One-stage Low-resolution Text Recognition with High-resolution Knowledge Transfer

Hang Guo
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
cshguo@gmail.com

Tao Dai*
College of Computer Science and Software Engineering, Shenzhen University
Shenzhen, China
daitao.edu@gmail.com

Mingyan Zhu
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
zmy20@mails.tsinghua.edu.cn

GuangHao Meng
Tsinghua Shenzhen International Graduate School, Tsinghua University
Research Center of Artificial Intelligence, Peng Cheng Laboratory
Shenzhen, China
mgh19@mails.tsinghua.edu.cn

Bin Chen
Department of Computer Science and Technology, Harbin Institute of Technology
Shenzhen, China
chenbin2021@hit.edu.cn

Zhi Wang
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
wangzhi@sz.tsinghua.edu.cn

Shu-Tao Xia
Tsinghua Shenzhen International Graduate School, Tsinghua University
Research Center of Artificial Intelligence, Peng Cheng Laboratory
Shenzhen, China
xiast@sz.tsinghua.edu.cn

ABSTRACT

Recognizing characters from low-resolution (LR) text images poses a significant challenge due to the information deficiency as well as the noise and blur in low-quality images. Current solutions for low-resolution text recognition (LTR) typically rely on a two-stage pipeline that involves super-resolution as the first stage followed by the second-stage recognition. Although this pipeline is straightforward and intuitive, it has to use an additional super-resolution network, which causes inefficiencies during training and testing. Moreover, the recognition accuracy of the second stage heavily depends on the reconstruction quality of the first stage, causing ineffectiveness. In this work, we attempt to address these challenges from a novel perspective: adapting the recognizer to low-resolution inputs by transferring the knowledge from the high-resolution. Guided by this idea, we propose an efficient and effective knowledge distillation framework to achieve multi-level knowledge transfer. Specifically, the visual focus loss is proposed to extract the character position knowledge with resolution gap reduction and character region focus, the semantic contrastive loss is employed to exploit the contextual semantic knowledge with contrastive learning, and the soft logits loss facilitates both local word-level and global sequence-level learning from the soft teacher label. Extensive experiments show that the proposed one-stage pipeline significantly outperforms super-resolution based two-stage frameworks in terms of effectiveness and efficiency, accompanied by favorable robustness. Code is available at https://github.com/csguoh/KD-LTR.

CCS CONCEPTS

• **Computing methodologies** → *Object recognition*; • **Computer systems organization** → Neural networks.

KEYWORDS

low-resolution, text recognition, knowledge distillation

*Corresponding author: Tao Dai.
This work is supported in part by the National Key Research and Development Program of China, under Grant 2022YFF1202104, National Natural Science Foundation of China, under Grant (6230070671, 62171248), Shenzhen Science and Technology Program (Grant No.RCYX20200714114523079, JCYJ20220818101014030,JCYJ20220818101012025), and the PCNL KEY project (PCL2021A07).

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/XXXXXX.XXXXXX

ACM Reference Format:
Hang Guo, Tao Dai, Mingyan Zhu, GuangHao Meng, Bin Chen, Zhi Wang, and Shu-Tao Xia. 2023. One-stage Low-resolution Text Recognition with High-resolution Knowledge Transfer. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

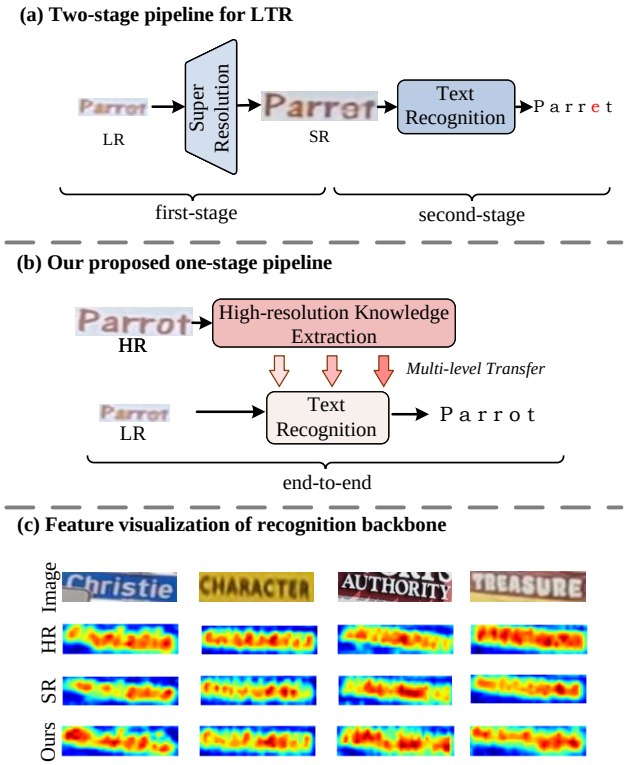

Figure 1: (a) The cascading in the two-stage pipeline leads to inefficiency and the error accumulation affects the effectiveness. For example, 'Parrot' is incorrectly reconstructed as 'Parret' in the first stage which then leads to subsequent misrecognition. (b) The proposed framework extracts multi-level knowledge from high-resolution images and transfers it to the recognizer. (c) The features of SR and HR differ in the character regions, e.g. the 'i' in 'Christie'.

1 INTRODUCTION

Scene text recognition (STR) has become increasingly popular in recent years due to its various applications, such as license plate recognition [26, 30], autonomous driving [5], and so on. However, current STR methods suffer from significant performance degradation when recognizing low-resolution images [8, 23, 51].

To address this issue, the mainstream approaches have adopted a two-stage pipeline (see Fig. 1 (a)). They split the whole LTR tasks into two separate tasks, generating recognition-friendly text images in the first stage, followed by common text recognition in the second stage. Guided by this two-stage strategy, several pioneering works [13, 47, 55] employ the generic single-image super-resolution model to generate SR for recognition. Recently, designing the text-oriented super-resolution model, also known as Scene Text Image Super-Resolution (STISR), has attracted the interest of many researchers [8, 9, 33–35, 37, 41, 51, 52, 59, 60, 62]. For example, the TextZoom dataset [51] has been introduced to facilitate real-world STISR research. Furthermore, recent STISR methods [33, 34, 60] have utilized pre-trained text recognizers [43] to inject linguistic knowledge as prior guidance for the super-resolution blocks. Due to the intuitiveness and simplicity, despite some alternative solutions such as multi-task learning [23, 35] have been developed, this two-stage pipeline is still prevalent.

As more advanced STISR models continue to be proposed, progress has been made in this two-stage framework. However, this two-stage pipeline also poses certain challenges. First, the two-stage approach necessitates an additional super-resolution network, resulting in inevitably high computational costs for both training and inference. For example, the current state-of-the-art STISR model [60] is even larger than the recognition model [3]. Moreover, due to severe information loss and noise in LR, even the use of text-customized super-resolution models may lead to wrong reconstruction and this reconstruction error will further be amplified due to the cascading design. We also visualize the features extracted from SR and HR in the recognizer backbone (see Fig. 1 (c)). It can be seen that even though the super-resolution model attempts to imitate HR in the pixel space, there are still differences between both in the feature space.

To break the limitations brought by the two-stage approach, this work explores a one-stage solution by directly adapting the text recognizer to low-resolution inputs without any super-resolution as pre-processing (see Fig. 1 (b)). In concrete, we design a LTR-customized knowledge distillation paradigm to mine the knowledge contained in high-resolution images and transfer it to the text recognizer. The key to designing such a paradigm is to find what knowledge should be used and how to transfer this knowledge. To this end, we develop KD-LTR, a novel Knowledge Distillation based framework for LTR, which can help the recognizer learn from high-resolution images. The proposed distillation pipeline extracts three distinct levels of knowledge to facilitate knowledge transfer. Specifically, we employ the visual focus loss to mine the character position knowledge using the resolution gap reduction techniques and mask distillation strategy. In addition, we introduce semantic contrastive loss which uses the contrastive learning scheme to enable the recognizer to acquire contextual semantic knowledge. Finally, the knowledge contained in the soft teacher label is learned from both local word-level and global sequence-level perspectives through the proposed soft logits loss. By leveraging the extensive knowledge encapsulated in high-resolution images from multiple levels, our approach achieves faster and more accurate performance than the super-resolution based two-stage framework and can be easily applied to various text recognition models.

Overall, our main contributions are three folds:

- We propose the first one-stage pipeline for LTR, which adapts the text recognizer to low-resolution inputs by transferring knowledge from high-resolution images.
- We propose three well-designed distillation losses to facilitate multi-level knowledge transfer.
- Extensive experiments show that the proposed one-stage pipeline sets new state-of-the-art for LTR tasks in terms of efficiency and effectiveness.

2 RELATED WORK

Although many vision tasks, including image classification[18, 46], object detection [6, 42], face recognition [2, 12], and text recognition

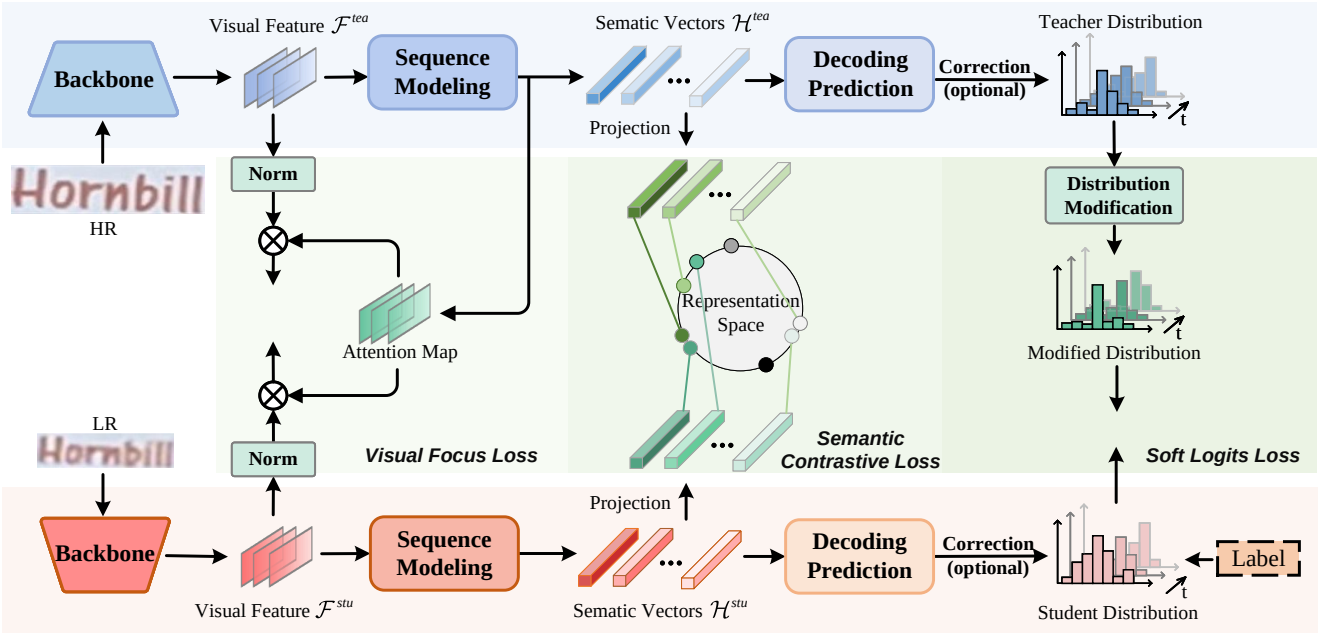

Figure 2: An overview of the proposed framework. The branch with blue background is the HR teacher and the branch with red background is the LR student. Three levels of loss are used to extract and transfer the knowledge in high-resolution images.

[3, 15] have achieved remarkable success, current methods still struggle with significant performance degradation when confronted with low-resolution images. Depending on the way in which the HR prior is transferred to the LR images, current solutions to this problem can be classified into two types: super-resolution based and knowledge distillation based approaches. Super-resolution based approaches [7, 16, 48, 52] learn the HR prior in the pixel space by employing an additional super-resolution model before visual recognition. In contrast, knowledge distillation based approaches [17, 39] transfer knowledge from HR teacher to LR student in the feature space using losses from different perspectives.

Super-resolution for Low-resolution Recognition. A natural idea to handle the low-resolution visual recognition task is to enhance the original low-quality input to an easily recognizable one by pre-processing. Guided by this idea, some early works borrow from generic Single Image Super-Resolution (SISR) [11, 29, 58] models to super-resolve the LR to SR before recognition. However, since these super-resolution models are usually trained with image quality as the objective, their effectiveness is limited due to the mismatch in objectives with the recognition task. Recently, there has been a surge of interest in designing task-specific super-resolution models. In the context of scene text recognition, several scene text image super-resolution methods have been proposed, which have shown promising results. For instance, TextSR [52] utilizes a GAN-based architecture to generate recognition-friendly SR images. To facilitate real-world STISR research, the TextZoom dataset [51] was introduced, accompanied by TSRN which considers the sequential nature of text image data. Moreover, STT [8] implicitly makes the model focus on the character regions by designing relevant loss functions. TATT [34] achieves spatial deformation robust STISR by using the proposed TP Interpreter. C3-STISR [60] uses three-level clues to guide the super-resolution block and obtains favorable results.

Knowledge Distillation for Low-resolution Recognition. The concept of Knowledge Distillation (KD) was first introduced by Hinton *et al.* [19] to transfer knowledge from the over-parameterized teacher to the compact student. More recently, resolution distillation has been developed to address the challenges of low-resolution visual recognition tasks. Unlike traditional KD, which focuses on model compression, resolution distillation transfers knowledge from the HR teacher to the LR student, enabling the student to recover lost information in LR with supervision from different perspectives. While this pipeline has been proven to be effective in image classification [61], object detection [39], and face recognition [44], it has not been explored for low-resolution text recognition task, whose data involves sequential nature.

3 METHOD

3.1 Overview

As shown in Fig. 2, we propose a knowledge distillation framework that can extract different levels of high-resolution knowledge and transfer them to a text recognizer to achieve low-resolution text recognition without additional super-resolution modules. The proposed framework consists of two branches: the HR teacher branch and the LR student branch. The HR teacher branch, which can be obtained from any off-the-shelf text recognizer, takes HR as input and is frozen during training to ensure optimal knowledge transfer. Meanwhile, the LR student branch works with low-resolution text images and aims to recover lost details in the LR input. We begin by

reviewing the generic framework for text recognition (Section 3.2). Then, we present the details of the loss functions in the proposed distillation framework (Section 3.3).

3.2 Base Text Recognizer

The text recognizer used in both the student and teacher branches exploits the prevalent encoder-decoder framework [3, 15, 36]. As shown in each branch of Fig. 2, the text recognition model is broadly divided into three parts: the backbone for feature extraction, the sequence modeling module, and the decoding prediction module. Given input text images I, the feature extraction backbone (Resnet [18] or ViT [14]) first extracts from I to obtain the visual features $\mathcal{F} \in \mathbb{R}^{N\times C\times H\times W}$, where N is the batch size, C is the number of channels, and H and W are the height and width of the features, respectively. Subsequently, the sequence modeling module captures the contextual dependencies with LSTM [20] or Transformer [49] to project the 2D visual features $\mathcal{F}$ into 1D semantic vectors $\mathcal{H} \in \mathbb{R}^{N\times T\times C}$, where T is the predefined maximum length of the character sequence. The $\mathcal{H}$ is then transformed into the character probability distribution by the decoding prediction module, which consists of projection layers and a softmax activation function. Linguistic knowledge can be optionally incorporated via an additional language model [15, 36].

3.3 High-resolution Knowledge Transfer

To efficiently transfer knowledge from high-resolution images, we propose three levels of loss functions. The first level, visual focus loss, is proposed to extract character position knowledge from the visual features. It can bridge the resolution gap between the two branches and give the character region more focus. The second level, semantic contrastive loss, is exploited to extract contextual semantic knowledge. It facilitates the generation of distinct semantic vectors by leveraging contrastive learning. Finally, the third level, soft logits loss, combines both word-level and sequence-level knowledge from the soft teacher label to produce meaningful recognition results. Further details of these three losses are as follows.

3.3.1 Visual Focus Loss. The visual features extracted from the backbone in the teacher branch contain rich character position knowledge to help the student model recover critical character region features which are useful for subsequent recognition. However, due to the resolution gap of the inputs between the two branches, the visual features of the two are not inherently identical. As such, aligning the student and teacher visual features with absolute numerical measures (e.g., L1 or L2) would be counterproductive. Following [44], we use the cosine similarity to measure the directional consistency between features. Furthermore, since statistics such as the mean and variance of a feature map can represent the corresponding domain [32], we thus first normalize visual features to achieve resolution-domain removal:

$$\tilde{\mathcal{F}} = \frac{\mathcal{F} - \mu}{\sigma}, \tag{1}$$

where $\mathcal{F}$ is a unified notation of the teacher features $\mathcal{F}^{tea}$ and the student features $\mathcal{F}^{stu}$. μ and $\sigma \in \mathbb{R}^{N\times C}$ are the results of global average pooling and standard deviation pooling of $\mathcal{F}$.

Then the cosine similarity can be used to measure the distance between the normalized features:

$$\mathcal{L}'_{visual} = 1 - \frac{1}{N}\frac{1}{C}\sum_{i=1}^{N}\sum_{j=1}^{C}\langle \tilde{\mathcal{F}}^{tea}_{i,j}, \tilde{\mathcal{F}}^{stu}_{i,j}\rangle, \tag{2}$$

where $\langle\cdot,\cdot\rangle$ denotes the vector cosine-similarity.

Further, based on the observation in Fig. 1 (c), character features are more difficult to learn as well as more important for recognition than background features. Drawing inspiration from [56], we utilize the mask distillation strategy to make the student more focused on the reconstruction of character features while blocking out the disturbance from irrelevant background noise. Specifically, we use the attention map from the teacher branch as a soft mask to reassign the weights of different pixels. Then the visual focus loss with mask distillation can be written as follow:

$$\mathcal{L}_{visual} = 1 - \frac{1}{N}\frac{1}{C}\sum_{i=1}^{N}\sum_{j=1}^{C}\langle \mathcal{M}\tilde{\mathcal{F}}^{tea}_{i,j}, \mathcal{M}\tilde{\mathcal{F}}^{stu}_{i,j}\rangle, \tag{3}$$

where $\mathcal{M}$ denotes the mask from the teacher attention map. Since most of the popular recognizers [15, 36, 40, 53, 54, 57] are attention-based, it is easy to obtain the attention map.

3.3.2 Semantic Contrastive Loss. While the visual focus loss can assist in recovering the lost spatial details in LR, the semantic vectors generated from the sequence modeling module can provide contextual information which is useful for sequential tasks. We therefore exploit this contextual semantic knowledge contained in the HR teacher's semantic vectors $\mathcal{H}^{tea}$.

Specifically, we first obtain the corresponding semantic vector of each character from the recognizer's sequence modeling module:

$$\mathcal{H} = \text{softmax}(\frac{QK^T}{\sqrt{C}})V, \tag{4}$$

where Q is the position query of character orders. K and V are the key and value generated from visual features $\mathcal{F}$. The above attentional sequence modeling can make the i-th element h_i in $\mathcal{H}$ represent the semantic information corresponding to the i-th character in the text image.

We then embed the contrastive learning scheme between $\mathcal{H}^{stu}$ and $\mathcal{H}^{tea}$ to facilitate the learning of more discriminative semantic knowledge. In concrete, we first construct positive and negative samples for contrastive learning. Given h_i in $\mathcal{H}^{stu}(\mathcal{H}^{tea})$ as the anchor, its corresponding positive sample is the i-th vector h'_i in the other set $\mathcal{H}^{tea}(\mathcal{H}^{stu})$, and its negative samples are all the elements from the union of $\mathcal{H}^{tea}$ and $\mathcal{H}^{stu}$ after dropping h_i and h'_i. Then the contrastive learning is implemented in the 1D semantic representation space:

$$\mathcal{L}_{semantic} = -\frac{1}{L}\sum_{i=1}^{L}\log\frac{\exp(sim(h_i, h'_i)/\tau)}{\sum_{h\in\mathcal{H}^{tea}\cup\mathcal{H}^{stu}\setminus h_i}\exp(sim(h_i, h)/\tau)}, \tag{5}$$

where L is the sum of valid character length over batch. $sim(\cdot,\cdot)$ is the distance metric and we use the cosine similarity here. τ is the distillation temperature.

Different from previous contrastive learning methods in sequential recognition [1, 31] which generate contrastive instances in

a predefined and fixed manner (e.g., sliding windows or image patches), we choose to perform contrastive learning on the semantic vectors generated from the sequence modeling module. Since the contrastive instances are already aligned in character order by the attention mechanism, we can obtain each contrastive instance in a data-dependent manner, which is more effective for images with arbitrary text orientations.

3.3.3 Soft Logits Loss. The utilization of the knowledge from the soft teacher label is appealing due to its capability to reflect the similarity between characters [22]. Previous studies [4] have enabled this knowledge transfer by minimizing the KL distance between the output distributions of the student and the teacher at each time step. Nonetheless, this word-level distillation is sub-optimal for sequential tasks as it lacks sequence-level supervision. To this end, we modify the teacher distribution to include both local word-level knowledge and global sequence-level knowledge.

Formally, we refer to the formula in [21], and the probability of k-th character at time step t in the teacher output p_t^k can be revised as the weighted sum of word-level and sequence-level probabilities:

$$\tilde{p}_t^k = (1-\alpha)p_t^k + \alpha \frac{\sum_{\pi_t=k} \prod_{t=1}^{T} p_t^{\pi_t}}{\sum_{\pi} \prod_{t=1}^{T} p_t^{\pi_t}}, \tag{6}$$

where π is all possible decoding paths, α is the hyper-parameter that balances the two distributions. The mathematical derivation can be found in the supplementary material.

However, applying the above equation directly for teacher distribution revision is intractable in practice considering the exponential number of all possible paths. Thus, in practical implementation, we apply some techniques for approximation. Specifically, we select paths with the TopK highest path likelihoods using beam search. Moreover, to ensure the representativity of the beam search results, we only use word-level teacher distribution as supervision when the maximum path likelihood is below a given threshold r.

The soft logits loss is then defined as the KL distance between the modified teacher distribution $\tilde{p}$ and the student distribution q:

$$\mathcal{L}_{logits} = \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{|\mathcal{A}|} \tilde{p}_t^k \log \frac{\tilde{p}_t^k}{q_t^k}, \tag{7}$$

where $|\mathcal{A}|$ is the size of alphabet. Similar to the traditional logits distillation [19], we also adopt a high distillation temperature to smooth the distribution.

Although some previous works [10, 27] also employ sequence-related knowledge in the soft teacher label, they treat the whole path likelihood as the atomic element and ignore the fine-grained information pertaining to different characters. For instance, easily-confused characters in one sequence should possess a lower confidence score. In contrast, the proposed sequence-level distribution uses the votes of all decoding paths at each character, which is integrated with word-level distribution to facilitate the acquisition of both global and local knowledge from the soft teacher.

3.4 Overall Loss

The overall loss is from the following parts: the task-related cross-entropy loss, the visual focus loss that aids in transferring character position knowledge, the semantic contrastive loss that facilitates contextual semantic knowledge, and the soft logits loss that transfers both word-level and sequence-level soft label knowledge.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{ce} + \lambda_2 \mathcal{L}_{visual} + \lambda_3 \mathcal{L}_{semantic} + \lambda_4 \mathcal{L}_{logits}, \tag{8}$$

where λ_1 λ_2, λ_3 and λ_4 are hyper-parameters.

4 EXPERIMENT

4.1 Datasets

TextZoom [51] contains 17367 LR-HR scene text image pairs for training and 4373 pairs for testing. The test set is divided into three subsets, with 1619 pairs for the easy subset, 1411 pairs for the medium subset, and 1343 pairs for the hard subset.
ICDAR2013 (IC13) [25] consists of 1015 images for testing, most of which are regular text images. Some of them are under uneven illumination.
ICDAR2015 (IC15) [24] consists of images taken from scenes and has two versions: 1,811 images ($IC15_S$) and 2,077 images ($IC15_L$). We use ($IC15_S$) for experiments.
CUTE80 [45] consists of 288 images. Most of them are heavily curved but with high resolution.
Street View Text (SVT) [50] has 647 images collected from Google Street View. Some of the images are severely corrupted by noise, blur, and low resolution.
Street View Text Perspective (SVTP) [38] contains 645 images of which texts are captured in perspective views.

4.2 Implementation Details

We use 4 NVIDIA TITAN X GPUs to train our model with batch size 128. We experimentally find that the contrastive learning in $\mathcal{L}_{semantic}$ is not sensitive to batch size. Adam [28] is used for optimization. We adopt a learning rate of 5e-5, with a decay factor of 0.1 every 25 epochs. We set the hyper-parameter in total loss $\lambda_1 = 4$, $\lambda_2 = 2$, $\lambda_3 = 0.025$, $\lambda_4 = 20$. The distillation temperatures in $\mathcal{L}_{semantic}$ and $\mathcal{L}_{logits}$ are 0.1 and 4, respectively. In the beam search, we use the paths with the top 6 likelihoods as approximations and set the threshold $r = 0.1$. We use the proposed distillation protocol on the currently prevalent SoTA text recognizers, namely ABINet [15], MATRN [36] and PARSeq [3]. Since the input images of the student and teacher branches are of different resolutions, we modified the convolution stride (for CNN backbone) or patch sizes (for ViT backbone) to ensure the consistency of the deep visual features between teacher and student. We use the released pre-trained weights to initialize the student and teacher, and fine-tune the student using the proposed distillation framework. We refer to the student model adapted to low-resolution as ABINet-LTR, MATRN-LTR and PARSeq-LTR, respectively.

4.3 Evaluation Metrics

We evaluate the model in terms of efficiency and effectiveness. Specifically, we adopt text recognition accuracy to demonstrate the effectiveness of different methods. We utilize Floating Point Operations (FLOPs) and the number of parameters (Params) to present the efficiency.

Table 1: Ablation on visual focus loss. 'cos' denotes cos-similarity loss, L2 loss is used when cos is removed. 'norm' denotes the mean-variance normalization. 'mask' denotes the mask distillation strategy.

cos	norm	mask	Recognition Accuracy↑			
			Easy	Medium	Hard	avgAcc
			86.16%	71.72%	55.17%	71.98%
✔			86.04%	72.22%	55.25%	72.12%
✔	✔		86.20%	71.89%	**55.72%**	72.22%
✔	✔	✔	**86.91%**	**72.36%**	55.10%	**72.45%**

Table 2: Ablation on semantic contrastive loss. We compare with L2 loss without contrastive learning and SeqCLR with manually predefined contrastive instance division manner.

semantic loss	Recognition Accuracy↑			
	Easy	Medium	Hard	avgAcc
L2	86.59%	72.09%	54.95%	72.19%
SeqCLR [1]	85.97%	72.09%	**55.92%**	72.26%
Ours	**86.91%**	**72.36%**	55.10%	**72.45%**

4.4 Ablation Study

In this section, we conduct an ablation study to demonstrate the effectiveness of each component. We use the widely adopted ABINet as the base recognizer on the TextZoom dataset.

4.4.1 Ablation on Visual Focus Loss. The proposed visual focus loss is employed to exploit the rich character position knowledge in the visual features of teacher branch. It contains the cosine similarity with the normalization operator to bridge the resolution gap, and the mask distillation strategy to prompt a character-focused feature learning. We conduct ablation to validate the effectiveness of different components in $\mathcal{L}_{visual}$. Table 1 shows the results. The direction-related metric results in an average accuracy improvement of 0.14% and the normalization operation improves accuracy by 0.1% through removing the resolution differences. The subsequent mask distillation further improves the average accuracy of 0.23% by character region focus.

4.4.2 Ablation on Semantic Contrastive Loss. We utilize the semantic contrastive loss to extract contextual semantic knowledge with contrastive learning. Since the contrastive instances have been aligned by the attentional sequence modeling module, it is feasible to adaptively decide the number of instances according to the text length in different images. We conduct experiments to verify its effectiveness (see Table 2). It can be seen that both contrastive based methods (SeqCLR and ours) outperform the non-contrastive one (L2), suggesting that contrastive learning can extract more discriminative context-aware semantic knowledge. Furthermore, the proposed method is more robust to spatially deformed text images benefiting from the adaptive instance division, leading to an average accuracy of 0.19% higher than SeqCLR which uses a fixed division manner.

Table 3: Ablation on soft logits loss. WKD denotes word-level distillation. SKD denotes the sequence-level distillation.

logits loss	Recognition Accuracy↑			
	Easy	Medium	Hard	avgAcc
WKD [4]	85.86%	72.22%	55.40%	72.10%
SKD [10]	85.92%	71.58%	54.43%	71.62%
Ours	**86.91%**	**72.36%**	**55.10%**	**72.45%**

Table 4: Joint effect ablation of the proposed loss functions.

visual	semantic	logits	Recognition Accuracy↑			
			Easy	Medium	Hard	avgAcc
		✔	81.35%	68.32%	51.08%	67.85%
	✔	✔	86.35%	72.01%	54.28%	71.87%
✔		✔	85.73%	72.01%	55.40%	71.99%
✔	✔	✔	86.21%	72.16%	**55.42%**	72.22%
✔	✔	✔	**86.91%**	**72.36%**	55.10%	**72.45%**

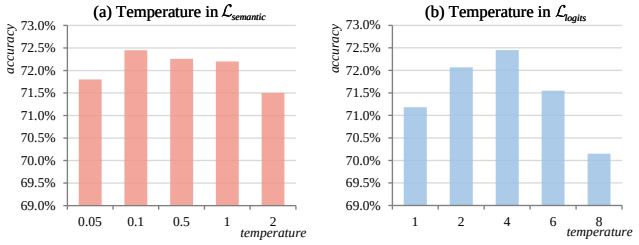

Figure 3: Ablation of distillation temperature on (a) semantic contrastive loss and (b) soft logits loss.

4.4.3 Ablation on Soft Logits Loss. The proposed soft logits loss introduces both local word-level and global sequence-level knowledge from the soft teacher label. To verify the validity, we replace the proposed loss with those used in previous work, i.e. word-level KL loss [4] with each time step modeled independently and sequence distillation loss [10] with the entire path likelihood as the target. Table 3 presents the results. It can be seen that only sub-optimal results can be achieved with only one single-level loss. By contrast, by providing both local and global knowledge, the proposed method outperforms others.

4.4.4 Combination of Different Losses. We analyze the combined effects of the three proposed distillation losses (refer to Table 4). The results show that using only the task-related cross-entropy loss yielded an accuracy of 67.85%. By incorporating the knowledge from the soft teacher label, there is a recognition accuracy improvement of 4.02%. The addition of semantic contrastive loss, which can migrate contextual semantic knowledge, led to an increase in accuracy of 0.12%. The visual focus loss, which transfers character position knowledge, further improved the accuracy by 0.35%. Finally, the best result of 72.45% was achieved when all three losses were utilized.

4.4.5 Ablation on Hyper-parameters. Both semantic contrastive loss and soft logits loss contain temperature hyper-parameters,

Table 5: Comparison with state-of-the-art two-stage methods on the TextZoom dataset regarding efficiency and effectiveness. SR refers to using the STISR model for super-resolution before recognition. KD refers to adapting the text recognizer to low resolution with knowledge distillation.

Text Recognizer	Method	Type	FLOPs↓ ($\times 10^9$)	Params↓ ($\times 10^6$)	Recognition Accuracy↑			
					Easy	Medium	Hard	avgAcc
ABINet [15]	Bicubic	-	-	-	77.52%	56.98%	42.81%	60.23%
	TSRN [51]	SR	6.38	39.42	76.10%	61.16%	45.57%	61.90%
	STT [8]	SR	6.66	39.95	79.80%	64.99%	48.47%	65.40%
	TATT [34]	SR	7.45	52.68	80.67%	65.77%	50.26%	66.52%
	C3-STISR [60]	SR	8.67	76.08	81.35%	66.90%	49.89%	67.03%
	Ours	KD	**5.46**	**36.74**	**86.91%**	**72.36%**	**55.10%**	**72.45%**
MATRN [36]	Bicubic	-	-	-	80.42%	58.97%	44.97%	62.61%
	TSRN [51]	SR	10.74	46.84	77.08%	62.65%	47.21%	63.25%
	STT [8]	SR	11.06	47.37	81.66%	65.98%	50.11%	66.91%
	TATT [34]	SR	11.85	60.10	81.10%	66.62%	51.68%	67.39%
	C3-STISR [60]	SR	13.07	83.50	81.90%	68.04%	51.08%	67.96%
	Ours	KD	**9.86**	**44.16**	**86.91%**	**73.14%**	**56.96%**	**73.27%**
PARSeq [3]	Bicubic	-	-	-	90.36%	75.34%	57.11%	75.30%
	TSRN [51]	SR	3.76	26.51	79.74%	61.09%	47.65%	63.87%
	STT [8]	SR	4.40	27.04	83.69%	66.69%	51.75%	68.40%
	TATT [34]	SR	4.83	39.77	82.21%	65.91%	52.12%	67.71%
	C3-STISR [60]	SR	6.06	63.17	84.25%	68.25%	50.86%	68.83%
	Ours	KD	**2.93**	**23.81**	**90.36%**	**78.88%**	**63.22%**	**78.23%**

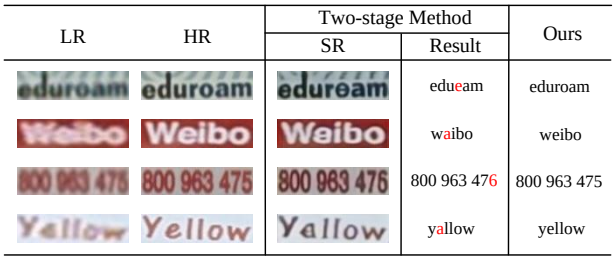

Figure 4: Qualitative comparison with the two-stage approach [60]. Erroneous reconstruction due to low quality of LR in the two-stage pipeline can affect subsequent text recognition.

here we focus on how distillation temperature affects model performance. Fig. 3 (a) shows the effects of distillation temperature in $\mathcal{L}_{semantic}$. It can be seen that as the distillation loss gets smaller, the recognition accuracy increases. This is due to the fact that a small temperature amplifies the differences between the semantic vectors in contrastive learning, forcing the model to focus on the subtle differences, thus facilitating the learning of more discriminative embeddings and leading to an increase in performance. However, too low a temperature can lead to an unhealthy gradient by over-amplifying the difference. In addition, the variation of model performance with temperature in $\mathcal{L}_{logits}$ is shown in Fig. 3 (b). A larger distillation temperature can make the distribution more uniform, thus aiding the learning of similarities between characters and enhancing the model's ability to differentiate between confusable characters. However, too large a temperature is detrimental due to the reduction of information.

4.5 Comparison to State-of-the-Arts

We first compare our method with super-resolution based two-stage methods. After that, we conduct the robustness test on five challenging STR benchmarks. Finally, we compare with other solutions to LTR.

4.5.1 Comparison with Two-stage Methods. We compare with the super-resolution based two-stage approaches which use the STISR model before recognition. Table 5 shows the results on the TextZoom dataset. Surprisingly, the recognition accuracy of bicubic even exceeds that of the SoTA STISR method [60] when applying a recognizer [3] with strong contextual modeling capabilities. We speculate that it is due to the manipulation of the original LR image by the super-resolution model. Moreover, by including only one text recognizer without any pre-processing model, the proposed framework is highly efficient, for example, PARSeq-LTR saves 3.13 GFLOPs and 39.36M parameters compared with C3-STISR. In addition, our method achieves a significant recognition accuracy improvement over the two-stage pipelines due to the use of joint optimization.

Table 6: Adaption to scene text recognition benchmarks. Images are manually downscaled to obtain LR for testing.

Method	Recognition Accuracy↑				
	IC13	IC15	CUTE80	SVT	SVTP
abinet					
TSRN	53.20%	36.44%	39.93%	42.66%	36.59%
STT	56.35%	39.37%	41.30%	41.42%	40.47%
TATT	53.60%	36.72%	40.97%	44.05%	37.36%
C3-STISR	55.17%	39.48%	36.81%	44.51%	39.22%
Ours	**57.44%**	**44.17%**	**41.32%**	**46.06%**	**42.64%**
matrn					
TSRN	54.09%	35.89%	39.93%	43.43%	38.76%
STT	56.65%	40.36%	42.36%	43.89%	42.02%
TATT	57.43%	39.76%	42.01%	43.12%	43.41%
C3-STISR	57.64%	42.19%	42.36%	44.51%	42.64%
Ours	**59.01%**	**43.57%**	**43.40%**	**53.31%**	**43.72%**
parsq					
TSRN	55.86%	38.98%	47.92%	44.20%	39.69%
STT	60.89%	42.79%	46.18%	46.68%	44.03%
TATT	55.96%	41.74%	48.26%	46.21%	41.24%
C3-STISR	58.72%	42.30%	44.10%	45.44%	42.33%
Ours	**62.36%**	**49.81%**	**53.82%**	**53.94%**	**49.61%**

Impressively, PARSeq-LTR even outperforms the two-stage pipeline which uses C3-STISR model for STISR by 9.4% on average recognition accuracy. A qualitative comparison is shown in Fig. 4.

4.5.2 Robustness Comparison. We first test the data distribution shift robustness of different methods by freezing the model trained on TextZoom and directly transferring them to five scene text recognition benchmarks. Since these STR datasets contain high-resolution images, we first manually degrade the raw images, including blur, noise, etc., before using the degraded images to test the robustness, see the supplementary material for details.

Table 6 shows the experimental results. It can be seen that the proposed method achieves SoTA performance on all five STR datasets even after transferring to other data distribution. For example, PARSeq-LTR achieves a 7.51% performance boost versus C3-STISR on IC15 dataset. The above results demonstrate the data distribution robustness of the proposed method.

Furthermore, we study the performance variation of different methods under different levels of Gaussian Blur and Gaussian Noise. Fig. 5 shows the results. For the gaussian blur, our method drops more slowly and outperforms the super-resolution based two-stage methods at all blur settings. For the gaussian noise, the difference among different methods is not significant when the noise is at a high setting, but our method still beats others.

4.5.3 Comparison with Other Solutions to LTR. In addition to the super-resolution based two-stage framework, multi-task learning, which handles LTR by learning common visual features for both super-resolution and text recognition, has also been proposed [23, 35]. In this work, we compare with two representative multi-task

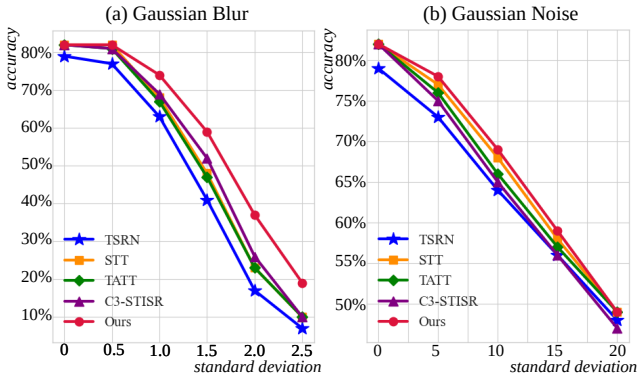

Figure 5: Recognition accuracy of different methods with varying (a) Gaussian Blur and (b) Gaussian Noise. ABINet is used on the IC15 dataset.

Table 7: Comparison with multi-task learning based pipelines. The proposed high-resolution knowledge transfer framework still achieves SoTA performance.

Method	Recognition Accuracy↑			
	Easy	Medium	Hard	avgAcc
PlugNet [35]	81.90%	68.89%	52.27%	68.60%
IFR [23]	82.58%	68.89%	52.87%	69.04%
Ours	**86.91%**	**72.36%**	**55.10%**	**72.45%**

learning based methods, i.e., PlugNet [35] and IFR [23]. To ensure a fair comparison, we apply the plug-and-play feature super-resolution modules of PlugNet and IFR to ABINet, and compare them with our ABINet-LTR. The results presented in Table 7 show that although this multi-task framework outperforms the super-resolution based two-stage one, it still falls behind the proposed distillation framework. This is because these methods are limited to pixel space learning as well as do not take into account supervision from different perspectives, such as semantic space and sequence modeling, which results in limited performance.

5 CONCLUSION

We propose a novel knowledge distillation framework that adapts text recognizers for the low-resolution to address challenges posed by previous super-resolution based two-stage pipelines. Three distillation losses are designed to extract multi-level knowledge from the high-resolution. The visual focus loss transfers the character position knowledge in a resolution-agnostic manner and reinforces the character focus with mask distillation. The semantic contrastive loss leverages contrastive learning to facilitate the learning of discriminative contextual semantic knowledge. The soft logits loss models both local word-level and global sequence-level knowledge in the soft teacher label for better supervision. Extensive experiments demonstrate that the proposed one-stage pipeline achieves state-of-the-art performance against two-stage counterparts in terms of efficiency and effectiveness, with favorable robustness. We hope that our work can inspire more studies on one-stage LTR.

REFERENCES

[1] Aviad Aberdam, Ron Litman, Shahar Tsiper, Oron Anschel, Ron Slossberg, Shai Mazor, R Manmatha, and Pietro Perona. 2021. Sequence-to-sequence contrastive learning for text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15302–15312.

[2] Mohamad Alansari, Oussama Abdul Hay, Sajid Javed, Abdulhaid Shoufan, Yahya Zweiri, and Naoufel Werghi. 2023. GhostFaceNets: Lightweight Face Recognition Model from Cheap Operations. *IEEE Access* (2023).

[3] Darwin Bautista and Rowel Atienza. 2022. 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*. Springer, 178–196.

[4] Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, and Yi-Zhe Song. 2021. Text is text, no matter what: Unifying text recognition using knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 983–992.

[5] Konstantin Bulatov, Nadezhda Fedotova, and Vladimir V Arlazarov. 2021. An approach to road scene text recognition with per-frame accumulation and dynamic stopping decision. In *Thirteenth International Conference on Machine Vision*, Vol. 11605. SPIE, 511–519.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 213–229.

[7] Hongyuan Chen, Yanting Pei, Hongwei Zhao, and Yaping Huang. 2022. Super-resolution guided knowledge distillation for low-resolution image classification. *Pattern Recognition Letters* 155 (2022), 62–68.

[8] Jingye Chen, Bin Li, and Xiangyang Xue. 2021. Scene Text Telescope: Text-Focused Scene Image Super-Resolution. *computer vision and pattern recognition* (2021).

[9] Jingye Chen, Haiyang Yu, Jianqi Ma, Bin Li, and Xiangyang Xue. 2022. Text gestalt: Stroke-aware scene text image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 285–293.

[10] Ying Chen, Liang Qiao, Zhanzhan Cheng, Shiliang Pu, Yi Niu, and Xi Li. 2022. 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*. Springer, 356–373.

[11] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. 2019. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11065–11074.

[12] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.

[13] C Dong, X Zhu, Y Deng, CC Loy, and Y Qia. 2015. Boosting optical character recognition: A super-resolution approach. *arXiv preprint arXiv:1506.02211* (2015).

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[15] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. 2021. 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*. 7098–7107.

[16] Jan Flusser, Sajad Farokhi, Cyril Höschl, Tomáš Suk, Barbara Zitova, and Matteo Pedone. 2015. Recognition of images degraded by Gaussian blur. *IEEE transactions on Image Processing* 25, 2 (2015), 790–806.

[17] Shiming Ge, Kangkai Zhang, Haolin Liu, Yingying Hua, Shengwei Zhao, Xin Jin, and Hao Wen. 2020. Look one and more: Distilling hybrid order relational knowledge for cross-resolution image recognition. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 10845–10852.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[20] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[21] Mingkun Huang, Yongbin You, Zhehuai Chen, Yanmin Qian, and Kai Yu. 2018. Knowledge Distillation for Sequence Model.. In *Interspeech*. 3703–3707.

[22] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. 2014. Speeding up convolutional neural networks with low rank expansions. *arXiv preprint arXiv:1405.3866* (2014).

[23] Zhiwei Jia, Shugong Xu, Shiyi Mu, Yue Tao, Shan Cao, and Zhiyong Chen. 2021. IFR: Iterative Fusion Based Recognizer for Low Quality Scene Text Recognition. In *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part II 4*. Springer, 180–191.

[24] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman K. Ghosh, Andrew D. Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. 2015. ICDAR 2015 competition on Robust Reading. *International Conference on Document Analysis and Recognition* (2015).

[25] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. 2013. ICDAR 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*. IEEE, 1484–1493.

[26] Vijeta Khare, Palaiahnakote Shivakumara, Chee Seng Chan, Tong Lu, Liang Kim Meng, Hon Hock Woon, and Michael Blumenstein. 2019. A novel character segmentation-reconstruction approach for license plate recognition. *Expert Systems with Applications* 131 (2019), 219–239.

[27] Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947* (2016).

[28] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv: Learning* (2014).

[29] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4681–4690.

[30] Hoang Danh Liem, Nguyen Duc Minh, Nguyen Bao Trung, Hoang Tien Duc, Pham Hoang Hiep, Doan Viet Dung, and Dang Hoang Vu. 2018. Fvi: An end-to-end vietnamese identification card detection and recognition in images. In *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*. IEEE, 338–340.

[31] Hao Liu, Bin Wang, Zhimin Bao, Mobai Xue, Sheng Kang, Deqiang Jiang, Yinsong Liu, and Bo Ren. 2022. Perceiving stroke-semantic context: Hierarchical contrastive learning for robust scene text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1702–1710.

[32] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*. PMLR, 97–105.

[33] Jianqi Ma, Shi Guo, and Lei Zhang. 2023. Text prior guided scene text image super-resolution. *IEEE Transactions on Image Processing* 32 (2023), 1341–1353.

[34] Jianqi Ma, Zhetong Liang, and Lei Zhang. 2022. A Text Attention Network for Spatial Deformation Robust Scene Text Image Super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5911–5920.

[35] Yongqiang Mou, Lei Tan, Hui Yang, Jingying Chen, Leyuan Liu, Rui Yan, and Yaohong Huang. 2020. PlugNet: Degradation Aware Scene Text Recognition Supervised by a Pluggable Super-Resolution Unit. *european conference on computer vision* (2020).

[36] Byeonghu Na, Yoonsik Kim, and Sungrae Park. 2022. Multi-modal text recognition networks: Interactive enhancements between visual and semantic features. In *European Conference on Computer Vision*. Springer, 446–463.

[37] Shimon Nakaune, Satoshi Iizuka, and Kazuhiro Fukui. 2021. Skeleton-aware text image super-resolution. In *Proceedings of the 32nd British Machine Vision Conference, Online*. 22–25.

[38] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. 2013. Recognizing Text with Perspective Distortion in Natural Scenes. *international conference on computer vision* (2013).

[39] Lu Qi, Jason Kuen, Jiuxiang Gu, Zhe Lin, Yi Wang, Yukang Chen, Yanwei Li, and Jiaya Jia. 2021. Multi-scale aligned distillation for low-resolution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14443–14453.

[40] Zhi Qiao, Yu Zhou, Jin Wei, Wei Wang, Yuan Zhang, Ning Jiang, Hongbin Wang, and Weiping Wang. 2021. Pimnet: a parallel, iterative and mimicking network for scene text recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2046–2055.

[41] Rui Qin, Bin Wang, and Yu-Wing Tai. 2022. Scene Text Image Super-Resolution via Content Perceptual Loss and Criss-Cross Transformer Blocks. *arXiv preprint arXiv:2210.06924* (2022).

[42] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[43] Baoguang Shi, Xiang Bai, and Cong Yao. 2016. 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* 39, 11 (2016), 2298–2304.

[44] Sungho Shin, Joosoon Lee, Junseok Lee, Yeonguk Yu, and Kyoobin Lee. 2022. 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*. Springer, 631–647.

[45] Palaiahnakote Shivakumara, Anhar Risnumawan, Chee Seng Chan, and Chew Lim Tan. 2014. A robust arbitrary text detection system for natural scene images. *Expert Systems With Applications* (2014).

[46] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[47] Hanh TM Tran and Tien Ho-Phuoc. 2019. Deep laplacian pyramid network for text images super-resolution. In *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*. IEEE, 1–6.

[48] Luan Tran, Xi Yin, and Xiaoming Liu. 2017. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1415–1424.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[50] Kai Wang, Boris Babenko, and Serge Belongie. 2011. End-to-end scene text recognition. *international conference on computer vision* (2011).

[51] Wenjia Wang, Enze Xie, Xuebo Liu, Wenhai Wang, Ding Liang, Chunhua Shen, and Xiang Bai. 2020. Scene Text Image Super-Resolution in the Wild. *european conference on computer vision* (2020).

[52] Wenjia Wang, Enze Xie, Peize Sun, Wenhai Wang, Lixun Tian, Chunhua Shen, and Ping Luo. 2019. Textsr: Content-aware text super-resolution guided by recognition. *arXiv preprint arXiv:1909.07113* (2019).

[53] Yuxin Wang, Hongtao Xie, Shancheng Fang, Jing Wang, Shenggao Zhu, and Yongdong Zhang. 2021. From two to one: A new scene text recognizer with visual language modeling network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14194–14203.

[54] Xudong Xie, Ling Fu, Zhifei Zhang, Zhaowen Wang, and Xiang Bai. 2022. Toward Understanding WordArt: Corner-Guided Transformer for Scene Text Recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*. Springer, 303–321.

[55] Xiangyu Xu, Deqing Sun, Jinshan Pan, Yujin Zhang, Hanspeter Pfister, and Ming-Hsuan Yang. 2017. Learning to super-resolve blurry face and text images. In *Proceedings of the IEEE international conference on computer vision*. 251–260.

[56] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. 2022. Masked generative distillation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XI*. Springer, 53–69.

[57] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. 2020. Towards accurate scene text recognition with semantic reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12113–12122.

[58] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. 2018. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*. 286–301.

[59] Cairong Zhao, Shuyang Feng, Brian Nlong Zhao, Zhijun Ding, Jun Wu, Fumin Shen, and Heng Tao Shen. 2021. Scene Text Image Super-Resolution via Parallelly Contextual Attention Network. *acm multimedia* (2021).

[60] Minyi Zhao, Miao Wang, Fan Bai, Bingjia Li, Jie Wang, and Shuigeng Zhou. 2022. C3-STISR: Scene Text Image Super-resolution with Triple Clues. *international joint conference on artificial intelligence* (2022).

[61] Mingjian Zhu, Kai Han, Chao Zhang, Jinlong Lin, and Yunhe Wang. 2019. Low-resolution visual recognition via deep feature distillation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3762–3766.

[62] Shipeng Zhu, Zuoyan Zhao, Pengfei Fang, and Hui Xue. 2023. Improving Scene Text Image Super-Resolution via Dual Prior Modulation Network. *arXiv preprint arXiv:2302.10414* (2023).