al.* [2] uses KD to unify scene text recognition and handwritten text recognition (Unified). Since this framework is designed for text data, we directly borrow the losses from it. (iv) The Dynamic Low-resolution Distillation (DLD) [3] adopts KD to achieve a better balance between accuracy and efficiency in text spotting and we also directly use the losses in DLD. All these modified distillation frameworks are applied to ABINet.

Table 1 shows the results. Interestingly, the designed distillation baselines all outperform the super-resolution based two-stage approaches, justifying the statement that using knowledge distillation for LTR tasks is a promising research direction. Moreover, the DFD and A-SKD methods, which focus on the recognition of the whole image and ignore the fine-grained sequential nature of text images, can only obtain limited performance. Although Unified and DLD are tailored to text data, they do not consider the resolution gap and also lack focus on character regions. By contrast, the proposed method is specifically designed for the LTR tasks and experimentally outperforms the other distillation based strong competitors.

Table 1. Comparison with knowledge distillation baselines.

Method	Recognition Accuracy↑			
	Easy	Medium	Hard	acgAcc
DFD [11]	83.32%	69.74%	53.24%	69.70%
A-SKD [8]	83.51%	68.89%	51.53%	68.97%
Unified [2]	84.93%	69.53%	54.43%	70.59%
DLD [3]	86.16%	71.16%	54.73%	71.67%
Ours	**86.91%**	**72.36%**	**55.10%**	**72.45%**

C Ablation on Hyper-parameters of Loss Function

Table 2 shows the impact of the different combinations of coefficients in the loss function on the performance. We only give experimental results about parameters λ_3 and λ_4 because we experimentally find other parameters are insensitive.

Table 2. Ablation on loss hyper-parameters. ABINet-LTR is used on the TextZoom [10] dataset.

λ_3	λ_4	Recognition Accuracy↑			
		Easy	Medium	Hard	avgAcc
0.025	15	86.16%	72.01%	**56.22%**	72.40%
0.025	25	85.61%	71.58%	55.62%	71.87%
0.1	20	85.67%	71.79%	55.99%	72.08%
0.01	20	85.55%	72.71%	55.70%	72.24%
0.025	20	**86.91%**	**72.36%**	55.10%	**72.45%**

D Extend to Other Low-resolution Text Recognition Task

We only show the application of the proposed distillation framework to scene text recognition in the main paper, however, as pointed out, the scheme can be applied to other text recognition tasks. Here, we initially explore the extension of low-resolution Handwritten Text Recognition (HTR) task. Specifically, we apply the proposed distillation framework to DAN [9] and obtain DAN-LTR. Since there are few low-resolution HTR datasets, we employ manual degradation on the IAM [6] datasets to obtain low-resolution images. Due to the few studies on handwritten text image super-resolution, we simply compare DAN-LTR with the Bicubic baseline. The results are shown in Table 3. It can be seen that the proposed method achieves a lower error rate. However, this is only a preliminary exploration, and we leave the extension of the proposed distillation framework on other LTR tasks for future work.

Table 3. Extend to low-resolution handwritten text recognition tasks. WER and CER denote Word Error Rate and Character Error Rate respectively.

Method	WER↓	CER↓
Bicubic	82.7%	56.8%
Ours	**59.9%**	**33.9%**

E More Details on Sequence Level Supervision

To enable the inclusion of sequence-level knowledge in the teacher distribution, we improve the word-level distribution of the soft teacher label by adding sequence-level knowledge. Here we give a rigorous mathematical derivation of sequence-level supervision [5].

Let $\boldsymbol{\pi}$ be all possible decoding paths obtained from the teacher distribution using beam search, and q_t^k be the probability that the student distribution predicts the k-th character at time step t. Note that this setting does not use the ground truth which only contains one path, because the purpose of logits loss is to make the model learn the similarity between different characters.

The objective function is defined as the negative path log-likelihood in the student distribution given the input image I:

$$\mathcal{L} = -\log P(\boldsymbol{\pi}|I) \tag{1}$$

Deriving the objective function to the output probability q_t^k gives:

$$\frac{\partial \mathcal{L}}{\partial q_t^k} = -\frac{1}{P(\boldsymbol{\pi}|I)} \frac{\partial P(\boldsymbol{\pi}|I)}{\partial q_t^k} \tag{2}$$

Substituting equation $P(\boldsymbol{\pi}|I) = \sum_{\pi} P(\pi|I) = \sum_{\pi} \prod_{t=1}^{T} q_t^{\pi_t}$ gives:

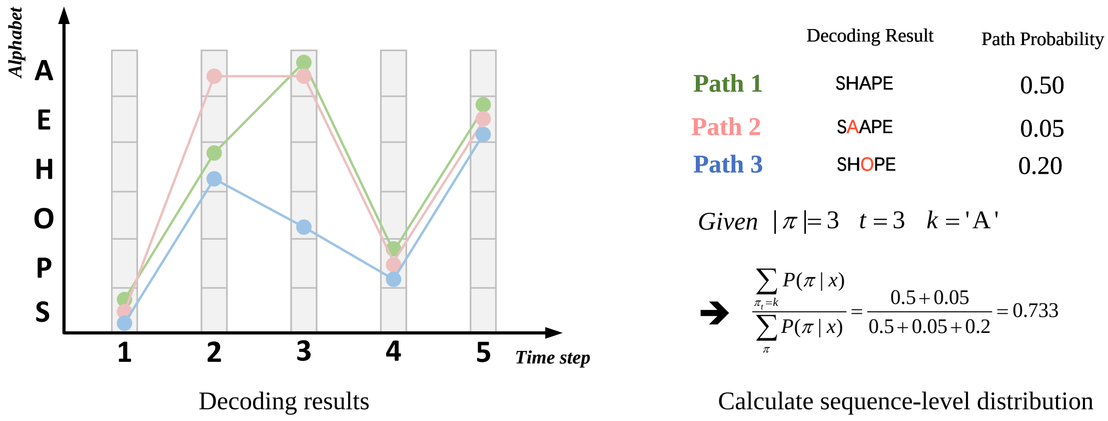

Fig. 1. An example of computing the sequence-level distribution. Suppose the number of paths in the beam search is 3 and the respective path likelihoods are 0.5, 0.05 and 0.2, then the probability of the sequence-level distribution at t=3 and character 'A' is 0.733.

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial q_t^k} &= -\frac{1}{P(\boldsymbol{\pi}|I)} \frac{\partial \sum_\pi \prod_{t=1}^T q_t^{\pi_t}}{\partial q_t^{\pi_t}} \\
&= -\frac{1}{P(\boldsymbol{\pi}|I)} \sum_\pi \frac{\partial \prod_{t=1}^T q_t^{\pi_t}}{\partial q_t^k} \\
&= -\frac{1}{P(\boldsymbol{\pi}|I)} \sum_{\pi_t=k} \frac{\prod_{t=1}^T q_t^{\pi_t}}{q_t^k}
\end{aligned}
\tag{3}
$$

Suppose the output logits in the student branch at the k-th character and t-h time step is a_t^k. Using the relationship between a_t^k and q_t^k, i.e. $q_t^k = \frac{\exp(a_t^k)}{\sum_{k'} \exp(a_t^{k'})}$, one can easily obtain the result of the derivation for softmax function:

$$
\frac{\partial q_t^{k'}}{\partial a_t^k} = \begin{cases} q_t^k \left(1 - q_t^k\right) & k = k' \\ -q_t^k \cdot q_t^{k'} & k \neq k' \end{cases}
\tag{4}
$$

Substituting the Eq. 4 into Eq. 3 yields:

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial a_t^k} &= \sum_{k'} \frac{\partial L}{\partial q_t^{k'}} \frac{\partial q_t^{k'}}{\partial a_t^k} \\
&= \sum_{k'} - \frac{1}{P(\boldsymbol{\pi}|I)} \sum_{\pi_t = k'} \frac{\prod_{t=1}^{T} q_t^{\pi_t}}{q_t^{k'}} \frac{\partial q_t^{k'}}{\partial a_t^k} \\
&= -\frac{1}{P(\boldsymbol{\pi}|I)} [\sum_{\pi_t=k} \frac{\prod_{t=1}^{T} q_t^{\pi_t}}{q_t^{k}} q_t^k (1 - q_t^k) + \sum_{k' \neq k} \sum_{\pi_t = k'} \frac{\prod_{t=1}^{T} q_t^{\pi_t}}{q_t^{k'}} (-q_t^{k'} q_t^k)] \\
&= -\frac{1}{P(\boldsymbol{\pi}|I)} [\sum_{\pi_t=k} \prod_{t=1}^{T} q_t^{\pi_t} (1 - q_t^k) + \sum_{k' \neq k} \sum_{\pi_t = k'} \prod_{t=1}^{T} q_t^{\pi_t} (-q_t^k)] \\
&= -\frac{1}{P(\boldsymbol{\pi}|I)} [\sum_{\pi_t=k} \prod_{t=1}^{T} q_t^{\pi_t} - q_t^k (\sum_{k'} \sum_{\pi_t = k'} \prod_{t=1}^{T} q_t^{\pi_t})] \\
&= -\frac{1}{P(\boldsymbol{\pi}|I)} [\sum_{\pi_t=k} \prod_{t=1}^{T} q_t^{\pi_t} - q_t^k \sum_{\pi} \prod_{t=1}^{T} q_t^{\pi_t}] \\
&= -\frac{1}{P(\boldsymbol{\pi}|I)} [\sum_{\pi_t=k} P(\pi|I) - q_t^k \sum_{\pi} P(\pi|I)] \\
&= q_t^k - \frac{\sum_{\pi_t=k} P(\pi|I)}{\sum_{\pi} P(\pi|I)}
\end{aligned}
\tag{5}
$$

It can be seen that optimizing the loss function after considering the sequence modeling is equivalent to making the probability of the student distribution fit the second term in Eq. 5. For the convenience of implementation, we use the teacher path likelihood which shares the same decoding path as an approximation to the student path likelihood. In this setting, sequence-level knowledge can be learned by modifying the teacher distribution and supervising training with KL divergence. An example of calculating the sequence-level distribution is shown in Fig. 1.

F Details of Manual Image Degradation on STR datasets

We use the STR benchmarks for robustness test in the main paper, however some of the images in these datasets are of high quality, so we perform manual degradation. Here we provide details of this degradation for reproduction. Specifically, we perform the degradation process using the `imgaug` library with the following code:

```
import imgaug.augmenters as iaa
sometimes = lambda aug: iaa.Sometimes(0.25, aug)
aug_list = [
    iaa.GaussianBlur(sigma=(0.0, 3.0)),
    iaa.AverageBlur(k=(1, 5)),
    iaa.MedianBlur(k=(3, 7)),
    iaa.BilateralBlur(d=(3, 9), sigma_color=(10, 250), sigma_space=(10, 250)),
    iaa.MotionBlur(k=3),
    iaa.MeanShiftBlur(),
    iaa.Superpixels(p_replace=(0.1, 0.5), n_segments=(1, 7)),
    iaa.AdditiveGaussianNoise(loc=0, scale=(0.0, 0.05 * 255), per_channel=0.5),
]
aug = iaa.Sequential([sometimes(a) for a in aug_list], random_order=True)
degraded_img = aug(img)
```

References

1. Bautista, D., Atienza, R.: Scene text recognition with permuted autoregressive sequence models. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII. pp. 178–196. Springer (2022)
2. Bhunia, A.K., Sain, A., Chowdhury, P.N., Song, Y.Z.: Text is text, no matter what: Unifying text recognition using knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 983–992 (2021)
3. Chen, Y., Qiao, L., Cheng, Z., Pu, S., Niu, Y., Li, X.: Dynamic low-resolution distillation for cost-efficient end-to-end text spotting. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII. pp. 356–373. Springer (2022)
4. Fang, S., Xie, H., Wang, Y., Mao, Z., Zhang, Y.: Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7098–7107 (2021)
5. Huang, M., You, Y., Chen, Z., Qian, Y., Yu, K.: Knowledge distillation for sequence model. In: Interspeech. pp. 3703–3707 (2018)
6. Marti, U.V., Bunke, H.: The iam-database: an english sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition **5**, 39–46 (2002)
7. Na, B., Kim, Y., Park, S.: Multi-modal text recognition networks: Interactive enhancements between visual and semantic features. In: European Conference on Computer Vision. pp. 446–463. Springer (2022)
8. Shin, S., Lee, J., Lee, J., Yu, Y., Lee, K.: Teaching where to look: Attention similarity knowledge distillation for low resolution face recognition. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII. pp. 631–647. Springer (2022)
9. Wang, T., Zhu, Y., Jin, L., Luo, C., Chen, X., Wu, Y., Wang, Q., Cai, M.: Decoupled attention network for text recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12216–12224 (2020)
10. Wang, W., Xie, E., Liu, X., Wang, W., Liang, D., Shen, C., Bai, X.: Scene text image super-resolution in the wild. european conference on computer vision (2020)
11. Zhu, M., Han, K., Zhang, C., Lin, J., Wang, Y.: Low-resolution visual recognition via deep feature distillation. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3762–3766. IEEE (2019)