Embedding Rotation Invariance for Provable Multi-Oriented Scene Text Recognition

Zhibin Ma[1,2], Pengwen Dai[1,2]✉, Yi Liu[3], Xugong Qin[4], Chenyun Yu[1], and Xiaochun Cao[1,2]

[1] Shenzhen Campus of Sun Yat-sen University, China
[2] Shenzhen Key Laboratory of Adversarial Artificial Intelligence, China
[3] Baidu Inc., China
[4] Nanjing University of Science and Technology, China
mazhb3@mail2.sysu.edu.cn, {daipw, yuchy35, caoxiaochun}@mail.sysu.edu.cn, liuyi22@baidu.com, qinxugong@njust.edu.cn

Abstract. Multi-oriented text is ubiquitous in real-world scenes and remains a major challenge for scene text recognition (STR). Existing rotation-aware methods explicitly estimate text orientation. However, due to the lack of theoretical guarantees, they are prone to error accumulation, increased computational cost, and strong reliance on data. In this work, we incorporate rotation invariance into the STR framework to address these limitations. Specifically, we adopt an encoder–decoder architecture, embedding rotation equivariance in the encoder and rotation invariance in the decoder to construct a fully rotation-invariant network. On the decoder side, we first identify and prove the rotation-invariant property of the cross-attention mechanism and use it to formulate a rotation-invariant text decoder that maps visual features to output text in a rotation-invariant manner. On the encoder side, we propose a rotation-equivariant local–global extraction network that integrates deep equivariant convolutions with self-attention, enabling rotation-equivariant feature extraction while modeling inter-character dependencies and preserving fine-grained visual details. By integrating the encoder and decoder, we obtain an end-to-end **R**otation-**I**nvariant **S**cene **TE**xt **R**ecognition network (RISTER). RISTER provides rotation invariance with theoretical guarantees, enhancing robustness on multi-oriented samples without introducing additional inference computation or relying on data-driven orientation correction. Experiments show that RISTER achieves state-of-the-art performance on both standard and multi-oriented benchmarks, surpassing the second-best model by 4.0% in accuracy on the general multi-oriented dataset.

Keywords: Scene Text Recognition, Rotation Invariance, Cross-Attention

1 Introduction

Scene Text Recognition (STR) aims to translate the text in images into a machine-readable form. Its wide applications in areas such as autonomous driv-

✉ Corresponding author.

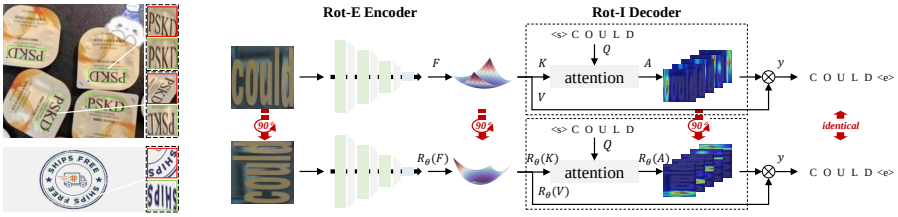

Fig. 1: (a) The detection results produced by LRANet [41]. The red boxes indicate results obtained by cropping with the minimum enclosing bounding boxes (arbitrary angles), while the green boxes indicate results obtained using TPS-based rectification (canonical angles). (b) In our proposed rotation-invariant network, rotating the input image induces equivariant transformations in both visual features and attention maps. After the attend operation, they lead to exactly the same prediction.

ing [58], visual understanding [36], and embodied intelligence [28] have made it a popular research topic. Existing STR methods primarily focus on enhancing the model's language modeling capability, either by introducing explicit language models [1, 12, 22, 54] or by integrating visual feature extraction and language modeling into a unified framework [9, 11, 46]. However, the orientation of text in images, which is one of the most common and important issues, has been largely overlooked [18]. In real-world scenarios, text appears in various orientations. While most popular text detectors regress boundary points to obtain tightly fitted text bounding boxes, they inevitably produce multi-oriented text instances, primarily with orientations of 0°, 90°, 180°, and 270°, as shown in Figure 1 (a). These samples, referred to as canonical-angle samples in this paper, pose significant challenges to the recognition task, while existing STR methods typically assume that the text in input images is arranged from left to right. When the text orientation deviates significantly from the horizontal direction, existing text recognizers are prone to producing incorrect outputs [18].

Although a limited number of studies have focused on recognizing multi-oriented text, they all face the following issues. (1) No theoretical guarantees: Methods based on multi-oriented encoding [3] or rectification modules [11,39,56] require the model to perceive text orientation explicitly. If the orientation perception network (e.g., the localization network in ASTER [39]) fails to estimate the text orientation correctly, the model struggles to recognize the text. (2) More computational cost: The introduction of an additional perception network incurs extra computation, leading to reduced inference efficiency. (3) Reliance on data augmentation: Rotation-based data augmentation during training becomes necessary, forcing the model to separately learn features for different orientations, which reduces feature utilization efficiency.

Inspired by existing rotation-equivariant networks [4, 40, 49], we attempt to provide a unified solution to the above problems by embedding rotation equivariance into STR networks. However, current rotation-equivariant networks are

not directly suitable for STR, as they are primarily designed for classification [4] or low-level vision tasks [42, 59]. In contrast, STR is a unique variable-length, multi-label recognition problem, which, unlike other tasks, requires achieving rotation invariance in an end-to-end manner rather than mere equivariance.

In this work, we incorporate rotation invariance into the STR network to address the aforementioned issues. Specifically, we identify a previously overlooked rotation-invariant property of the cross-attention mechanism shared by a class of attention-based STR decoders. Our study demonstrates that the cross-attention mechanism exhibits a rotation-invariant property when operating on 2d feature inputs, namely, the attended features remain identical before and after arbitrary spatial rotations of the visual features. We further analyze this distinctive property and provide a theoretical proof. Based on this observation, we derive a set of design principles for constructing *Rotation-Invariant Text Decoder (RITD)*. By minimally modifying existing decoder architectures, RITD realizes a rotation-invariant mapping from visual features to the text output space, producing consistent variable-length predictions and confidence scores under input rotations (exact at canonical orientations and approximate at other orientations).

While RITD ensures rotation-invariant decoding, achieving an end-to-end rotation-invariant STR system further requires rotation-equivariant visual feature extraction. This motivates us to combine RITD with rotation-equivariant visual encoders. However, existing rotation-equivariant backbones are not well-suited for STR. Specifically, purely convolutional designs [4,33,49] fail to capture inter-character dependencies, while attention-based counterparts [17,32,51] typically rely on shallow convolutional tokenization, leading to early downsampling and the loss of fine-grained details. To address these issues, we propose a hierarchical *Rotation-Equivariant Local–Global Extraction (RELG)* network that integrates deep equivariant convolutions with self-attention, enabling rotation-equivariant feature extraction while preserving detailed geometric information and modeling inter-character dependencies and encodes positional information. Finally, by combining RELG and RITD into an encoder-decoder network, we construct a ***R**otation-**I**nvariant S**TR** network (RISTER)*. The inherent rotation invariance of RISTER ensures identical recognition results before and after rotations of the input image (As shown in Figure 1 (b)). This property allows RISTER to recognize multi-oriented text without relying on any correction modules or data augmentation strategies, maintaining stability across various orientations while avoiding additional computational overhead. Furthermore, integrating rotation invariance boosts data utilization and strengthens the aggregation of key features, thereby improving the model's robustness across diverse and challenging benchmarks. Our contributions are summarized as follows:

- We incorporate rotation invariance into STR, which leads to significantly improved robustness to multi-oriented text and mitigates the computational and data-efficiency limitations of prior rotation-aware methods.
- A theoretical proof of the rotation-invariant property of cross-attention is provided. Based on this analysis, we derive the design principles for con-

structing *Rotation-Invariant Text Decoder* (RITD) that establishes a rotation-invariant mapping from the feature space to the output text space.
- A *Rotation-Equivariant Local–Global Extraction* (RELG) network is proposed as the encoder backbone, characterized by learning a rotation-equivariant image-to-feature mapping while capturing inter-character dependencies and preserving fine-grained visual details.
- By combining RELG with RITD, a *Rotation-Invariant STR network* is built, which exhibits identical recognition results before and after image rotation. Experimental results demonstrate that RISTER consistently achieves state-of-the-art performance on both standard and multi-oriented benchmarks.

2 Related Works

2.1 Multi-oriented Text Recognition

Existing STR methods typically recognize multi-oriented text by combining 2d feature extraction with attention mechanisms [1,10,18,21,26]. Unlike 1d feature extractors [38,39,55] or CTC-based methods [9,11,37], these models retain spatial structures and use attention during decoding to localize and identify each character, enhancing recognition of irregular text layouts. However, their performance is highly dependent on data augmentation and the precision of the attention mechanism. Once attention fails, accuracy drops sharply. Besides, AON [3] encodes images from four directions to capture directional features and fuses them via a filter gating module, which increases computational cost and may introduce inconsistent spatial semantics across directions. SLOAN [5] transforms images from Cartesian to polar coordinates, converting rotations into translations to leverage CNNs' translation equivariance. Yet, the polar transformation causes spatial distortion that weakens feature extraction. Other approaches, such as ASTER [39], ESIR [56], and SVTRv2 [11], attempt to rectify rotated text to a horizontal orientation via interpolation or feature rearrangement. These methods add extra computation and lack theoretical guarantees, leading to unstable performance on multi-oriented samples.

2.2 Rotation Equivariant Deep Networks

Incorporating rotation equivariance into neural networks has become a prominent research topic in computer vision. G-CNN [4] first introduced a theoretical framework for equivariant convolutions under $\pi/2$ rotations, inspiring a variety of Rot-E approaches [16,33,34,49]. Methods such as [47,48] achieved continuous rotation equivariance through filter parameterization, while [33, 34] linked convolutions with differential operators for formal error analysis. Subsequent works extended equivariance to transformers [15, 17, 32, 51], and theoretical studies demonstrated that group-equivariant networks exhibit improved generalization capabilities [40]. In applications, F-Conv [49] introduced Fourier-based filters to maintain rotational consistency in the super-resolution task, with subsequent

works confirming enhanced stability in related vision tasks such as image inpainting [23], image fusion [59], or reconstruction [42].

Despite these advances, rotation equivariant networks remain unexplored in STR. Unlike classification or low-level tasks, STR is a variable-length, multi-label sequence generation task, which requires rotation equivariance during feature extraction but rotation invariance during decoding to ensure orientation-independent semantic output. In addition, the unique requirements of STR tasks for backbones to model inter-character dependencies and perform progressive downsampling makes existing equivariant networks unsuitable for direct use as encoders. Therefore, achieving efficient equivariant feature extraction along with rotation-invariant decoding is the main focus of this work.

3 Rotation-Invariant Text Recognition Network

In this section, we first present the overall architecture of RISTER and describe how it realizes a rotation-invariant mapping from the input image to the output text. We then detail the architecture of the proposed Rotation-Equivariant Local–Global Extraction network. Finally, we introduce the rotation-invariant property of cross-attention and demonstrate how it can be exploited to achieve rotation-invariant variable-length decoding.

3.1 Overall Architecture

RISTER follows the widely adopted encoder–decoder paradigm in STR. In particular, we achieve an overall rotation-invariant architecture by embedding rotation equivariance into the encoder and rotation invariance into the decoder, respectively. During visual encoding, the input image x is fed into the *Rotation-Equivariant Local–Global Extractor (RELG)* to perform feature extraction and sequence relationship modeling, producing the visual representation F:

$$F = RELG(x) \in \mathbb{R}^{c \times \frac{h}{8} \times \frac{w}{8}}, \tag{1}$$

where c denotes the feature dimension, while h and w represent the height and width of the input image, respectively. During decoding, F together with the previously decoded sequence V is input into the *Rotation-Invariant Text Decoder (RITD)*, which autoregressively generates the complete text sequence $y \in \mathbb{R}^L$:

$$y^t = RITD(F, V^{0:t-1}) \in \mathbb{R}^1, \tag{2}$$

$$V^t = Embedding(y^t) \in \mathbb{R}^{c \times 1}, \tag{3}$$

where $t = 1, \ldots, L$ denotes the time step in the autoregressive decoding process, y^t represents the character decoded at time step t, and y^0 denotes the start-of-sequence token.

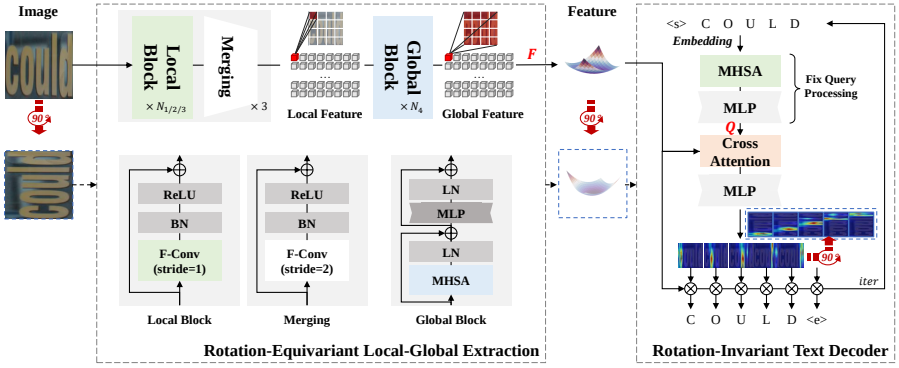

Fig. 2: An overview of RISTER. It employs RELG for feature representation learning and decodes character sequences from the resulting visual features using RITD. When the image undergoes spatial rotation, we obtain equivariant visual features and attention maps (as shown in the blue dash boxes); after being attended in the decoder, the model produces rotation-invariant predictions.

3.2 Rotation-Equivariant Local-Global Extractor

The RELG is designed with three objectives: (1) to achieve rotation-equivariant visual feature extraction; (2) to preserve fine-grained local features through progressive downsampling; and (3) to model inter-character dependencies and positional information. Although existing Rot-E networks [17, 32, 34, 48, 49, 51] can satisfy the first objective, they cannot simultaneously fulfill the latter two objectives, which limits their applicability to text recognition tasks. In RELG, we address these two objectives through Rot-E local extraction and Rot-E global extraction, respectively.

Rot-E Local Extraction We construct the Local Block using F-Conv [49], a type of group-equivariant convolution, to enable rotation-equivariant local feature extraction while maximizing the preservation of fine-grained features through progressive downsampling. Specifically, RELG consists of three stages. In each stage, several local blocks are first applied for feature extraction, followed by a merging operation for spatial downsampling and channel expansion:

$$F_k = Merging(\mathcal{LB}(F_{k-1})) \in \mathbb{R}^{c_k \times \frac{w}{2^k} \times \frac{h}{2^k}}, k = 1, 2, 3 \tag{4}$$

where F_k denotes the feature output of the k-th stage, and $F_0 = x$. $\mathcal{LB}(\cdot)$ is composed of N_k stacked local blocks. In each local block, we employ F-Conv to perform local feature extraction. The convolution kernels are of size 3×3 with a stride of 1. For the choice of the rotation group, we adopt the discrete rotation group C_4 as [49] does, corresponding to rotations by canonical angles. The merging operation is similar to a local block, except that it uses a stride of 2 and increases the number of output channels. After each stage, the resolution

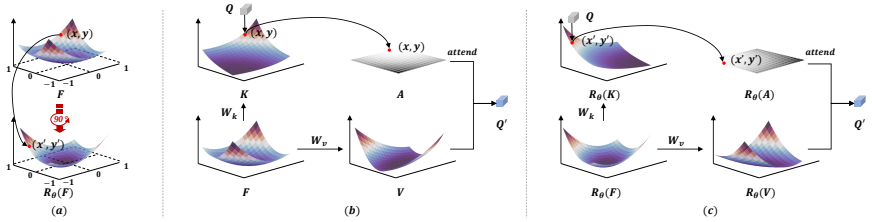

Fig. 3: Illustration of the rotation invariant of cross-attention. Keeping the query token Q fixed, the attended feature Q' is identical before and after spatial rotation of F.

of the input image is reduced to 1/2, 1/4, and 1/8 of the original, respectively, while the number of channels is expanded to 96, 192, and 384.

Rot-E Global Extraction Attention-based STR networks require visual features to encode inter-character dependencies, which necessitates global relationship modeling over these features. We achieve this through a self-attention mechanism. Notably, existing works have already demonstrated that self-attention layers exhibit equivariance to rotations [32, 51]. Specifically, global blocks take the visual feature F_3 as input and finally outputs the enhanced visual feature representation F:

$$F = \mathcal{GB}(F_3) \in \mathbb{R}^{c_3 \times \frac{w}{8} \times \frac{h}{8}}, \tag{5}$$

where $\mathcal{GB}(\cdot)$ denotes the stack of global blocks composed of multi-head self-attention and MLP layers.

3.3 Design principles of Rotation-Invariant Text Decoder

Rotation-invariant variable-length decoding is a core component of RISTER, achieved by exploiting the rotation invariance of cross-attention. We first describe the rotation-invariant property of cross-attention. Next, we outline the principles for constructing a rotation-invariant decoder, and finally present an autoregressive decoding architecture that realizes these principles.

Rotation invariant of cross attention. Figure 3 illustrates the rotation-invariant property of cross-attention, which can be described as:

$$\mathcal{C}(Q, R_\theta(F)) \equiv \mathcal{C}(Q, F), \tag{6}$$

where $\mathcal{C}$ denotes the cross-attention operation, Q denotes the query tokens, F denotes the source tokens, and R_θ denotes a spatial rotation of F by an angle θ. As illustrated in Fig. 3, with fixed query tokens, a spatial rotation applied to F results in equivariant transformations of the corresponding keys K, values V, and attention map A, thereby guaranteeing that the attended output remains rotation-invariant. The mathematical proof of Equation 6 is provided in Appendix A, where it is shown to hold exactly at canonical angles and approximately at other orientations.

Design of the proposed decoder. The validity of Eq. 6 reveals a fundamental property of cross-attention: it defines a mapping over unordered feature sets and is inherently insensitive to geometric structures. Therefore, as long as this invariance is carefully preserved, a decoder constructed using cross-attention can naturally enable rotation-invariant variable-length decoding. The design of RITD must satisfy the following two principles: (1) cross-attention is used to perform modality interaction, and (2) the query tokens fed into the cross-attention remain invariant under image rotation. Following these two principles, we adopt an autoregressive decoding scheme to implement our RITD, as illustrated in the right part of Figure 2. Specifically, at time step t, we first establish dependencies among previously decoded characters via fixed query processing, thereby obtaining the query Q:

$$Q_{0:t-1} = MLP(MHSA(Embedding(y_{0:t-1}))), \tag{7}$$

where $y_{0:t-1}$ denotes the previously decoded characters, and y_0 represents the start-of-sequence token. Subsequently, Q_{t-1} is used as the query token, and the visual features F are treated as source tokens and fed into the cross-attention module to compute the attended output. After passing through an MLP for feature mapping, the predicted character is obtained by applying the $argmax$ operator over the output logits:

$$y_t = argmax(MLP(\mathcal{C}(Q_{t-1}, F))). \tag{8}$$

Subsequently, y_t is used as a decoded character to compute the query token for the next time step, and the decoder iteratively outputs the subsequent predicted characters. Since the query token Q fed into the cross-attention remains invariant under image rotation, the decoder consistently produces identical predictions at the same time step before and after feature rotation.

3.4 Optimization Objective

The training objective follows the traditional AR Loss, formulated as:

$$\min_{\mathcal{W}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[-\sum_{t=1}^{L} \log p_{\mathcal{W}} \left(y_t \mid y_{<t}, x \right) \right], \tag{9}$$

where $\mathcal{W}$ denotes the model parameters and L is the maximum decoding length.

3.5 Architecture Variants

We design different variants of RISTER by adjusting its hyperparameters, including the number of local blocks and global blocks at each stage, as well as the number of attention heads in both the encoder and decoder. Detailed configurations are shown in Table 1. By designing different RISTER variants, flexible choices can be offered for different application scenarios.

Table 1: Architecture specifications of RISTER variants. *Heads* denote the number of self-attention heads in both the global blocks/decoder. Measured on the IC13 dataset.

	$[N_1, N_2, N_3, N_4]$	*Heads*	FLOPs ($\times 10^9$)	*fps*	Params ($\times 10^6$)
Tiny	[3,3,3,6]	6	7.22	42.1	15.9
Small	[9,9,9,9]	6	12.99	31.4	21.8
Base	[12,12,12,15]	12	18.48	26.8	32.8
Large	[18,18,18,18]	12	24.25	22.7	38.6

4 Experiments

4.1 Datasets and Implementation Details

For training, we adopt the large-scale real-world dataset Union14M-Filter [11], a refined version of Union14M-L [18], which contains 3.2M training images from 17 datasets. For evaluation, we test our method on 14 English benchmarks, including: (1) six common STR benchmarks (**CoB**): IIIT5k (3000) [25], SVT (647) [43], IC13 (857) [20], IC15 (1811) [19], SVTP (645) [27], and CUTE80 (288) [31]; (2) **Union14M-benchmarks (U14M-B)** [18], covering seven challenging categories: Curved (2426), Multi-Oriented (1369), Artistic (900), Contextless (779), Salient (1585), Multi-Words (829), and General (400k); (3) **ASOT** (3000) [5], consisting of 3,000 text images with diverse orientations and lengths.

During training, we use AdamW [24] with weight decay of 0.05. The learning rate is set to 3.2×10^{-5} for RISTER-T / S and 1.6×10^{-5} for RISTER-B / L, with a batch size of 512. A OneCycleLR scheduler with 1.5 epochs of linear warm-up is applied over 20 epochs. All images are resized to 128×128. Owing to the intrinsic rotation invariance of our method, rotation-based augmentation is omitted, whereas the remaining data augmentation settings follow those of PARSeq [1]. The maximum text length L is 25, and the character set size N is 96, including letters, digits, punctuation, and special tokens (start, end, pad). The results are evaluated using the Word Accuracy Ignore Cases (WAIC) metric. All experiments are conducted on a single 80 GB A100 GPU.

4.2 Effectiveness of the proposed framework

RISTER exhibits end-to-end rotation invariance, from the input image to the final predicted text. To verify this property, we rotate the input images and measure the similarity between the logits produced by the model. As illustrated in Figure 4(a), RISTER yields exactly the same logits for images before and after rotation, thereby confirming its strict rotation invariance under canonical-angle rotations. Consequently, the model demonstrates strong zero-shot rotation generalization, accurately recognizing images with diverse semantic orientations despite being trained exclusively on horizontally oriented images. The rotation invariance of RISTER arises from the combination of a rotation-equivariant encoder and a rotation-invariant decoder. To verify the roles of these two components, we design two ablation variants: (a) replacing the equivariant convolutions

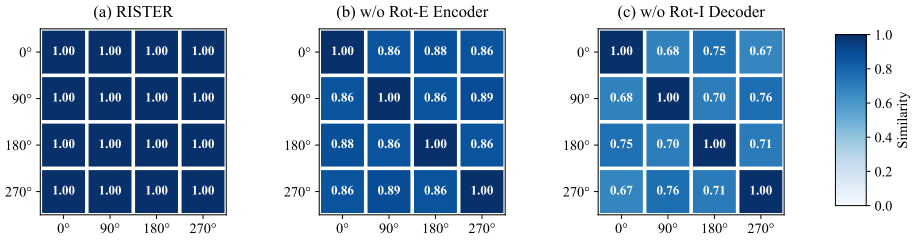

Fig. 4: Cosine similarity between the output logits before and after image rotation. A higher similarity indicates greater consistency between the model outputs. A similarity of 1 means that the predictions and their corresponding confidence scores are exactly identical before and after image rotation. Measured on the IC13 dataset.

Table 2: Ablation study on the encoder–decoder framework. Measured on IC13.

Method	0°	90°	180°	270°	Avg.
(a): w/o Rot-E Encoder	**98.4**	97.9	95.6	96.0	96.98
(b): w/o Rot-I Decoder	98.0	91.2	95.6	90.5	93.82
(c): (a) + random-rotation	97.5	96.9	97.7	96.9	97.25
(d): (b) + random-rotation	97.3	93.6	97.9	93.1	95.48
(e): RISTER-S (Ours)	**98.4**	**98.4**	**98.4**	**98.4**	**98.40**

in RELG with standard convolutions, thereby breaking the rotation-equivariant property of the encoding process; (b) replacing RITD with a CTC-based decoder [13,37], thereby destroying the rotation-invariant property of the decoding process. As shown in Figure 4(b, c), the absence of either the Rot-E encoder or the Rot-I decoder causes the model's outputs to vary with the rotation of the input image, indicating that the model loses end-to-end rotation invariance. The quantitative results presented in Table 2 further corroborate the effectiveness of RISTER. Owing to its intrinsic rotation invariance, RISTER attains consistent recognition accuracy across images of varying orientations. This equivariant recognition capability enables robust performance on multi-oriented text. In contrast, the baseline models (a, b) exhibit a substantial decline in recognition accuracy when subjected to image rotations. We also experimented with incorporating random rotation augmentation during training, as adopted in AON [3]. As shown in Table 2, random rotation augmentation effectively improves the model's performance on multi-oriented samples. However, with the model capacity fixed, reducing the proportion of horizontally oriented samples in the training data encourages the model to learn more orientation-invariant features, which in turn leads to a degradation in recognition accuracy on horizontal samples. In contrast, our RISTER achieves the highest recognition accuracy at every orientation. These observations suggest that incorporating rotation invariance into the model architecture is a more principled approach to multi-oriented recognition than relying on data-driven strategies.

Table 3: Results of encoders comparison. † indicates that the global blocks in RELG are replaced with an equal number of local blocks.

Encoder	IC13				SVT				U14M-Mul-Ori	Params ($\times 10^6$)
	0°	90°	180°	270°	0°	90°	180°	270°		
(a): Resnet45	97.1	94.5	89.5	93.2	95.5	89.8	82.8	88.7	90.8	14.4
(b): (a) + G-CNN [4]	97.2	97.2	97.2	97.2	94.9	94.9	94.9	94.9	90.9	14.4
(c): (a) + PDO-eConv [33]	97.0	97.0	97.0	97.0	94.8	94.8	94.8	94.8	91.2	14.4
(d): (a) + F-Conv [49]	97.3	97.3	97.3	97.3	95.7	95.7	95.7	95.7	94.9	14.4
(e): Vision Transformer [6]	98.5	98.1	95.2	96.1	97.4	96.6	97.3	94.0	93.9	26.9
(f): (e) + Stand-Alone [32]	98.3	98.3	98.3	98.3	97.1	97.1	97.1	97.1	94.5	27.2
(g): (e) + LieTransformer [17]	97.9	97.9	97.9	97.9	96.4	96.4	96.4	96.4	94.0	28.5
(h): SVTR [9]	**98.8**	98.2	97.4	97.0	98.1	96.9	93.2	95.8	94.6	28.2
(i): RELG†	98.0	98.0	98.0	98.0	96.4	96.4	96.4	96.4	95.0	28.4
(j): RELG (Ours)	98.6	**98.6**	**98.6**	**98.6**	**98.3**	**98.3**	**98.3**	**98.3**	**96.6**	32.8

4.3 Effectiveness of the proposed RELG

RELG is an efficient rotation-equivariant backbone tailored for STR. To verify its effectiveness, we compare it with existing Rot-E backbone networks [4,32,33,49]. Since most existing Rot-E networks are primarily designed for classification tasks, their model scales differ by orders of magnitude from those commonly used in STR. To ensure a fair comparison, we incorporate their rotation-equivariant parameterized layers into the standard STR backbone ResNet-45 [14] or ViT [6], thereby constructing Rot-E encoders with comparable model capacity. All variants adopt the RITD design proposed in Section 3.3 as the decoder. As shown in Table 3 (a-d), by incorporating Rot-E convolutions into existing convolution-based networks, the models exhibit rotation equivariance at the encoding stage. When combined with RITD, this property further enables end-to-end rotation invariance, resulting in consistent recognition performance across images with different orientations, and consequently improves the recognition accuracy on multi-oriented samples. However, due to the lack of global modeling of visual features, purely convolutional backbones are not the optimal choice for STR. As shown in Table 3 (e, h), backbones equipped with self-attention, such as ViT and SVTR, achieve significantly higher performance on images with 0° orientation compared to those in Table 3 (a–d). This observation motivates us to incorporate attention mechanisms into the Rot-E backbone. Specifically, We reconstruct the ViT by leveraging existing group-equivariant self-attention methods [17,32]. As illustrated in Table 3 (f, g), after introducing rotation equivariance into ViT, the models achieve improved recognition performance on multi-oriented samples; however, their accuracy on 0° samples decreases. We attribute this behavior to the fact that an overly shallow feature extraction process compresses the representational degrees of freedom of the model, while the early downsampling leads to the loss of local details, making it difficult for subsequent self-attention layers to recover the missing local geometric information.

Unlike the aforementioned backbones, our proposed RELG adopts a deeper equivariant convolutional network and employs progressive downsampling, which preserves geometric details of the input images to the greatest extent. Meanwhile, the incorporation of self-attention enables the model to effectively capture intra-character patterns in text images. As a result, our model achieves strong performance on both 0° and multi-oriented samples. Compared with existing backbones, RELG attains an accuracy of 96.6% on U14M-Multi-Oriented, outperforming the second-best encoder SVTR by 2%, while achieving comparable recognition accuracy to SVTR on 0° samples. Furthermore, we validate the role of self-attention by replacing the Global Blocks in RELG with Local Blocks. As shown in Table 3 (i), removing self-attention leads to degraded feature extraction capability, resulting in performance drops of 0.6%, 1.9%, and 1.6% on IC13, SVT, and U14M-Multi-Oriented, respectively. These results further demonstrate the effectiveness of our design.

4.4 Comparison of Computational Cost and Efficiency

We analyze the computational efficiency of RISTER. Using group-equivariant convolutions (a) incurs the same inference cost as standard convolutions (b), because after training, the parameters are fixed through cyclic shifting, making the inference process identical to that of standard convolutions. Consequently, the FLOPs, parameter count, and FPS of RISTER are comparable to those of existing AR-based STR methods (e.g., NRTR), as reported in the Table 4. This demonstrates that RISTER achieves rotation-invariant recognition without introducing additional computational overhead, highlighting the advantage of our model in terms of inference efficiency.

Table 4: Comparison of computational cost and inference speed. Measured on IC13.

	FLOPs ($\times 10^9$)	Params ($\times 10^6$)	FPS
(a) RISTER-S (with F-Conv)	12.99	21.8	31.4
(b) RISTER-S (with standard conv.)	12.99	21.8	31.4
(c) NRTR	13.06	44.3	28.9

4.5 Influence of Input Resolution

Most STR models adopt an input resolution of 32×128, which results in visual feature maps with a height of only 4 pixels after feature extraction. When handling multi-oriented images, this leads to insufficient feature resolution for vertically oriented text, thereby degrading recognition accuracy [18]. In contrast, methods designed for multi-oriented text recognition typically employ square input resolutions, with equal height and width (e.g., 100×100 in AON [3]).

We experimentally investigate the impact of input resolution on model performance. As shown in Table 5, we investigate the performance of two recent

Table 5: Model performance under different input resolutions

Method (Resolution)	Common				Oriented		
	IC13	SVT	CUTE	Avg.	Mul-Ori	ASOT	Avg.
(a) SVTRv2 (32×128) [11]	98.7	98.0	99.0	98.57	89.5	86.5	88.00
(b) SVTRv2 (128×128) [11]	98.6	98.1	98.8	98.50	90.2	87.1	88.65
(c) IGTR (32×128) [10]	98.6	98.3	97.6	98.17	92.6	90.1	91.35
(d) IGTR (128×128) [10]	98.6	98.4	97.7	98.20	92.9	90.5	91.70
(e) RISTER-S (128×128)	98.6	98.3	98.3	98.40	96.0	93.3	94.65

Table 6: Comparison results with Oriented Text Recognizers.

Method	Common				Oriented		
	IC13	SVT	CUTE	Avg.	Mul-Ori	ASOT	Avg.
AON [3]	91.5	82.8	76.8	83.70	-	-	-
SLOAN [5]	96.0	96.1	92.4	94.83	91.1	90.2	90.65
ASTER [39]	96.2	94.6	93.4	94.80	78.8	80.8	79.80
SVTRv2 [11]	**98.7**	98.0	**99.0**	**98.57**	89.5	86.5	88.00
RISTER-S	98.6	**98.3**	98.3	98.40	**96.0**	**93.3**	**94.65**

STR models, SVTRv2 [11] and IGTR [10], under different input resolutions. The experimental results indicate that increasing the resolution to 128×128 improves recognition performance on multi-oriented samples, while the recognition accuracy on horizontally oriented images remains almost unchanged. Therefore, we follow this setting and set the input resolution of RISTER to 128×128. Under the same input resolution, our RISTER still outperforms the latest models on multi-oriented samples by 6.00% and 2.95% in accuracy, while achieving very comparable performance on common benchmarks. This demonstrates the superiority of integrating rotation invariance into the model.

4.6 Comparison with Oriented Text Recognizers

Some existing STR methods are specifically designed for multi-oriented text, such as multi-orientation feature extraction [3], polar transformation [5], or rectification modules [11, 39]. We compare these methods with our proposed RISTER, as shown in Table 6. The results show that on common samples, RISTER achieves recognition performance very close to the existing state-of-the-art SVTRv2 [11]. On oriented samples, our method significantly outperforms existing approaches, achieving a 4.00% improvement over the second-best method. This demonstrates that the proposed RISTER not only handles oriented samples effectively but also delivers impressive performance on common samples.

4.7 Comparison with State-of-the-arts

To evaluate the performance of our RISTER on general benchmarks, we compare it with existing and widely used STR methods. The results are shown in Table 7.

Table 7: Comparison on Union14M-Benchmarks & Common Benchmarks. **Bold** and underlined values denote the 1st and 2nd results in each column.

Method	Union14M-Benchmarks							Common Benchmarks						Avg.	Params ($\times 10^6$)
	CUR	MO	ART	CTL	SAL	MTW	GEN	IC13	SVT	IIIT	IC15	SVTP	CUTE		
CRNN [37]	19.4	4.5	34.2	44.0	16.7	35.7	60.4	91.8	83.8	90.8	71.8	70.4	80.9	54.18	8.3
ASTER [39]	74.2	78.8	61.3	65.9	75.9	70.6	76.9	96.2	94.6	97.6	87.4	88.7	93.4	81.66	19.1
NRTR [35]	67.9	42.4	66.5	73.6	66.4	77.2	78.3	97.8	96.8	98.1	88.9	93.3	94.4	80.12	44.3
SAR [21]	73.2	63.5	60.1	68.8	64.2	75.4	74.7	96.7	94.0	97.7	84.8	84.5	94.4	82.67	57.5
RoScanner [55]	79.4	68.1	70.5	79.6	71.6	82.5	80.8	97.7	95.8	98.5	88.2	90.1	97.6	84.65	48.0
SRN [54]	78.1	63.2	66.3	65.3	71.4	58.3	76.5	97.5	96.3	97.2	87.9	90.9	96.9	80.44	51.7
SEED [30]	69.1	80.9	56.9	63.9	73.4	61.3	76.5	94.2	93.2	96.5	87.5	88.7	93.4	79.66	24.0
VisionLAN [45]	79.6	71.4	67.9	73.7	76.1	73.9	79.1	97.1	95.8	98.2	88.6	91.2	96.2	83.75	32.9
PIMNet [29]	80.3	79.8	68.4	75.9	77.8	68.3	80.8	98.4	97.8	98.9	89.6	94.3	97.9	85.24	24.2
ABINet [12]	82.7	77.4	68.8	69.2	77.9	72.0	77.9	97.8	97.7	98.1	89.4	93.2	97.6	84.59	36.9
SVTR [9]	79.3	69.4	67.1	68.8	75.8	76.8	77.0	96.7	96.7	98.2	88.6	91.3	96.5	83.27	18.1
PARSeq [1]	87.6	88.3	72.3	77.4	84.0	80.8	82.6	97.8	97.2	98.7	90.6	94.6	96.8	88.40	23.8
MATRN [26]	82.2	73.0	73.4	76.9	79.4	77.4	81.0	97.9	**98.3**	98.8	90.3	95.2	97.2	86.24	44.3
MGP-STR [44]	85.2	83.7	72.6	75.1	79.8	71.1	83.1	97.1	97.8	97.9	89.6	95.2	96.9	86.54	148.0
LPV [57]	86.2	78.7	75.8	80.2	82.9	81.6	82.9	98.1	97.8	98.6	89.8	93.6	97.6	88.01	30.5
MAERec [18]	81.4	71.4	72.0	80.0	78.5	82.4	82.5	97.6	96.8	98.0	87.1	93.2	97.9	86.22	35.7
LISTER [2]	87.1	88.2	72.5	78.3	79.7	81.3	80.3	97.3	96.6	98.5	88.2	90.7	96.5	86.32	20.5
CAM [52]	85.4	89.0	72.0	75.4	84.0	74.8	83.1	96.6	96.1	98.2	89.0	93.5	96.2	87.18	58.7
CDistNet [60]	81.7	77.1	72.6	78.2	79.9	79.7	81.1	97.8	98.1	98.7	89.6	93.5	96.9	86.45	43.3
BUSNet [46]	83.0	82.3	70.8	77.9	78.8	71.2	82.6	97.8	98.1	98.3	90.2	95.3	96.5	86.39	32.1
OTE [50]	87.9	82.8	75.3	73.4	81.8	68.9	80.9	97.7	97.8	98.2	89.6	94.4	97.9	86.66	20.0
IGTR [10]	90.3	92.6	77.5	78.8	85.6	81.8	82.9	98.6	**98.3**	98.9	90.4	94.7	97.6	89.85	24.1
CPPD [8]	87.9	81.7	75.5	76.2	83.5	81.5	81.9	98.2	98.1	98.8	**91.1**	93.8	97.6	88.14	26.9
IPAD [53]	85.2	83.3	72.1	78.4	81.4	73.7	82.2	98.1	97.7	99.0	90.8	95.5	97.9	87.33	24.7
SVTRv2 [11]	91.0	89.5	78.7	**81.3**	85.9	85.1	82.3	98.7	98.0	**99.2**	91.0	93.6	**99.0**	90.30	21.0
SMTR [7]	90.5	92.6	75.5	80.0	84.9	85.2	82.6	98.4	98.2	98.8	90.0	93.3	97.6	88.93	15.8
RISTER-T	89.4	94.5	72.0	74.0	85.2	79.7	80.0	98.3	96.6	98.0	89.2	93.8	95.1	88.13	15.9
RISTER-S	92.5	96.0	76.7	77.8	87.8	84.6	81.9	98.4	97.8	98.6	90.0	95.5	97.2	90.37	21.8
RISTER-B	93.0	**96.6**	78.2	78.6	89.1	86.5	83.0	98.6	**98.3**	98.9	**91.1**	95.2	98.3	91.18	32.8
RISTER-L	**94.1**	**96.6**	**79.4**	80.6	**90.3**	**87.4**	**83.7**	**98.9**	98.1	99.0	90.6	**95.7**	97.6	**92.46**	38.6

As shown by the experimental results, our RISTER outperforms existing STR methods even on the standard 0° benchmarks. Moreover, it achieves superior performance on more challenging datasets, such as U14M-B, reaching state-of-the-art results. Specifically, RISTER-T achieves an average accuracy of 88.13% while maintaining a relatively small number of parameters. Meanwhile, RISTER-S and RISTER-B, as medium-sized models in terms of parameter count, outperform existing STR methods on both CoB and U14M-B benchmarks. When the parameter size is further increased, the resulting RISTER-L becomes a powerful recognition model, achieving the highest accuracy on 8 out of 13 test benchmarks, thereby setting new state-of-the-art performance. Meanwhile, it still maintains a parameter count below 40M. The superiority of RISTER on these 0° test benchmarks can be attributed to two main factors: (1) the well-designed four-stage RELG architecture, which enables efficient feature extraction while maintaining a relatively small number of parameters; and (2) the integration of rotation invariance into the model, which allows RISTER to learn effectively without relying on rotated training samples. This not only improves the efficiency of train-

ing data utilization but also enhances the aggregation of key features, thereby strengthening the model's robustness.

5 Conclusion

In this work, we propose RISTER, a system-level, rotation-invariant framework for STR. RISTER provides theoretical guarantees for multi-oriented text recognition and addresses common limitations of existing methods, such as increased computational cost and heavy reliance on data augmentation. Experimental results demonstrate that RISTER consistently achieves superior performance across both multi-oriented benchmarks and horizontally oriented samples. In future work, we plan to explore extending RISTER-like rotation-invariant frameworks to non-square input settings to support applications such as mathematical expression recognition and long-text recognition.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62302532), Guangdong Basic and Applied Basic Research Foundation (2025A15-15011224), Shenzhen Science and Technology Program (KQTD20221101093559018, SYSRD20250529113401002), Open Research Fund of The State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS2025052), Basic Research Program of Jiangsu (BK20251441), and sponsored by CCF-Baidu Open Fund (OF202519). The model is implemented based on PaddlePaddle.

References

1. Bautista, D., Atienza, R.: Scene text recognition with permuted autoregressive sequence models. In: ECCV. pp. 178–196 (2022)
2. Cheng, C., Wang, P., Da, C., Zheng, Q., Yao, C.: LISTER: neighbor decoding for length-insensitive scene text recognition. In: ICCV. pp. 19484–19494 (2023)
3. Cheng, Z., Xu, Y., Bai, F., Niu, Y., Pu, S., Zhou, S.: AON: towards arbitrarily-oriented text recognition. In: CVPR. pp. 5571–5579 (2018)
4. Cohen, T., Welling, M.: Group equivariant convolutional networks. In: ICML. vol. 48, pp. 2990–2999 (2016)
5. Dai, P., Zhang, H., Cao, X.: SLOAN: scale-adaptive orientation attention network for scene text recognition. IEEE Trans. Image Process. **30**, 1687–1701 (2021)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
7. Du, Y., Chen, Z., Jia, C., Gao, X., Jiang, Y.: Out of length text recognition with sub-string matching. In: AAAI. pp. 2798–2806 (2025)
8. Du, Y., Chen, Z., Jia, C., Yin, X., Li, C., Du, Y., Jiang, Y.: Context perception parallel decoder for scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell. **47**(6), 4668–4683 (2025)
9. Du, Y., Chen, Z., Jia, C., Yin, X., Zheng, T., Li, C., Du, Y., Jiang, Y.: SVTR: scene text recognition with a single visual model. In: IJCAI. pp. 884–890 (2022)
10. Du, Y., Chen, Z., Su, Y., Jia, C., Jiang, Y.: Instruction-guided scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell. **47**(4), 2723–2738 (2025)
11. Du, Y., Chen, Z., Xie, H., Jia, C., Jiang, Y.G.: Svtrv2: Ctc beats encoder-decoder models in scene text recognition. In: ICCV. pp. 20147–20156 (2025)
12. Fang, S., Xie, H., Wang, Y., Mao, Z., Zhang, Y.: Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In: CVPR. pp. 7098–7107 (2021)
13. Graves, A., Fernández, S., Gomez, F.J., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: ICML. pp. 369–376 (2006)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
15. He, L., Chen, Y., Shen, Z., Dong, Y., Wang, Y., Lin, Z.: Efficient equivariant network. In: NeurIPS. pp. 5290–5302 (2021)
16. Hoogeboom, E., Peters, J.W.T., Cohen, T.S., Welling, M.: Hexaconv. In: ICLR (2018)
17. Hutchinson, M.J., Lan, C.L., Zaidi, S., Dupont, E., Teh, Y.W., Kim, H.: Lietransformer: Equivariant self-attention for lie groups. In: ICML. vol. 139, pp. 4533–4543 (2021)
18. Jiang, Q., Wang, J., Peng, D., Liu, C., Jin, L.: Revisiting scene text recognition: A data perspective. In: ICCV. pp. 20486–20497 (2023)
19. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S.K., Bagdanov, A.D., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., Shafait, F., Uchida, S., Valveny, E.: ICDAR 2015 competition on robust reading. In: ICDAR. pp. 1156–1160 (2015)
20. Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L.G., Mestre, S.R., Mas, J., Mota, D.F., Almazán, J., de las Heras, L.: ICDAR 2013 robust reading competition. In: ICDAR. pp. 1484–1493 (2013)

21. Li, H., Wang, P., Shen, C., Zhang, G.: Show, attend and read: A simple and strong baseline for irregular text recognition. In: AAAI. pp. 8610–8617 (2019)
22. Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florêncio, D.A.F., Zhang, C., Li, Z., Wei, F.: Trocr: Transformer-based optical character recognition with pre-trained models. In: AAAI. pp. 13094–13102 (2023)
23. Li, S., Davies, M., Yaghoobi, M.: Equivariant imaging for self-supervised hyperspectral image inpainting. CoRR **abs/2404.13159** (2024)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
25. Mishra, A., Alahari, K., Jawahar, C.V.: Scene text recognition using higher order language priors. In: BMVC. pp. 1–11 (2012)
26. Na, B., Kim, Y., Park, S.: Multi-modal text recognition networks: Interactive enhancements between visual and semantic features. In: ECCV. pp. 446–463 (2022)
27. Phan, T.Q., Shivakumara, P., Tian, S., Tan, C.L.: Recognizing text with perspective distortion in natural scenes. In: ICCV. pp. 569–576 (2013)
28. Posner, I., Corke, P., Newman, P.M.: Using text-spotting to query the world. In: IROS. pp. 3181–3186 (2010)
29. Qiao, Z., Zhou, Y., Wei, J., Wang, W., Zhang, Y., Jiang, N., Wang, H., Wang, W.: Pimnet: A parallel, iterative and mimicking network for scene text recognition. In: ACM MM. pp. 2046–2055 (2021)
30. Qiao, Z., Zhou, Y., Yang, D., Zhou, Y., Wang, W.: SEED: semantics enhanced encoder-decoder framework for scene text recognition. In: CVPR. pp. 13525–13534 (2020)
31. Risnumawan, A., Shivakumara, P., Chan, C.S., Tan, C.L.: A robust arbitrary text detection system for natural scene images. Expert Syst. Appl. **41**(18), 8027–8048 (2014)
32. Romero, D.W., Cordonnier, J.: Group equivariant stand-alone self-attention for vision. In: ICLR (2021)
33. Shen, Z., He, L., Lin, Z., Ma, J.: Pdo-econvs: Partial differential operator based equivariant convolutions. In: ICML. vol. 119, pp. 8697–8706 (2020)
34. Shen, Z., Shen, T., Lin, Z., Ma, J.: Pdo-es2cnns: Partial differential operator based equivariant spherical cnns. In: AAAI. pp. 9585–9593 (2021)
35. Sheng, F., Chen, Z., Xu, B.: NRTR: A no-recurrence sequence-to-sequence model for scene text recognition. In: ICDAR. pp. 781–786 (2019)
36. Shenoy, A., Lu, Y., Jayakumar, S., Chatterjee, D., Moslehpour, M., Chuang, P., Harpale, A., Bhardwaj, V., Xu, D., Zhao, S., Zhao, L., Ramchandani, A., Dong, X.L., Kumar, A.: Lumos: Empowering multimodal llms with scene text recognition. In: KDD. pp. 5690–5700 (2024)
37. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(11), 2298–2304 (2017)
38. Shi, B., Wang, X., Lyu, P., Yao, C., Bai, X.: Robust scene text recognition with automatic rectification. In: CVPR. pp. 4168–4176 (2016)
39. Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: ASTER: an attentional scene text recognizer with flexible rectification. IEEE Trans. Pattern Anal. Mach. Intell. **41**(9), 2035–2048 (2019)
40. Sokolić, J., Giryes, R., Sapiro, G., Rodrigues, M.R.D.: Generalization error of deep neural networks: Role of classification margin and data structure. In: SampTA. pp. 147–151 (2017)
41. Su, Y., Chen, Z., Shao, Z., Du, Y., Ji, Z., Bai, J., Zhou, Y., Jiang, Y.G.: Lranet: Towards accurate and efficient scene text detection with low-rank approximation

network. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4979–4987 (2024)
42. Terris, M., Moreau, T., Pustelnik, N., Tachella, J.: Equivariant plug-and-play image reconstruction. In: CVPR. pp. 25255–25264 (2024)
43. Wang, K., Babenko, B., Belongie, S.J.: End-to-end scene text recognition. In: ICCV. pp. 1457–1464 (2011)
44. Wang, P., Da, C., Yao, C.: Multi-granularity prediction for scene text recognition. In: ECCV. pp. 339–355 (2022)
45. Wang, Y., Xie, H., Fang, S., Wang, J., Zhu, S., Zhang, Y.: From two to one: A new scene text recognizer with visual language modeling network. In: ICCV. pp. 14174–14183 (2021)
46. Wei, J., Zhan, H., Lu, Y., Tu, X., Yin, B., Liu, C., Pal, U.: Image as a language: Revisiting scene text recognition via balanced, unified and synchronized vision-language reasoning network. In: AAAI. pp. 5885–5893 (2024)
47. Weiler, M., Cesa, G.: General e(2)-equivariant steerable cnns. In: NeurIPS. pp. 14334–14345 (2019)
48. Weiler, M., Hamprecht, F.A., Storath, M.: Learning steerable filters for rotation equivariant cnns. In: CVPR. pp. 849–858 (2018)
49. Xie, Q., Zhao, Q., Xu, Z., Meng, D.: Fourier series expansion based filter parametrization for equivariant convolutions. IEEE Trans. Pattern Anal. Mach. Intell. **45**(4), 4537–4551 (2023)
50. Xu, J., Wang, Y., Xie, H., Zhang, Y.: OTE: exploring accurate scene text recognition using one token. In: CVPR. pp. 28327–28336 (2024)
51. Xu, R., Yang, K., Liu, K., He, F.: E(2)-equivariant vision transformer. In: UAI. vol. 216, pp. 2356–2366 (2023)
52. Yang, M., Yang, B., Liao, M., Zhu, Y., Bai, X.: Class-aware mask-guided feature refinement for scene text recognition. Pattern Recognit. **149**, 110244 (2024)
53. Yang, X., Qiao, Z., Zhou, Y.: IPAD: iterative, parallel, and diffusion-based network for scene text recognition. Int. J. Comput. Vis. **133**(8), 5589–5609 (2025)
54. Yu, D., Li, X., Zhang, C., Liu, T., Han, J., Liu, J., Ding, E.: Towards accurate scene text recognition with semantic reasoning networks. In: CVPR. pp. 12110–12119 (2020)
55. Yue, X., Kuang, Z., Lin, C., Sun, H., Zhang, W.: Robustscanner: Dynamically enhancing positional clues for robust text recognition. In: ECCV. pp. 135–151 (2020)
56. Zhan, F., Lu, S.: ESIR: end-to-end scene text recognition via iterative image rectification. In: CVPR. pp. 2059–2068 (2019)
57. Zhang, B., Xie, H., Wang, Y., Xu, J., Zhang, Y.: Linguistic more: Taking a further step toward efficient and accurate scene text recognition. In: IJCAI. pp. 1704–1712 (2023)
58. Zhang, C., Tao, Y., Du, K., Ding, W., Wang, B., Liu, J., Wang, W.: Character-level street view text spotting based on deep multisegmentation network for smarter autonomous driving. IEEE Trans. Artif. Intell. **3**(2), 297–308 (2022)
59. Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Zhang, K., Xu, S., Chen, D., Timofte, R., Gool, L.V.: Equivariant multi-modality image fusion. In: CVPR. pp. 25912–25921 (2024)
60. Zheng, T., Chen, Z., Fang, S., Xie, H., Jiang, Y.: Cdistnet: Perceiving multi-domain character distance for robust text recognition. Int. J. Comput. Vis. **132**(2), 300–318 (2024)