electronics

MDPI

Article

Towards Robust Scene Text Recognition: A Dual Correction Mechanism with Deformable Alignment

Yajiao Feng [1,*] and Changlu Li [2,*]

[1] School of Microelectronics, Tianjin University, Tianjin 300072, China
[2] School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
* Correspondence: fyj3398@tju.edu.cn (Y.F.); changlu@tju.edu.cn (C.L.)

Abstract

Scene Text Recognition (STR) faces significant challenges under complex degradation conditions, such as distortion, occlusion, and semantic ambiguity. Most existing methods rely heavily on language priors for correction, but effectively constructing language rules remains a complex problem. This paper addresses two key challenges: (1) The over-correction behavior of language models, particularly on semantically deficient input, can result in both recognition errors and loss of critical information. (2) Character misalignment in visual features, which affects recognition accuracy. To address these problems, we propose a Deformable-Alignment-based Dual Correction Mechanism (DADCM) for STR. Our method includes the following key components: (1) We propose a visually guided and language-assisted correction strategy. A dynamic confidence threshold is used to control the degree of language model intervention. (2) We designed a visual backbone network called SCRTNet. The net enhances key text regions through a channel attention module (SENet) and applies deformable convolution (DCNv4) in deep layers to better model distorted or curved text. (3) We propose a deformable alignment module (DAM). The module combines Gumbel-Softmax-based anchor sampling and geometry-aware self-attention to improve character alignment. Experiments on multiple benchmark datasets demonstrate the superiority of our approach. Especially on the Union14M-Benchmark, where the recognition accuracy surpasses previous methods by 1.1%, 1.6%, 3.0%, and 1.3% on the Curved, Multi-Oriented, Contextless, and General subsets, respectively.

Keywords: scene text recognition; dual correction; fusion; deformable; alignment

Academic Editor: Dah-Jye Lee

Received: 9 September 2025
Revised: 2 October 2025
Accepted: 7 October 2025
Published: 9 October 2025

Citation: Feng, Y.; Li, C. Towards Robust Scene Text Recognition: A Dual Correction Mechanism with Deformable Alignment. Electronics **2025**, *14*, 3968. https://doi.org/10.3390/electronics14193968

1. Introduction

Text recognition is an important research area in computer vision and natural language processing. In particular, STR focuses on accurately extracting text information from images. This task plays a key role in many AI applications, such as automated document processing, license plate recognition, intelligent surveillance, and autonomous driving [1–3]. In recent years, with the rapid development of deep learning, STR has achieved remarkable progress across various fields [4–6].

However, in complex natural scenes, text images often present various challenging features. As shown in Figure 1, these include complex fonts, blurriness, occlusion, distortion, and lack of semantic content. These factors make scene text recognition difficult. As shown in Figure 2, existing STR methods can be divided into two main categories: visual feature learning [7–12] and semantic understanding [13–20].

Figure 1. Various types of scene text images.

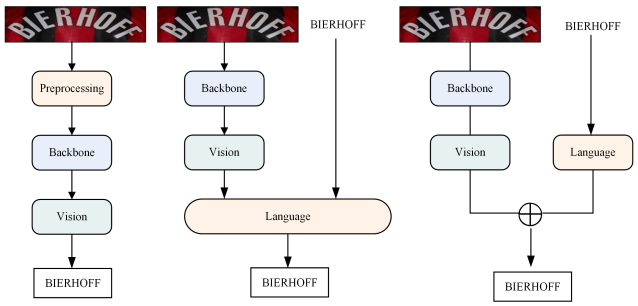

Figure 2. Different types of STR methods. From left to right: the pipelines of a pre-processing visual model, a model with separate visual and linguistic modules, and a vision–language fusion model.

Visual methods mainly use preprocessing to correct distortions and introduce some spatial variations to enhance the image feature extraction capability [9,13,21]. However, these methods often fail to accurately model semantic dependencies between characters when the text is blurry or structurally degraded.

On the other hand, semantic understanding methods [14,15,22–24] introduce semantic information implicitly or explicitly to address the shortcomings of visual models in the presence of blurry or incomplete visual inputs. These methods significantly improve the accuracy of scene text recognition.

Despite these advances, existing methods still have the following limitations: (1) Over-integration of language models makes it difficult to distinguish between visual and language signals, leading to poor interpretability. (2) Over-correction by language models introduces too much semantic information from model architecture or dataset corpus, which can cause the model to predict text that follows grammatical rules, even for non-semantic text, such as abbreviations or random combinations of letters, numbers, and symbols. (3) The lack of targeted correction in language models: some methods' iterative corrections negatively impact inference speed.

To address the aforementioned challenges, we propose DADCM, a method designed from three key perspectives. The overall workflow of our approach is illustrated in Figure 3. Specifically, DADCM aims to achieve the following: (1) Visual feature extraction is more accurate, improving the visual model's fine-grained ability to model complex text and enhancing its robustness to blurry and distorted characters. (2) Language correction becomes more intelligent, with visual information as the primary guide and the language model providing auxiliary correction, avoiding misjudgment of non-semantic text caused by excessive reliance on language priors. (3) Features and characters are better aligned to handle various forms of text images.

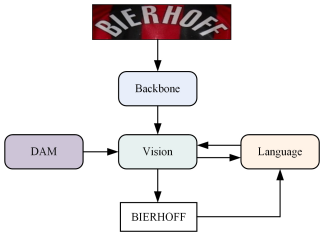

Figure 3. DADCM. The pipeline of our DADCM, which integrates a deformable alignment module for bidirectional correction.

Therefore, the main contributions of this paper are as follows:

(1) We propose a novel visually guided and language-assisted framework that balances visual predictions and linguistic priors in a structured manner. This framework improves recognition robustness under noisy, ambiguous, or incomplete visual inputs, and its design is compatible with potential extensions to more complex STR systems.

(2) To enhance feature extraction, we designed SCRTNet, a lightweight backbone that integrates channel attention with deformable convolution (DCNv4) [25]. This task-driven combination strengthens feature discrimination in distorted, curved, or perspective-deformed text regions, allowing the model to capture complex text patterns more effectively.

(3) We also introduce the Dynamic Anchor-based Alignment Module (DAM), which combines Gumbel-Softmax-based dynamic anchor sampling, self-attention refinement, and Gaussian-weighted feature extraction. This module explicitly addresses character–feature misalignment and significantly improves recognition performance for irregular text shapes.

(4) Extensive experiments across multiple STR benchmarks demonstrate the practical effectiveness of our approach. Especially on the Union14M-Benchmark, recognition accuracy surpasses previous methods by 1.1%, 1.6%, 3.0%, and 1.3% on the Curved, Multi-Oriented, Contextless, and General subsets, respectively, highlighting both the conceptual novelty and empirical impact of our method.

2. Methods

The architecture of DADCM is shown in Figure 4. It consists of four main components: the Visual Feature Extraction Backbone Network, the Language Correction Module, the Character Inference Feature Module, and the Vision–Language Joint Optimization.

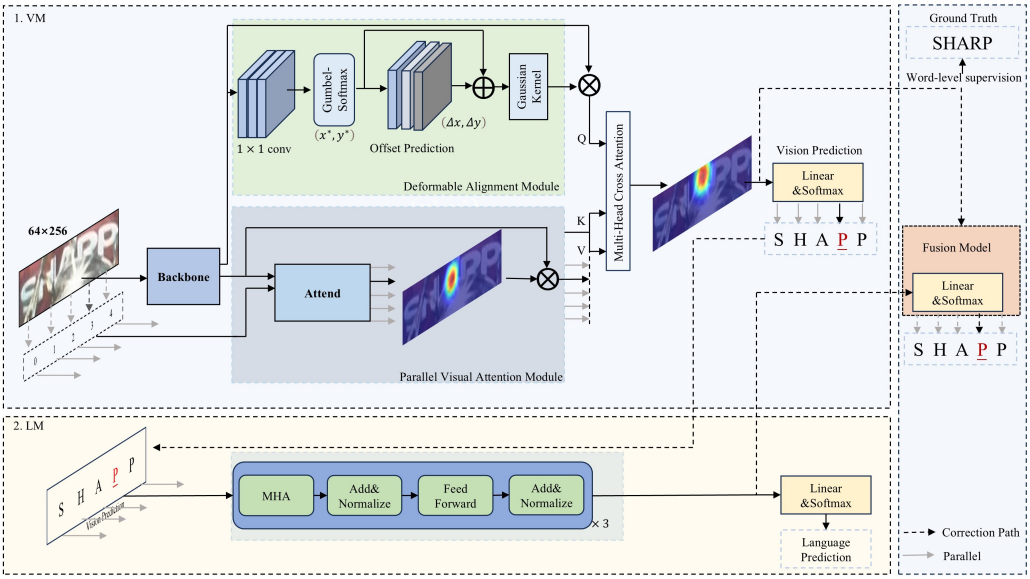

Figure 4. The architecture of the proposed DADCM. The Offset Prediction module consists of 1 × 1 convolutions, self-attention mechanism, and a fully connected (FC) layer.

2.1. Visual Model

We propose a novel visual feature extraction backbone, named SCRTNet, which integrates local and global information. The detailed structure is shown in Figure 5. SCRTNet is built upon a hybrid architecture that combines ResNet [20,26] and Transformer-based modules [22–24,27,28], enabling more powerful and robust visual feature representation.

In the early stage of feature extraction, specifically after the second stage of ResNet, we introduce a channel attention module (Squeeze-and-Excitation Module, SENet) [29]. This module

models the importance of each channel and reweights the responses accordingly. It enhances the responses related to text regions while suppressing redundant background information.

Furthermore, in the fourth stage of ResNet, we apply DCNv4 in the second convolutional layer of each residual block. Unlike standard convolutions, DCNv4 samples features with learnable offsets, allowing the receptive field to adapt to irregular or distorted text patterns. This significantly improves the network's ability to handle common text variations, such as rotation, curvature, and perspective distortion in natural scenes.

The Transformer unit has two layers, each incorporating positional encoding, multi-head self-attention (MHSA), and a feed-forward network (FFN). This design enhances the model's ability to capture global spatial structures.

By combining the above modules, we construct a visual feature extractor that integrates local precision and global context awareness. The detailed formulation of the extracted visual feature $F_{v_{i,j}}$ is as follows:

$$F_{v_{i,j}} = T(\mathcal{R}(X)) \in \mathbb{R}^{B \times P \times C} \tag{1}$$

where $X \in \mathbb{R}^{H \times W \times 3}$ denotes the input text image, H and W represent the height and width, respectively. $\mathcal{R}$ denotes the feature representation extracted by SCResNet-46, and T refers to the output of the Transformer encoder. B is the batch size, and C is the number of channels. The length of the resulting visual feature sequence is defined $P = \frac{H}{32} \times \frac{W}{8}$. The final visual features $F_{v_{i,j}}$ subsequently fed into the recognition module for inference.

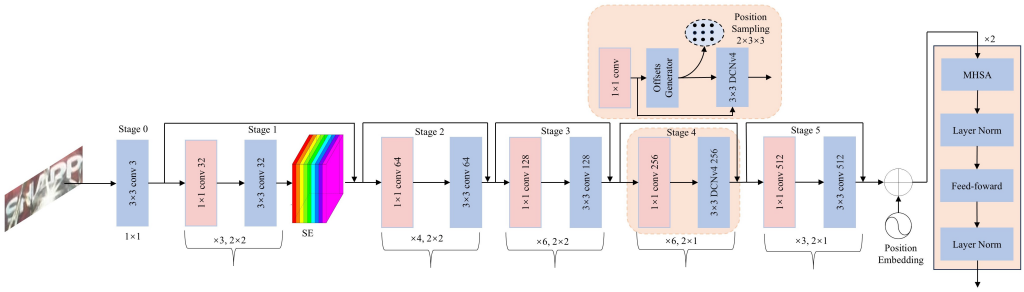

Figure 5. Structure of the Visual Module. Each ResBlock contains Conv-BN-ReLU layers. For clarity, batch normalization and activation functions are omitted in the diagram.

2.2. Recognition and Inference Module

To address alignment challenges caused by text deformation, irregular layout, and uneven character spacing in complex scenes, we propose a Deformable Alignment Module (DAM). This module is integrated with parallel visual attention (PVA) [14,24] to enhance the recognition process. The overall structure is illustrated in Figure 4. DAM derives alignment-enhanced features from $F_{v_{i,j}}$ by employing Gumbel-Softmax [30] differentiable sampling, offset refinement, and Gaussian kernel-based feature aggregation. The detailed process is illustrated in Figure 6.

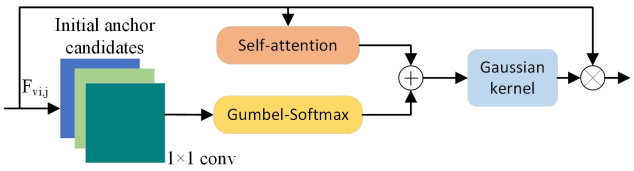

Figure 6. Architecture of the DAM.

First, at each spatial location of the feature map, we apply a three-layer 1×1 convolutional network to predict N_d candidate anchor points (x_i, y_i), along with their confidence scores p_i. Then, we apply Gumbel-Softmax sampling to select the final anchor point in

a differentiable manner. Next, we compute character-level geometric relations using a self-attention-based offset generation module, which dynamically predicts position shifts and yields the refined anchor locations $(\Delta x, \Delta y)$. Finally, we extract the aligned feature at (x', y'). Using a Gaussian kernel-weighted aggregation approach as follows:

$$(x^*, y^*) = \sum_{i=1}^{N_d} p_i \cdot (x_i, y_i) \tag{2}$$

$$\Delta = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \cdot W_{\text{offset}} \tag{3}$$

$$(x', y') = (x^* + \Delta x,\ y^* + \Delta y) \tag{4}$$

$$G(x, y) = \exp\left(-\frac{(x - x')^2 + (y - y')^2}{2\sigma^2}\right) \tag{5}$$

$$G_t = G(x', y') \cdot F_{v_{i,j}} \tag{6}$$

where Q, K, and V are obtained through linear projections of the feature map. d_k is a scaling factor, and W_{offset} is a learnable parameter matrix. The parameter σ adaptively controls the spatial extent of the weighting function. In addition, we obtain the attention map using the Parallel Visual Attention module as follows:

$$p_t = \text{Att}_t^{\text{T}} F_{v_{i,j}} \tag{7}$$

$$\text{Att} = \text{Softmax}\left(G\left(F_{v_{i,j}}\right)\right) \tag{8}$$

$$G\left(F_{v_{i,j}}\right) = W_1 \tanh\left(W_2 O_c + W_3 F_{v_{i,j}}\right) \tag{9}$$

where $Att \in \mathbb{R}^{hw \times N}$ denotes the attention map, h and w are the height and width of the feature map, respectively, and d is the feature dimension.N represents the maximum decoding length. $O_c \in \mathbb{R}^{T \times C}$ is the position encoding of character sequences [31]. W_1, W_2, and W_3 learnable weights, and t refers to the current decoding time step.

Finally, we use the output of DAM as the Q and take the attention-enhanced features from PVA as the K and V. The final fusion is performed as follows:

$$F_a = \text{Softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d}}\right) V \tag{10}$$

The result is passed through a linear layer to produce the visual prediction probability distribution $P_v(y_i)$. Figure 7 shows the visual attention maps generated by the visual module for recognition.

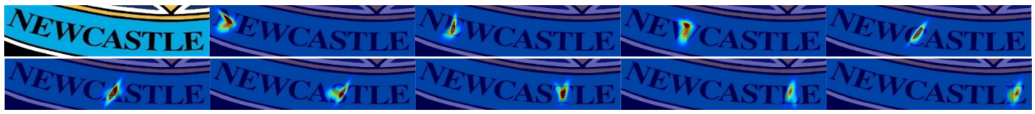

Figure 7. The visual attention maps generated by the visual module for recognition.

2.3. Language Module

In this work, we decouple the visual and language models by introducing an independent language model as a correction module. In our architecture, the context characters $y_{1:i-1}$ are directly derived from visual features. To eliminate potential bias during backpropagation, we apply Blocking Gradient Flow (BGF) [22,23] to isolate learning between different modalities. This ensures the independence of visual and linguistic representa-

tions. Within the correction module, the visual model provides predictions $P(y_i|y_{1:i-1})$ for each character. The visual confidence reflects the reliability of character recognition. Low-confidence predictions may be caused by blur, occlusion, or noise, while high-confidence predictions are typically accurate. Based on these confidence scores, we generate a binary mask that determines whether a character should be corrected. If a character's visual confidence falls below a predefined threshold, it is masked and refined by the language model using contextual information. The attention mechanism within the multi-head blocks is described as follows:

$$M(y_i) = \begin{cases} 0 & \text{if } P(y_i) \geq \tau \\ -\infty & \text{otherwise} \end{cases} \tag{11}$$

$$F_{\text{mha}} = \text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d}} + M\right)V \tag{12}$$

where $Q \in \mathbb{R}^{t\times d}$ denotes either the positional encoding of character sequences from the first layer or the output of the final layer. The K and V are derived from the visual character probabilities $P(y_i)$. $M \in \mathbb{R}^{t\times t}$ is the attention mask used to prevent attention to low-confidence positions. By stacking multiple BCN layers, we obtain a deep architecture that models the bidirectional representation of text F_l [22,23]. In this design, visual confidence is used to dynamically generate the attention mask. The language model performs correction only when necessary, serving as a flexible auxiliary module in scene text recognition.

2.4. Vision-Language Fusion

Scene text contains rich semantic information. Incorporating language models to guide and fuse with visual models can significantly improve recognition accuracy. However, in STR, a significant challenge of introducing language models is preventing the over-correction of originally correct character sequences, especially in cases with weak or no semantic context. To address this issue, we propose a confidence-based correction mechanism. Specifically, we combine outputs from both the visual and language models, and use a predefined confidence threshold τ to dynamically determine which characters should be corrected by the language model. In this way, only low-confidence predictions are refined by the language model. Furthermore, the fused features are used to assess the reliability of the correction.

To integrate features from both visual and linguistic modalities, we adopt a gating mechanism [14,16,22] to fuse the visual features with the language-corrected representations. The gating mechanism employs learnable weights to balance the contributions from each modality. The fusion process is defined as follows:

$$G = \sigma([F_a, F_l])W_f \tag{13}$$

$$F_f = G \odot F_a + (1 - G) \odot F_l \tag{14}$$

where W_f is a learnable parameter. F_a, F_l, and F_f represent the visual features, language-corrected features, and fused features, respectively. σ denotes the sigmoid activation function. $G \in \mathbb{R}^{t\times d}$ is the gated feature, dynamically selected from F_a and F_l.

2.5. Training Objective

The final objective function is defined as follows:

$$\mathcal{L} = \lambda_v \mathcal{L}_v + \lambda_l \mathcal{L}_l + \lambda_f \mathcal{L}_f \tag{15}$$

$$\mathcal{L}_* = -\frac{1}{N}\sum_{t=1}^{N}\log(y_t \mid g_t) \tag{16}$$

$$\mathcal{L}_v = \mathcal{L}_{\text{main}} + \lambda\mathcal{L}_{\text{aux}} \tag{17}$$

$$\mathcal{L}_{\cos} = \begin{cases} 1-\cos(G_t, F_{v_{i,j}}), & \text{if } q = 1 \\ \max(0, \cos(G_t, F_{v_{i,j}}) - m), & \text{if } q = -1 \end{cases} \tag{18}$$

where $\mathcal{L}_v$, $\mathcal{L}_l$, and $\mathcal{L}_f$ represent the losses from the visual module, the language module, and the final fusion module, respectively. y_t and g_t represent the prediction and ground truth. λ_v, λ_l, and λ_f are balancing coefficients. Both $\mathcal{L}_l$ and $\mathcal{L}_f$ adopt the cross-entropy loss as defined in Equation (16). The visual loss $\mathcal{L}_v$ follows a stage-wise hybrid training strategy. The auxiliary loss ($\mathcal{L}_{aux}$) uses cosine embedding loss $\mathcal{L}_{\cos}$ to align the features extracted by PVA and DAM. The main visual loss $\mathcal{L}_{main}$ employs the cross-entropy loss from $\mathcal{L}_*$.

3. Experiment

3.1. Datasets

To enable a more comprehensive comparison with existing STR methods, we train our model on two synthetic datasets, SynthText (ST) [32] and SynthText90K (90K) [33], as well as a real-world training set, Union14M-L [19]. We evaluate DADCM on multiple benchmark datasets that cover a wide range of STR scenarios. These include (1) six commonly used STR benchmarks—ICDAR 2013 (IC13) [34], Street View Text (SVT) [35], IIIT5K-Words (IIIT5K) [36], ICDAR 2015 (IC15) [37], Street View Text-Perspective (SVTP) [38], and CUTE80 [39]. Specifically, we use the version of IC13 that contains 857 images. For IC15, the standard test set with 1811 images is used. (2) The real-world benchmark Union14M-L, which contains seven challenging subsets: Curved, Multi-Oriented (MO), Artistic, Context-less (Cless), Salient, Multi-Word (MW), and General. These subsets present more complex and diverse scene text conditions.

3.2. Implementation Details

The parameter settings used in our experiments are as follows: (1) The feature dimension d of the recognition model is set to 512. (2) The number of candidate points N_d in DAM is set to 5. (3) The offset prediction module uses three layers, each with four self-attention heads. (4) The language model consists of 3 layers, with six attention heads in each layer. (5) The balance factors λ_v, λ_l, and λ_f are set to 0.3, 0.2, and 0.5, respectively.

We resize all input images to 64 × 256, and apply the data augmentation strategy from CDisNet [28]. The model is trained to recognize 94 character classes, including 10 digits, 52 case-sensitive letters, 31 punctuation marks, and one special "END" token. We use case-insensitive word accuracy as the evaluation metric. All STR models, including our DADCM, are trained on the same datasets. The training is conducted in two stages using four NVIDIA RTX 4090 GPUs (NVIDIA, Santa Clara, CA, USA). The operating system was Ubuntu (Canonical, London, UK), the deep learning framework used was PyTorch (v2.2.0, Meta Platforms, Menlo Park, CA, USA), and CUDA (v11.8, NVIDIA, Santa Clara, CA, USA) was employed for GPU acceleration.

In the pretraining stage, we train the visual module separately. We use the AdamW optimizer [40] with a weight decay of 0.05 to pre-train on SynthText (ST) and SynthText90K (90 K) for 20 epochs. The batch size is set to 384, and the initial learning rate is 1.5×10^{-4}. We adopt a cosine learning rate scheduler with two epochs of linear warm-up. Additionally, during the first 6000 iterations, the DAM is frozen, and only the backbone network is trained to ensure stable feature extraction.

In the fine-tuning stage, we incorporate a pre-trained language model [22] into the training process. To keep consistency with the pretraining setup, we continue to use the AdamW optimizer with a weight decay of 0.01. The model is fine-tuned on the Union14M-L training set for 10 epochs.The initial learning rate is set to 1×10^{-4}, and we use a cosine scheduler without warm-up. The batch size is set to 256.

All experiments are conducted on large-scale datasets including ST, 90K, and Union14M-L. Due to the sufficient size of these datasets, training results are stable and not sensitive to random initialization. In addition, all baseline models were retrained under the same training settings to ensure a fair comparison, rather than directly adopting results from the literature.

3.3. Ablation Study on Visual Model

3.3.1. Effectiveness of SCRTNet

To evaluate the impact of different backbone architectures on recognition accuracy, we compare SCRTNet with ResNet+TFs, ConvNeXtV2 [41], and ViT [42] under the same experimental settings as pretraining. For a fair comparison, we disable both the language and alignment modules in all models. The inference process and the overall framework remain the same. We conduct two sets of experiments: one trained only on synthetic datasets and another trained only on Union14M-L. Word accuracy on the challenging IC15 and Union14M-L benchmark subsets is used as the evaluation metric.

Moreover, the results in Table 1 indicate the promising effectiveness of the proposed SCRTNet. We choose SCRTNet as the visual backbone for its balance between feature extraction capability and computational efficiency. The combination of channel attention and DCNv4 allows the network to better capture multi-scale and deformable text features, which is crucial for downstream DAM-based alignment and recognition. Compared to general-purpose encoders such as Swin Transformer, SCRTNet achieves competitive performance while being more lightweight and task-specific. Compared with the other three feature extraction backbones, SCRTNet achieves higher recognition accuracy on each benchmark dataset. When trained on synthetic data, real-world data, and a combination of both, SCRTNet achieves recognition accuracy improvements of 0.8%, 3.3%, and 2.2%, respectively, compared to using ResNet+TFs as the backbone. These results indicate that SCRTNet exhibits better adaptability in recognizing curved and multi-oriented text.

Table 1. Ablation study of different feature extraction networks trained with synthetic datasets and Union14M-L. The Union14M-Benchmark is divided into seven subsets from left to right: Curve, Multi-Oriented, Artistic, Contextless, Salient, Multi-Words, and General. S denotes training with synthetic datasets ST and 901 K, while R denotes training with the Union14M-Train dataset. ViT-S uses six multi-head attention layers, and TF3 indicates the use of three Transformer blocks.

Encoder	Train Data	IC15	Curve	Multi-Oriented	Artistic	Contextless	Salient	Multi-Words	General	Avg
ResNet	S	70.2	67.1	59.6	50.8	70.2	59.8	61.2	60.5	62.4
ResNet+TFs	S	71.0	68.0	60.5	51.9	71.0	60.9	62.5	62.1	63.5
ConvNeXtV2	S	70.8	67.8	60.6	50.9	71.3	59.9	62.3	61.7	63.2
ViT-S	S	69.5	67.7	60.1	50.7	70.1	58.8	61.0	60.1	62.3
SCRTNet (Ours)	S	71.6	68.9	61.2	52.6	72.6	61.1	62.8	63.2	64.3
ResNet	R	82.3	80.7	68.7	60.1	81.3	70.1	76.0	71.2	73.8
ResNet+TFs	R	83.2	81.5	72.6	61.2	82.5	70.9	77.3	72.0	75.1
ConvNeXtV2	R	83.0	81.2	72.2	61.1	82.2	70.7	76.9	71.7	74.9
ViT-S	R	83.2	81.0	71.9	60.2	81.4	70.4	76.2	71.5	74.5
SCRTNet (Ours)	R	85.9	84.1	76.8	64.9	85.4	73.7	81.9	74.6	78.4

Table 1. *Cont.*

Encoder	Train Data	IC15	Curve	Multi-Oriented	Artistic	Contextless	Salient	Multi-Words	General	Avg
ResNet	S+R	83.6	82.2	73.1	61.7	83.3	73.2	77.8	73.5	76.1
ResNet+TFs	S+R	84.9	83.1	74.8	62.3	84.7	74.6	78.5	75.1	77.3
ConvNeXtV2	S+R	84.3	82.9	74.1	62.9	83.6	73.3	77.4	74.9	76.7
ViT-S	S+R	82.7	81.9	73.1	61.6	82.9	72.9	76.6	72.9	75.8
SCRTNet (Ours)	S+R	86.1	85.7	77.2	65.1	86.9	75.1	83.2	76.3	79.5

While our primary comparisons are with ResNet, ConvNeXtV2, and ViT, SCRTNet was selected for its superior trade-off between recognition accuracy and efficiency. Future work could include comparisons with more recent lightweight CNN-Transformer hybrids or Swin-based encoders.

3.3.2. Effectiveness of DAM

DAM is designed to address the challenges posed by the diversity of scene text data. Therefore, it is expected to bring improvements across different types of datasets. To ensure a fair comparison with SCRTNet, we apply the same training settings from Table 1 to the visual backbone enhanced with DAM, using both synthetic datasets and Union14M-L. As shown in Figure 8, the word-level average accuracy improves by 2.3% after integrating DAM. The overall trend of the results clearly demonstrates the effectiveness of DAM.

Interestingly, the gains from DAM vary across subsets. It provides larger improvements on curved and multi-oriented text because these cases suffer from significant character–feature misalignment, which DAM is specifically designed to correct. In contrast, for contextless text, the main challenge is the lack of linguistic cues rather than visual misalignment, while DAM still improves alignment, its effect is limited in low-context scenarios. These observations highlight that DAM is particularly effective for addressing spatial alignment issues in irregular text.

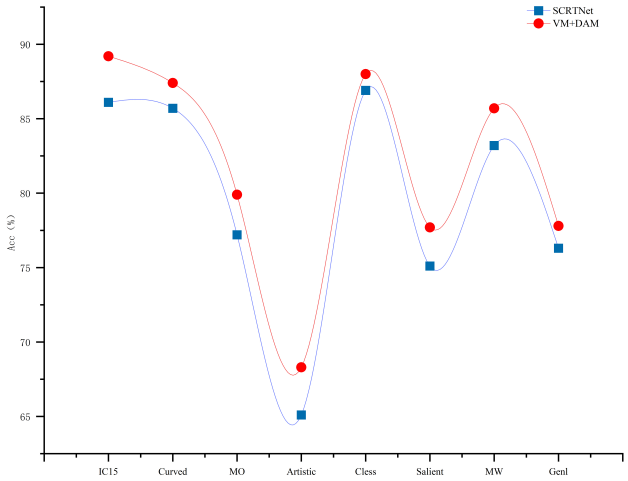

Figure 8. Effectiveness of DAM. Word accuracy comparison between models with and without DAM, trained using both synthetic datasets and Union14M-L10.

3.4. Vision-Language Fusion

3.4.1. Ablation Study on Vision–Language Fusion

The language model is introduced to complement and correct characters that are difficult to extract through visual features alone. The fusion between language and vision also plays a key role in model performance. To evaluate the effectiveness of our dynamic vision–language fusion, we conduct three comparison experiments: (1) visual-only infer-

ence, (2) language-only inference, and (3) fused inference. The results are shown in Table 2. To ensure consistency in evaluation, all three models undergo a two-stage training scheme, including pre-training and fine-tuning.

To further investigate the contribution of the Deformable Alignment Module (DAM), we conducted an ablation study by integrating it into the model with VM, LM, and VLM. As shown in Table 2, adding DAM improves recognition accuracy on the irregular text subsets, including IC15, Curve, Multi-Oriented, and Artistic, demonstrating its effectiveness in aligning visual features for challenging text shapes. The performance on Contextless and general text remains stable, indicating that DAM does not negatively affect simpler or context-free cases. Overall, the inclusion of DAM yields a slight increase in average accuracy, confirming its complementary role in enhancing the model's robustness for complex scene text recognition.

Table 2. Ablation study of the contributions of different modules (VM: Vision Module, LM: Language Module, VLM: Visual–Language Module, DAM/Deformable Alignment Module).

VM	LM	VLM	DAM	IC15	Curve	Multi-Oriented	Artistic	Contextless	Salient	Multi-Words	General	Avg
✓				89.2	87.4	79.9	76.6	86.5	77.7	85.7	77.0	82.5
✓	✓			90.9	89.7	83.4	79.3	85.3	79.8	86.1	79.8	84.3
✓	✓	✓		92.1	89.9	85.5	79.8	88.5	81.3	86.5	83.1	85.8
✓	✓	✓	✓	90.9	89.7	83.4	79.3	85.3	79.8	86.1	79.8	84.3

As shown in Table 2, the Visual–Language Fusion Module (VLM) improves performance over using only the Vision Module, with gains of 2.9%, 2.5%, 5.6%, 3.2%, 2.0%, 3.6%, 0.8%, and 6.1% on IC15, Curve, Multi-Oriented, Artistic, Contextless, Salient, Multi-Words, and General datasets, respectively, and an average improvement of 3.3%. These results demonstrate that the model effectively integrates visual and language information. To better understand the contribution of the Language Module itself, we compare it with the Vision Module alone. The Language Module improves recognition accuracy on most test subsets, with gains of 1.7%, 2.3%, 3.5%, 2.7%, 0.4%, and 2.8% for IC15, Curve, Multi-Oriented, Artistic, Salient, and Multi-Words, respectively. However, a slight decrease of 1.2% is observed on the Contextless subset. This is likely because the Language Module may "over-correct" text that was already correctly recognized, relying on semantic context, while it effectively aids in correcting semantically meaningful text and complements the visual model on difficult cases, scene text often contains visually challenging content without semantic cues. Therefore, the application of the Language Module requires careful consideration.

Therefore, we should not rely solely on language correction, as some scene texts lack semantic content. On the other hand, visual predictions may also be unreliable. A robust vision–language fusion strategy is essential for handling complex text in real-world scenarios. In Table 2, compared to using the visual model alone, the fusion module brings further improvements of 1.2%, 0.2%, 2.1%, 0.5%, 3.2%, 1.5%, 0.4%, and 3.3% on each test subset.

Figure 9 shows the character-level confidence predictions from the visual, linguistic, and fusion modules in DADCM. As each module is added, the confidence scores for character predictions increase to varying degrees. This indicates that the modules complement each other and help the model better understand the text images. The fusion module, in particular, enhances the interaction between visual features and semantic information. As a result, the overall recognition becomes more stable and accurate. These results demonstrate the effectiveness of DADCM in handling complex scene text recognition tasks.

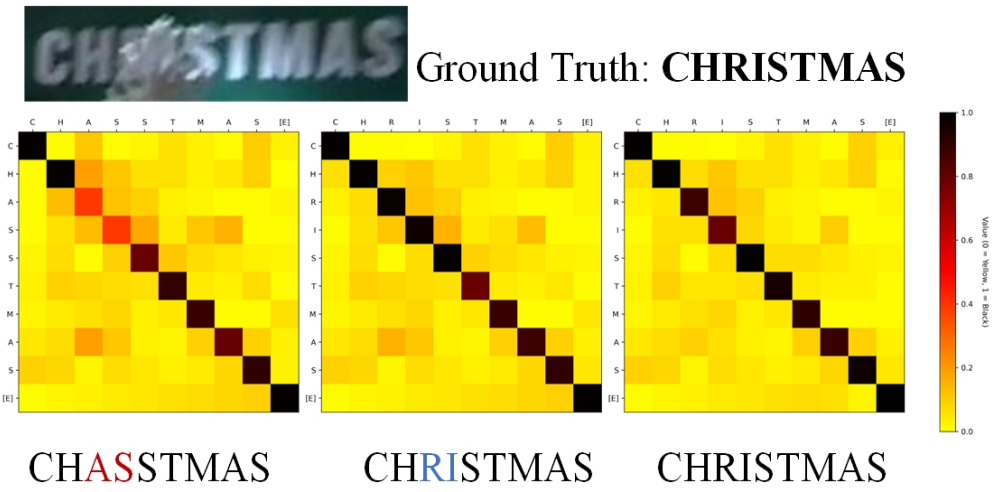

Figure 9. Character prediction results from the visual module, linguistic module, and the final DADCM output (from left to right). Confidence scores are in the range [0, 1]. [E] denotes the end-of-sequence ("END") token. Red indicates incorrect predictions; blue indicates corrected and accurate characters.

3.4.2. Sensitivity to Confidence Threshold

To further investigate the effect of the language correction confidence threshold τ, we conducted a sensitivity analysis on the Union14M-Benchmark sub-datasets, as shown in Table 3. The results indicate that recognition accuracy remains stable across a wide range of τ values, with fluctuations within ±0.2% for most sub-datasets. This demonstrates that the dynamic vision–language fusion mechanism consistently enhances performance without relying on a specific threshold.

Notably, the Contextless subset exhibits slightly lower accuracy and minimal variation across different τ values, which is consistent with the observation in Table 2 that the language module may over-correct text lacking semantic content. This suggests that the model successfully avoids excessive correction on contextless text through the confidence-based selection.

When τ takes moderate values near 0.5, the resulting performance is similar to that of the model employing both visual and language branches. Incorporating the complete visual-linguistic fusion module provides an additional gain in accuracy. These findings indicate that the proposed dynamic fusion strategy effectively leverages complementary visual and linguistic information, improving recognition accuracy for semantically meaningful and visually challenging text while mitigating potential errors on contextless text.

Table 3. Recognition accuracy under different language correction thresholds τ on Union14M-Benchmark sub-datasets, reflecting the effect of dynamic vision–language fusion.

τ	Curve	Multi-Oriented	Artistic	Contextless	Salient	Multi-Words	General	Avg
0.1	89.5	85.2	79.7	85.1	79.7	86.0	79.7	84.0
0.2	89.6	85.3	79.8	85.2	79.8	86.1	79.8	84.2
0.3	89.7	85.4	79.9	85.2	79.9	86.2	79.9	84.3
0.4	89.7	85.5	79.9	85.3	79.9	86.2	79.9	84.4
0.5	89.8	85.5	80.0	85.3	80.0	86.3	80.0	84.5
0.6	89.8	85.5	80.0	85.3	80.0	86.3	80.0	84.5
0.7	89.7	85.4	79.9	85.2	79.9	86.2	79.9	84.4
0.8	89.6	85.3	79.8	85.2	79.8	86.1	79.8	84.3
0.9	89.5	85.2	79.7	85.1	79.7	86.0	79.7	84.2

3.5. Comparison with SOTA

We evaluate DADCM on six commonly used STR benchmarks and the seven subsets of the Union14M-L benchmark. The results are compared with three different types of STR methods. Table 3 shows the accuracy comparison across all 13 test datasets. All methods are trained under the same settings as our model.

As shown in Table 4, the comparison results with various STR methods suggest that attention-based approaches and models incorporating language information show better generalization in challenging scenarios. This demonstrates the effectiveness of attention mechanisms and language models for complex scene text recognition.

Table 4. Word accuracy comparison with other STR methods on 13 benchmark datasets.

IIIT	*SVT*	*IC13(857)*	*IC15 (1811)*	*SVTP*	*CUTE*		*Curved*	*Multi-Oriented*	*Artistic*	*Contextless*	*Salient*	*Multi-Word*	*General*		
Type	**Method**	**Common Benchmarks**					**Avg**	**Union14M**						**Avg**	**Avg**	**Param /($\times 10^6$)**
CTC	CRNN [21]	90.8	83.8	92.8	71.8	70.5	81.2	81.8	29.3	12.6	34.3	44.2	16.8	35.6	60.3	33.3	55.7	8.3
Attention	SATRN [43]	97.1	95.2	**98.8**	87.3	91.0	93.9	93.9	74.9	64.8	67.2	76.3	72.2	74.2	75.8	72.2	82.2	67.0
	MGP-STR [44]	97.2	**97.9**	98.0	91.4	93.0	98.1	95.9	85.9	80.8	73.6	76.1	78.4	72.8	84.4	78.9	86.7	148.0
	LISTER [45]	98.1	97.5	98.6	89.6	94.0	97.2	95.8	78.4	65.6	74.7	82.9	73.4	84.1	84.8	77.7	86.1	51.1
	MAERec [19]	**98.5**	97.8	98.3	89.5	94.4	**98.6**	96.2	88.8	83.9	**80.0**	85.5	**84.9**	**87.6**	85.9	85.2	90.3	35.7
LM	SRN [14]	95.5	89.3	95.6	79.2	83.9	91.5	89.2	49.7	20.1	50.7	61.1	43.9	51.6	62.7	48.5	67.3	51.7
	VisionLAN [24]	96.3	91.5	96.1	83.6	85.4	92.4	90.9	70.9	57.2	56.7	63.8	67.6	47.5	74.2	62.6	75.6	32.8
	ABINet++ [23]	97.2	95.7	97.9	87.6	92.2	94.5	94.2	75.1	61.5	65.3	71.2	72.9	59.2	79.4	69.2	80.7	36.7
	BUSNet [46]	97.6	97.5	97.9	89.3	**95.4**	97.8	95.9	82.7	79.1	71.8	79.2	77.4	72.9	83.9	78.1	86.3	32.1
	CdisNet [28]	98.0	97.1	97.9	88.7	93.6	97.2	95.4	81.4	73.9	73.6	79.1	78.5	81.4	82.4	78.6	86.4	65.5
	DADCM	98.3	**97.9**	98.7	**92.1**	95.3	**98.6**	**96.8**	**89.9**	**85.5**	79.8	**88.5**	81.3	86.5	**87.2**	**85.5**	**90.7**	31.8

The bolded values indicate the best performance within each dataset. The underlined values indicate the second-best performance. The italicized text in the first row of the table represents the specific sub-datasets corresponding to Common Benchmarks and Union14M, respectively.

In addition, existing STR methods have already achieved strong performance on the six common benchmarks. Therefore, improvements on these datasets are relatively small. Our method achieves a 0.4% average gain, with the highest improvement of 0.7% on the challenging IC15 dataset. On the Union14M-L benchmark, our method performs better, especially on subsets with irregular layouts, varying orientations, and limited semantic context. Specifically, accuracy improves by 0.7% on IC15, 1.1% on Curve, 1.6% on Multi-Oriented, and 3% on Contextless.

However, our method still faces limitations in scenarios involving multi-line or multi-word text. This may be attributed to the tendency of the offset predictions in the alignment module to align with dominant visual features. In addition, the model may focus more on character-level alignment within single words, sometimes treating multiple words as a single unit.

As shown in Figure 10, the prediction results of the DADCM model from the visual branch, the linguistic branch, and the visual–linguistic fusion branch are presented. By comparing the outputs of different modules, we observe that DADCM performs robustly and consistently in complex and challenging scene text recognition tasks.

For blurry, distorted, or irregularly shaped text images, the visual branch alone can produce accurate predictions. This suggests that the proposed visual modeling component has strong representational capacity in capturing visual cues. In cases where some characters are missing but the overall semantic content remains clear, the linguistic branch helps correct and complete the text based on context. For example, in the case of “Chang,” the language model successfully infers the correct result. Moreover, For contextless text images, such as proper names (“Evelyn”) or short words (“mei”), language-based correction can actually reduce recognition accuracy. In our experiments, these cases benefit from the hybrid model, which leverages visual alignment and attention mechanisms to achieve more accurate predictions than purely language-assisted methods.

However, some failure cases remain. When severe occlusion occurs, the model sometimes produces incorrect predictions, even though the output confidence scores are high across all modules. For instance, the prediction of "CAP" under occlusion could not be corrected. This indicates that the current model still has limited robustness in such conditions. Future work could focus on improving the model's ability to reason over occluded regions. Enhancing semantic modeling and context understanding in these cases may help further improve the generalization performance of the system.

Figure 10. Prediction results of DADCM from the visual, linguistic, and fusion modules. From left to right and bottom to top, the characters represent: ground-truth labels (GT), visual (VM), linguistic (LM), and fusion predictions (FM). Red: Incorrect characters, Blue: Corrected and accurate characters.

3.6. Analysis of Inference Speed

As shown in Table 5, we compare the average inference time and FLOPs of the proposed DADCM with several representative scene text recognition models on the seven subsets of Union14M. The proposed DADCM achieves an average inference time of 19.2 ms and requires 2.65 G FLOPs. Compared with MAERec, which requires 91.0 ms and 2.97 G FLOPs, as well as SRN, ABINet++, and BUSNet, DADCM provides faster or comparable inference while maintaining high recognition accuracy. Although slightly slower than CRNN, which runs at 6.5 milliseconds, and VisionLAN at 17.8 milliseconds, DADCM demonstrates superior robustness in handling complex scenes and challenging text samples. Overall, these results indicate that DADCM achieves a favorable balance between recognition accuracy and computational efficiency, making it well suited for practical applications.

Table 5. Average Inference Time and Computational Cost of STR Methods on the Union14M Benchmark.

Method	CRNN	MAERec	SRN	VisionLAN	ABINet++	BUSNet	DADCM
Time (ms)	6.5	91.0	19.1	17.8	29.6	19.8	19.2
FLOPs (G)	0.69	2.97	4.30	2.73	3.05	2.67	2.65

4. Discussion

This paper presents a scene text recognition method based on a dual-alignment correction mechanism with deformable alignment (DADCM). The proposed approach integrates SCRTNet, the DAM, and feature fusion to address challenges such as irregular layouts, occlusion, and lack of semantic information.

Extensive experiments across multiple benchmark datasets demonstrate the effectiveness of our method. In particular, ablation studies highlight the importance of establishing robust visual–language fusion rules in improving recognition accuracy, confirming and extending findings from previous STR research.

Limitations and Future Work: Although DADCM achieves competitive performance on benchmarks such as ICDAR2015, Total-Text, and CTW1500, challenges remain under general, lexicon-free settings. Recognition stability can fluctuate in images with high content

randomness, complex text layouts, or significant visual distractions. The model also shows limitations in feature preservation and sequence alignment for long character sequences.

Future research should focus on enhancing model generalization in complex environments, improving adaptability to multi-scale texts, and achieving a better trade-off between accuracy and inference efficiency. Exploring model compression or other optimization techniques may help improve inference speed while maintaining high accuracy.

5. Conclusions

In this work, we proposed DADCM, a scene text recognition method based on a dual-alignment correction mechanism with deformable alignment. By integrating SCRTNet, the DAM, and feature fusion, the method effectively addresses challenges such as irregular text layouts, occlusion, and limited semantic information. Extensive experiments on multiple benchmark datasets confirm the effectiveness of DADCM, and ablation studies highlight the importance of robust visual–language fusion for improving recognition accuracy. These findings provide insights for future development of more accurate and generalizable scene text recognition models.

Author Contributions: Conceptualization, Y.F. and C.L.; methodology, Y.F.; software, Y.F.; validation, Y.F. and C.L.; formal analysis, Y.F.; investigation, Y.F.; resources, Y.F.; data curation, Y.F.; writing—original draft preparation, Y.F.; writing—review and editing, C.L.; visualization, Y.F.; supervision, C.L.; project administration, Y.F.; funding acquisition, C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China, grant number 2023YFF0905603.

Data Availability Statement: The datasets used in this study, including IIIT5K, SVT, IC13, IC15, SVTP, CUTE , are publicly available at their respective repositories. The additional data generated during the experiments, such as model predictions and intermediate feature maps, are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wan, Z.; Zhang, J.; Zhang, L.; Luo, J.; Yao, C. On Vocabulary Reliance in Scene Text Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11425–11434.
2. Long, S.; He, X.; Yao, C. Scene Text Detection and Recognition: The Deep Learning Era. *Int. J. Comput. Vis.* **2021**, *129*, 1–24. [CrossRef]
3. Wang, Y.; Xie, H.; Zha, Z.; Tian, Y.; Fu, Z.; Zhang, Y. R-net: A Relationship Network for Efficient and Accurate Scene Text Detection. *IEEE Trans. Multimed.* **2021**, *23*, 1316–1329. [CrossRef]
4. 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.* **2021**, *3*, 297–308. [CrossRef]
5. Ouali, I.; Ben Halima, M.; Ali, W. Augmented Reality for Scene Text Recognition, Visualization and Reading to Assist Visually Impaired People. *Procedia Comput. Sci.* **2022**, *207*, 158–167. [CrossRef]
6. Mei, Q.; Hu, Q.; Yang, C.; Zheng, H.; Hu, Z. Port Recommendation System for Alternative Container Port Destinations Using a Novel Neural Language-Based Algorithm. *IEEE Access* **2020**, *8*, 199970–199979. [CrossRef]
7. Liu, W.; Chen, C.; Wong, K.-Y.K.; Su, Z.; Han, J. STAR-Net: A Spatial Attention Residue Network for Scene Text Recognition. In Proceedings of the British Machine Vision Conference, York, UK, 19–22 September 2016.
8. Shi, B.; Wang, X.; Lyu, P.; Yao, C.; Bai, X. Robust Scene Text Recognition with Automatic Rectification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4168–4176.
9. Zhan, F.; Lu, S. ESIR: End-to-End Scene Text Recognition via Iterative Image Rectification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2059–2068.
10. Luo, C.; Jin, L.; Sun, Z. MORAN: A Multi-Object Rectified Attention Network for Scene Text Recognition. *Pattern Recognit.* **2019**, *90*, 109–118. [CrossRef]

11. Li, H.; Wang, P.; Shen, C.; Zhang, G. Show, Attend and Read: A Simple and Strong Baseline for Irregular Text Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 8610–8617.
12. Yang, X.; He, D.; Zhou, Z.; Kifer, D.; Giles, C.L. Learning to Read Irregular Text with Attention Mechanisms. In Proceedings of the International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 3280–3286.
13. 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.* **2018**, *41*, 2035–2048. [CrossRef] [PubMed]
14. Yu, D.; Li, X.; Zhang, C.; Liu, T.; Han, J.; Liu, J.; Ding, E. Towards Accurate Scene Text Recognition with Semantic Reasoning Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12113–12122.
15. Zhao, Z.; Tang, J.; Lin, C.; Wu, B.; Huang, C.; Liu, H.; Tan, X.; Zhang, Z.; Xie, Y. Multi-Modal In-Context Learning Makes an Ego-Evolving Scene Text Recognizer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 15567–15576.
16. Yue, X.; Kuang, Z.; Lin, C.; Sun, H.; Zhang, W. RobustScanner: Dynamically Enhancing Positional Clues for Robust Text Recognition. In Proceedings of the European Conference on Computer Vision, Virtual, 23–28 August 2020; pp. 135–151.
17. Baek, J.; Kim, G.; Lee, J.; Park, S.; Han, D.; Yun, S.; Seong, J.O.; Lee, H. What Is Wrong with Scene Text Recognition Model Comparisons? Dataset and Model Analysis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4715–4723.
18. Bautista, D.; Atienza, R. Scene Text Recognition with Permuted Autoregressive Sequence Models. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 178–196.
19. Jiang, Q.; Wang, J.; Peng, D.; Liu, C.; Jin, L. Revisiting Scene Text Recognition: A Data Perspective. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 20543–20554.
20. Qiao, Z.; Zhou, Y.; Yang, D.; Zhou, Y.; Wang, W. SEED: Semantics Enhanced Encoder-Decoder Framework for Scene Text Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 13528–13537.
21. 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.* **2016**, *39*, 2298–2304. [CrossRef] [PubMed]
22. Fang, S.; Xie, H.; Wang, Y.; Mao, Z.; Zhang, Y. Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 7098–7107.
23. Fang, S.; Mao, Z.; Xie, H.; Wang, Y.; Yan, C.; Zhang, Y. ABINet++: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Spotting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 7123–7141. [CrossRef] [PubMed]
24. 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 Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 14194–14203.
25. Xiong, Y.; Li, Z.; Chen, Y.; Wang, F.; Zhu, X.; Luo, J.; Wang, W.; Lu, T.; Li, H.; Qiao, Y.; et al. Efficient Deformable ConvNets: Rethinking Dynamic and Sparse Operator for Vision Applications. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 5652–5661.
26. Wang, T.; Zhu, Y.; Jin, L.; Luo, C.; Chen, X.; Wu, Y.; Wang, Q.; Cai, M. Decoupled Attention Network for Text Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12216–12224.
27. Na, B.; Kim, Y.; Park, S. Multi-Modal Text Recognition Networks: Interactive Enhancements between Visual and Semantic Features. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 446–463.
28. Zheng, T.; Chen, Z.; Fang, S.; Xie, H.; Jiang, Y. CDistNet: Perceiving Multi-Domain Character Distance for Robust Text Recognition. *Int. J. Comput. Vis.* **2024**, *132*, 300–318. [CrossRef]
29. Hu, J.; Li, S.; Sun, G. Squeeze-and-Excitation Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
30. Jang, E.; Gu, S.; Poole, B. Categorical Reparameterization with Gumbel-Softmax. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; N-Gomez, A.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
32. Gupta, A.; Vedaldi, A.; Zisserman, A. Synthetic Data for Text Localisation in Natural Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2315–2324.
33. Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition. In Proceedings of the Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014.

34. Karatzas, D.; Shafait, F.; Uchida, S. ICDAR 2013 Robust Reading Competition. In Proceedings of the International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 1484–1493.
35. Wang, K.; Babenko, B.; Belongie, S. End-to-End Scene Text Recognition. In Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1457–1464.
36. Mishra, A.; Karteek, A.; Jawahar, C.V. Scene Text Recognition Using Higher Order Language Priors. In Proceedings of the British Machine Vision Conference, Swansea, UK, 7–10 September 2015; pp. 1–11.
37. Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V.; Lu, S.; et al. ICDAR 2015 Competition on Robust Reading. In Proceedings of the International Conference on Document Analysis and Recognition, Tunis, Tunisia, 23–26 August 2015; pp. 1156–1160.
38. Phan, T.-Q.; Shivakumara, P.; Tian, S.; Tan, C.-L. Recognizing Text with Perspective Distortion in Natural Scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–27 June 2013; pp. 569–576.
39. Anhar, R.; Palaiahnakote, S.; Chan, C.-S.; Tan, C.-L. A Robust Arbitrary Text Detection System for Natural Scene Images. *Expert Syst. Appl.* **2014**, *41*, 8027–8048. [CrossRef]
40. Loshchilov, I.; Hutter, F. Fixing Weight Decay Regularization in Adam. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
41. Yang, M.; Yang, B.; Liao, M.; Zhu, Y.; Bai, X. Class-Aware Mask-Guided Feature Refinement for Scene Text Recognition. *Pattern Recognit.* **2024**, *149*, 110244. [CrossRef]
42. Atienza, R. Vision Transformer for Fast and Efficient Scene Text Recognition. In Proceedings of the International Conference on Document Analysis and Recognition, Lausanne, Switzerland, 5–10 September 2021; pp. 319–334.
43. Lee, J.; Park, S.; Baek, J.; Seong, J.-O.; Kim, S.; Lee, H. On Recognizing Texts of Arbitrary Shapes with 2D Self-Attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 546–547.
44. Wang, P.; Da, C.; Yao, C. Multi-Granularity Prediction for Scene Text Recognition. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 1484–1493.
45. Cheng, C.; Wang, P.; Da, C.; Zheng, Q.; Yao, C. LISTER: Neighbor Decoding for Length-Insensitive Scene Text Recognition. In Proceedings of the International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 19484–19494.
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 Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; pp. 5885–5893.