OTE: Exploring Accurate Scene Text Recognition Using One Token

Jianjun Xu, Yuxin Wang*, Hongtao Xie, Yongdong Zhang
University of Science and Technology of China
xujj1998@mail.ustc.edu.cn, {wangyx58,htxie,zhyd73}@ustc.edu.cn

Abstract

In this paper, we propose a novel framework to fully exploit the potential of a single vector for scene text recognition (STR). Different from previous sequence-to-sequence methods that rely on a sequence of visual tokens to represent scene text images, we prove that just ***one token*** *is enough to characterize the entire text image and achieve accurate text recognition. Based on this insight, we introduce a new paradigm for STR, called* ***O****ne* ***T****oken r****E****cognizer (****OTE****). Specifically, we implement an image-to-vector encoder to extract the fine-grained global semantics, eliminating the need for sequential features. Furthermore, an elegant yet potent vector-to-sequence decoder is designed to adaptively diffuse global semantics to corresponding character locations, enabling both autoregressive and non-autoregressive decoding schemes. By executing decoding within a high-level representational space, our vector-to-sequence (V2S) approach avoids the alignment issues between visual tokens and character embeddings prevalent in traditional sequence-to-sequence methods. Remarkably, due to introducing character-wise fine-grained information, such global tokens also boost the performance of scene text retrieval tasks. Extensive experiments on synthetic and real datasets demonstrate the effectiveness of our method by achieving new state-of-the-art results on various public STR benchmarks. Our code is available at https://github.com/Xu-Jianjun/OTE.*

1. Introduction

Scene Text Recognition (STR) is an important task in computer vision that aims to read the text content of a given cropped scene image. Due to the provision of rich semantic information, it is widely applied in fields such as autonomous driving, visual question answering, and augmented reality.

Most deep learning methods [5, 8, 9, 12, 44, 47] regard scene text recognition as a sequence labeling or sequence transcription problem. Such sequence-based pipelines generally employ image-to-sequence encoders for extracting visual feature sequences, which are subsequently decoded into text sequences using sequence-to-sequence (S2S) decoders, typically implemented in two forms: attention-based [5, 9, 37, 41] and CTC-based [8, 31], as shown in Fig. 1. Specifically, attention-based decoders utilize the cross-attention mechanism to intricately focus on different visual feature sequence segments while predicting different characters. For example, in recognizing the word "Kevin", the decoder sequentially identifies each character—'K', 'e', 'v', 'i', 'n'—by actively querying specific visual features. However, these methods are critically contingent upon the accuracy of the attention map, and attention drift can drastically undermine the performance, leading to significant accuracy losses. In contrast, CTC-based methods typically allocate a unique output token for each anticipated character, integrating a special 'blank' label to address alignment issues. However, this approach frequently necessitates substantial post-processing to manage blank and repeated char-

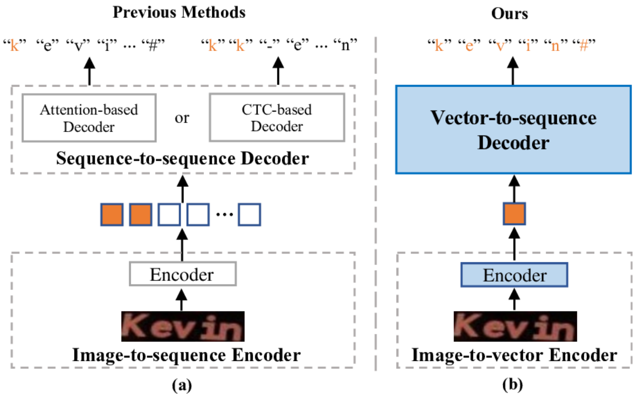


Figure 1. Comparison between previous methods and our OTE. (a) The previous methods typically use an image-to-sequence encoder to extract visual features and a sequence-to-sequence decoder to align character embeddings and visual tokens. (b) Our method utilizes an image-to-vector encoder to extract the multi-granularity global semantic and employs a vector-to-sequence decoder to predict the sequence by reusing this global semantic, avoiding the process of aligning character embeddings with partial visual tokens.

*Corresponding author

acters, proving challenging to handle in complex scenarios. Thus, **is there another paradigm to represent text images efficiently and decode text sequences accurately while avoiding alignment issues between visual features and character embeddings?**

Research in general image understanding has demonstrated that Vision Transformer [7] (ViT) architectures can distill complex and fine-grained semantic features into a single token, achieving remarkable results. For instance, plain ViT [7] employs a single *CLS* token to classify over 20,000 categories accurately. Similarly, CLIP [29] has proven that using an additional token can effectively distinguish a vast array of images, aligning them with textual descriptions. Inspired by this, we explored the application of this 'single-token representation for fine-grained semantic perception' approach to text recognition. Due to the uniqueness of cropped text images, two primary challenges emerge: (i) **Sequence Prediction**. Unlike image recognition, which predicts independent category labels, text recognition is a sequential prediction task, requiring not only the prediction of existing characters but also an understanding of their sequence and combination. (ii) **Application of Linguistic Rules.** As scene text carries rich linguistic information, judicious utilization of language rules can significantly boost recognition performance.

Driven by this analysis, we present a simple, effective, and adaptable **O**ne **T**oken r**E**cognizer (OTE) for Scene Text Recognition, as illustrated in Fig. 1. The OTE comprises a ViT-based image-to-vector (I2V) encoder designed to extract global semantic features and a vector-to-sequence (V2S) decoder for transcribing these global semantic features into character sequences.

Firstly, the image-to-vector (I2V) encoder capitalizes on the Vision Transformer's long-range perceptual abilities and detailed representation capabilities to generate a rich semantic vector. According to Information Bottleneck Theory [34], such an encoding strategy improves the effectiveness in extracting and compressing essential semantic features from images while filtering noise or irrelevant details. Our experiments also demonstrate that a single token can encode a comprehensive semantic representation, applicable across a spectrum of ViT variants. Secondly, based on the image-to-vector encoder, we have designed a novel vector-to-sequence (V2S) paradigm to decode character-wise predictions from the global token. Distinct from conventional methods that analyze features within a 2-D spatial framework, our V2S strategy reuses global semantics and decodes character information across the channel dimension. Furthermore, we have introduced sequence language modeling into this framework, implementing both autoregressive and non-autoregressive decoding strategies via a masked multi-head self-attention mechanism. By executing decoding within a high-level representational space, V2S demonstrates considerable robustness in character-wise representation, marking a significant improvement over traditional sequence-to-sequence (S2S) paradigms. The efficiency of our method is further enhanced by its streamlined post-processing.

In addition, we explore the potential of our method for scene text retrieval tasks. By introducing the character-wise fine-grained information, the multi-grained global semantics also help the retrieval task to obtain a strong representation of input images.

Our main contributions can be summarized as follows:

- By capturing global multi-grained semantics, we **first** prove that **One token** is enough for accurate scene text recognition.
- A new solution is introduced to eliminate the need for sequential tokens in scene text recognition. Furthermore, a concise vector-to-sequence decoder is designed to be capable of decoding text sequences, whether in an auto-regressive or non-auto-regressive manner.
- We prove that character-wise fine-grained semantics will benefit the alignment process in scene text retrieval. The OTE provides a unified solution for both scene text recognition and retrieval tasks, utilizing the global token.
- Experimental results across various training datasets and diverse testing benchmarks verify the effectiveness of our framework, which achieves state-of-the-art performance.

2. Related Work

2.1. Scene Text Recognition

Scene text recognition typically involves extracting visual features using a backbone and aligning them with textual representations via a sequence-to-sequence (S2S) decoder. There are generally two implementation forms for S2S: CTC (Connectionist Temporal Classification) and attention. CTC-based decoder [8, 31] aims to maximize the probability of all paths for the final prediction, employing blank labels and post-processing to address alignment issues, while attention-based methods [9, 37, 41, 43] localize the position of each character using a cross-attention mechanism. Most attention-based methods employ a position-attention mechanism to query corresponding visual features based on position. Building on this, many recent methods [5, 41] integrate language modeling concepts into the character decoding process, yielding impressive results. Unlike these methods, our approach extracts high-level global semantics and decodes character sequences via a vector-to-sequence decoder. Our method effectively circumvents the reliance on sequential features and the need to align character representations with low-level visual features.

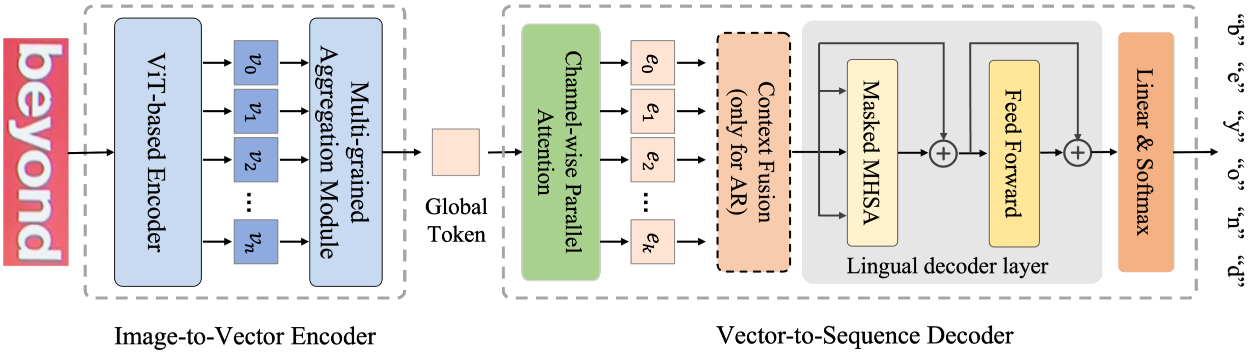


Figure 2. The pipeline of OTE. The image-to-vector encoder aims to aggregate both fine-grained and global semantics of the input image into the global token. Then, a concise vector-to-sequence decoder is designed to be capable of decoding text sequences, whether in an auto-regressive or non-auto-regressive manner.

2.2. Scene Text Retrieval

Scene text retrieval is another important task for understanding scene text images [1, 2, 10, 38]. Mishra *et al.* [24] first adopts two independent steps for character-wise detection and classification. Then, the probability of query texts is used for retrieval prediction. To provide an end-to-end scene text retrieval model, Gomez *et al.* [10] predicts the text proposals and PHOCs simultaneously, and the images are ranked by calculating the distance between the query word and text proposal. Based on [33], Wang *et al.* [38] introduce a well-designed alignment loss function to enhance the retrieval capability. In this paper, we argue that not only global semantics but also character-wise fine-grained information is important for retrieval tasks. Furthermore, our consistent representation first provides a unified solution for both scene text recognition and retrieval tasks.

3. Our Method

3.1. Pipeline

The pipeline of OTE is shown in Fig. 2, which contains a ViT-based image-to-vector encoder and a vector-to-sequence decoder. In the ViT-based image-to-vector encoder, we use various vision transformers as our backbone and construct a multi-grained aggregation module to generate a token containing both character-wise fine-grained and image-level global semantics. After the multi-grained token generation, an effective and flexible vector-to-sequence decoder is crafted to decode specific text sequences from the global semantic vector. Significantly, through the optional use of context fusion operations and attention mask mechanisms, our V2S decoder adeptly handles both autoregressive and non-autoregressive decoding.

3.2. Image-to-vector Encoder

The motivation of the ViT-based image-to-vector encoder is to gather the fine-grained and global semantics into a single token, which is realized based on the self-attention mechanism (ViT) and a multi-grained aggregation module (MAM).

As shown in Fig. 2, we first obtain the multi-grained semantics by calculating both local and long-range dependency through a ViT-based encoder. We choose different scales (ViT-S/ViT-B) and different architectures (plain ViT/multi-scale SVTR) vision transformers as our backbone. Then, a multi-grained aggregation module (MAM) is used for multi-grained token generation. In our experiment, we demonstrate that simple aggregation schemes (*e.g.* global average pooling or CLS token) is sufficient is sufficient to extract robust multi-grained token. The image-to-vector encoder is highly adaptable, easily conforming to various model scales and structural designs.

Overall, the I2V process is shown as the following:

$$\mathbf{z} = \mathbf{MAM}(\mathrm{Enc}(\mathbf{x})); \quad \mathbf{z} \in \mathbb{R}^{1\times d} \tag{1}$$

where x is the input images, z is the output global token, and d is the dimensional of the model.

3.3. Vector-to-sequence Decoder

The vector-to-sequence decoder comprises two common components: channel-wise parallel attention and a lingual decoder layer, along with a context fusion module specific to autoregressive (AR) decoding.

Channel-wise parallel attention: Given that the global semantic token encapsulates all information, intuitively, spreading the entire semantic information to specific positions can achieve sequence decoding. Inspired by SENet [13], we propose channel-wise parallel attention (CPA), as depicted in Fig. 3. The CPA constructs a parallel generation process of channel attention maps to improve efficiency. Initially, the global token z undergoes a transformation via a linear projection layer ($\mathbf{W}_1$), subsequently combined with positional embeddings to integrate ordering information. The generation of the attention matrix is orchestrated through the employment of a hyperbolic tangent ac-

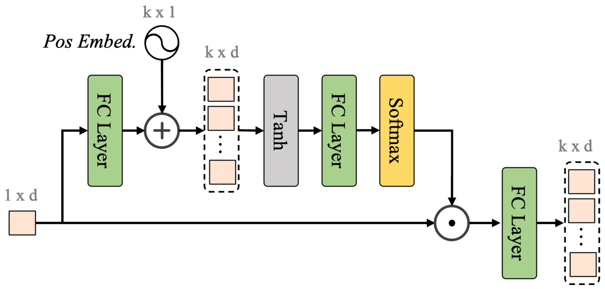


Figure 3. The structure of Channel-wise Parallel Attention(CPA).

tivation function (σ), succeeded by a linear transformation ($\mathbf{W}_2$) and normalization via a Softmax operation.

$$attn = \text{Softmax}\left(\sigma\left(z * \mathbf{W}_1 + \mathbf{P}\right) * \mathbf{W}_2\right), \tag{2}$$

where $\mathbf{P} \in \mathbb{R}^{k \times d}$ and $\mathbf{W_1}, \mathbf{W_1} \in \mathbb{R}^{d \times d}$ represent the positional embedding and the weights of the linear layer, respectively. k is the max length of predicted text.

The attention map is first generated using a fully connected layer, a Tanh activation layer (σ), and a Softmax layer. Subsequently, the attention map is element-wise multiplied with the global token and then passed through another fully connected layer ($\mathbf{W}_3$) to obtain the distinctive features f_o.

$$f_o = (\mathcal{T}(z) \cdot attn) * \mathbf{W}_3 \tag{3}$$

where $\mathcal{T}$ signifies the *tile* operation.

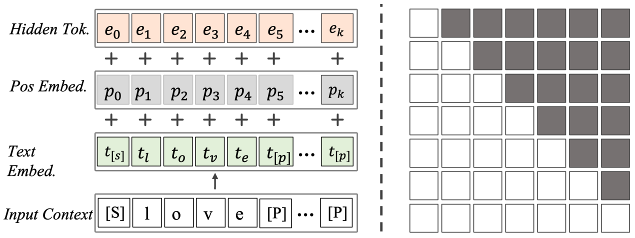
Context fusion: For autoregressive (AR) decoding, we incorporate a straightforward context fusion strategy to effectively integrate the context information of already predicted characters, as shown in Fig. 4(a). Specifically, we merge the embeddings outputted by the CPA with the context embeddings and position embedding, thus obtaining the fused features. Notably, during training, we utilize the right-shifted ground truth as the context embedding, while in the testing phase, we employ the characters already predicted.

Lingual decoder layer: As illustrated in Figure Fig. 2, the lingual decoder consists of two components: the masked multi-head self-attention (MHSA) module and the feed-forward module. In the following equations, LayerNorm and Dropout are omitted for brevity. The masked MHSA module captures the semantic dependencies between characters, taking the feature embedding f_o as input.

$$[\mathbf{Q}, \mathbf{K}, \mathbf{V}] = f_o * \mathbf{W}_l, \tag{4}$$

$$f_{\text{mha}} = f_o + \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}} + \mathbf{M}\right)\mathbf{V} \tag{5}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are obtained from the feature embedding f_o, and $\mathbf{W}_l \in \mathbb{R}^{d \times 3 * d}$ is a learnable mapping matrix. The attention mask ($\mathbf{M} \in \mathbb{R}^{k \times k}$) controls the flow of information at specific positions, thereby facilitating language modeling. Specifically, a causal self-attention mask is employed for autoregressive decoding, ensuring that future tokens are conditioned on past tokens, as shown in Fig. 4(b). For non-autoregressive decoding, no attention mask is utilized. The output state is the output of the Feed-Forward Network (FFN).


Hidden Tok.
e_0 e_1 e_2 e_3 e_4 e_5 ··· e_k
Pos Embed.
p_0 p_1 p_2 p_3 p_4 p_5 ··· p_k
Text Embed.
$t_{[s]}$ t_l t_o t_v t_e $t_{[p]}$ ··· $t_{[p]}$
Input Context
[S] l o v e [P] ··· [P]

(a) Context Fusion (b) Attention mask for AR

Figure 4. (a) The context fusion process for autoregressive decoding, which merges the hidden features outputted by the CPA with the context embeddings and position embedding. (b) The causal self-attention mask is used for autoregressive decoding.

$$f_{\text{output}} = f_{\text{mha}} + \text{FFN}(f_{\text{mha}}) \tag{6}$$

Finally, the output logits $p = Linear(f_{\text{output}}) \in \mathbb{R}^{k \times S}$, where S is the size of the character set.

3.4. Training Objective

The training objective of scene text recognition is formulated in Eq. (7). Generally, we use the cross-entropy loss for character learning. p_t and g_t are prediction and ground truth at time step t.

$$L_{rec} = -\frac{1}{N}\sum_{t=1}^{N}\log(p_t|g_t) \tag{7}$$

3.5. Global semantic token for Scene Text Retrieval

The scene text retrieval task aims to search the text instances from an image gallery, which can be regarded as an alignment process between texts and instance-level images. The OTE, by aggregating the visual representation into a single token, is inherently suited for retrieval tasks, offering a unified feature representation for both scene text recognition and scene text retrieval tasks. Specifically, we employ a scene text detector (same as [38]) without fine-tuning to crop text patches and use the frozen image-to-vector encoder for extracting global semantic tokens as visual representations. These tokens are mapped to a visual-text joint space via a linear layer, facilitating matching with text queries, like CLIP. Besides, the structure of the text encoder is the same as that of CLIP's text encoder, while the input consists of split character sequences rather than entire words. To align visual and textual representations accurately, we use the contrastive loss [29] in the training stage. In the inference, we assign the image to the most similar word query for retrieval prediction.

$$L_{ret} = \text{CLIP}(Token, word\_embedding) \quad (8)$$

4. Experiment

In this section, we first introduce the experimental setup, including the datasets, implementation details, and evaluation metrics. Next, we discuss the impact of different components and settings through ablation studies. Finally, we present our results and compare OTE to SOTA methods.

4.1. Datasets

To comprehensively evaluate the performance of our method, we trained our model on a wide range of datasets, including synthetic and real-world datasets, and conducted tests across multiple benchmarks. To explore OTE's potential in other domains, we also performed training and testing on text-image retrieval datasets.

Trained on synth dataset: Following [5, 9], we use two synthetic datasets (MJ [14] and ST [11]) for training and evaluate our method on six standard datasets (IIIT [23], SVT [39], IC13 [16], IC15 [17], SVTP [27], CUTE [30]). Moreover, we also introduce three additional challenging datasets for further evaluation, including ArT [6], COCO-Text (COCO) [36], Uber-Text (Uber) [49].

Trained on real-world dataset: For real-world data, we select the Union14M-L [15] dataset for our experiments, which comprises more than four million labeled images from a wide array of real-life scenarios. Specifically, addressing the current challenges in Scene Text Recognition (STR), Union14M-L includes an extensive challenge-driven Benchmark. This benchmark consists of six subsets, totaling 409,393 images, characterized by both complexity and diversity. These subsets include curve text, multi-oriented text, artistic text, contextless text, salient text, and multi-word text.

Retrieval dataset: To evaluate the effectiveness of our method in scene text retrieval, IIITSTR [24] is used for evaluation. IIITSTR [24] consists of 10k images and 50 query words. Due to the various styles, fonts, and viewpoints, it is a challenging dataset and can effectively reflect the retrieval performance.

4.2. Implementation Details

We construct the plain ViT [7] and SVTR [8] as our backbone. Images are resized to 32 × 128. Following [5], the RandAugment is utilized for data augmentation, including Sharpness, Invert, GaussianBlur, and PoissonNoise. We choose the AdamW as the optimizer and set the learning rate to 3e-4. The cosine learning rate decay is used to degrade the learning rate. The experiments are conducted on 2 NVIDIA 4090 GPUs with batch size 512 per GPU for 20 epochs. The max length N is set to 25. For retrieval models, we resize the input image to 32×128 and use MJ [14] and MLT-5k [26] to train the model, which follows the setup of [38]. The experiments are conducted on 4 NVIDIA 4090 GPUs with batch size 384 per GPU for 20 epochs. Specially, we directly use the scene text detector provided by Wang *et al.* [38] without fine-tuning to crop the text patches.

4.3. Evaluation Metric

We set the size of the recognition character to 36, including a-z and 0-9. Word accuracy is the metric for STR benchmarks. A prediction is considered correct if and only if characters at all positions match.

4.4. Ablation study

The Evaluation of Multi-grained Aggregation Module. Leveraging the multi-grained capabilities of the ViT-based backbone [7], we demonstrate that a simple Multi-grained Aggregation Module (MAM) is effectively adequate for robust multi-grained token generation. We explored two distinct methods to generate the multi-grained token: using a class token and implementing global average pooling (GAP). In the first implementation, an additional class token from ViT [7] represents the global token. In the second, we utilize only visual tokens as input for ViT, with a GAP layer aggregating these tokens into a global token. Consistently, we employ vit-s as the backbone and utilize autoregressive decoding in our baseline models for comparative analysis. These models are trained on synthetic datasets (MJ and ST). As shown in Tab. 2, these two implementations achieve similar performance, with the additional class token slightly outperforming pooling all output features. Therefore, unless otherwise specified, we default to adding an extra token to aggregate multi-grained information. Specifically, for customized ViT variants like SVTR [8], which are challenging to augment with additional tokens, we use global average pooling on the output features to obtain the global token.

The impact of different backbones. Our One Token Recognizer (OTE) is highly compatible with most ViT-based backbones. Intuitively, the stronger a backbone's ability to capture multi-grained semantic information, the more information the global token contains, leading to higher recognition accuracy. To this end, we conducted experiments using backbones of different scales (ViT-S vs. ViT-B) and structures (ViT-S vs. SVTR), with results illustrated in Tab. 1. Two conclusions can be drawn: (1) Larger-scale backbones generally yield better performance. (2) Models designed explicitly for text images tend to perform better. Regarding the first point, larger models typically imply more robust representational capabilities, a hypothesis our experiments corroborate. OTE with ViT-B significantly leads OTE with ViT-S in most datasets, with an average accuracy improvement of 0.7%. For the second point, even with equal theo-

Table 1. Comparison with SOTA models trained on synthetic datasets (**MJ** and **ST**) on six common STR benchmarks. **N** and **A** respectively represent non-autoregressive and autoregressive decoding. **Bold** and underlined values denote the 1st and 2nd results in each column.

Type	Methods	Lang.	Regular Text			Irregular Text			Avg	Params(M)
			IIIT	SVT	IC13	IC15	SVTP	CUTE		
CTC	CRNN [31]	×	82.9	81.6	91.9	69.4	70.0	65.5	78.6	8.3
	SVTR [8]	×	96.0	91.5	97.1	85.2	89.9	91.7	92.3	24.6
Attention	TRBA [4]	×	87.9	87.5	93.6	77.6	79.2	74.0	84.6	-
	DAN [42]	×	94.3	89.2	93.9	74.5	80.0	84.4	87.2	-
	RobustScanner [46]	×	95.3	88.1	94.8	77.1	79.5	90.3	88.4	-
	TextScanner [37]	×	93.9	90.1	92.9	79.4	84.3	83.3	88.5	-
	ViTSTR [3]	×	88.4	87.7	93.2	78.5	81.8	81.3	85.6	-
	ABINet-Vision [9]	×	94.6	94.9	90.4	81.7	84.2	86.5	89.8	23.5
	Parseq_N [5]	×	95.7	92.6	96.3	85.1	87.9	91.4	92.0	23.8
LM	SEED [28]	✓	93.8	89.6	92.8	80.0	81.4	83.6	88.3	-
	SRN [45]	✓	94.8	95.5	91.5	82.7	85.1	87.8	90.4	55
	VisionLAN [43]	✓	95.8	91.7	95.7	83.7	86.0	88.5	91.2	32.8
	ABINet [9]	✓	96.2	93.5	97.4	86.0	89.3	89.2	92.6	36.7
	Parseq_A [5]	✓	**97.0**	93.6	97.0	86.5	88.9	92.2	93.3	23.8
	ConCLR [48]	✓	96.5	94.3	97.7	85.4	89.3	91.3	92.8	37.0
	MGP [40]	✓	95.3	93.5	96.4	86.1	87.3	87.9	92.0	52.6
Ours	OTE_N / ViT-S	×	95.8	94.6	96.5	85.2	88.2	89.0	92.2	24.0
	OTE_A / ViT-S	✓	96.2	93.5	97.6	85.9	89.6	91.7	92.8	24.0
	OTE_N / ViT-B	×	95.7	95.4	97.0	85.4	89.3	90.3	92.5	94.2
	OTE_A / ViT-B	✓	96.4	**95.5**	**97.9**	86.8	**91.9**	90.3	**93.5**	94.2
	OTE_N / SVTR	×	95.9	94.4	97.8	86.0	88.5	90.3	92.6	25.2
	OTE_A / SVTR	✓	96.4	**95.5**	97.4	**87.2**	89.6	**92.4**	93.4	25.2

Table 2. The Evaluation of Multi-grained Aggregation Module. ViT-small and auto-regressive decoding are used in this experiment. GAP means global average pooling, and CLS means using the class token.

Strategy	Regular Text			Irregular Text			Avg
	IIIT	SVT	IC13	IC15	SVTP	CUTE	
CLS	**96.2**	93.5	**97.6**	**85.9**	**89.6**	**91.7**	**92.8**
GAP	96.0	**94.7**	96.5	85.5	88.5	91.2	92.5

retical information capacity (the dimension of the global token being 384 in both ViT-S and SVTR), OTE with SVTR outperforms OTE with ViT-S by 0.6%. We attribute this to SVTR's stronger representational ability, compared to ViT, to focus on stroke features and capture local features within individual characters and long-distance global dependencies between characters. Without introducing extra parameters, a stronger backbone can generate more potent multi-grained semantics, affirming the generalization of our multi-grained aggregation module and vector-to-sequence paradigm.

4.5. Comparisons with State-of-the-Arts

To comprehensively validate the performance of our model, we train our series of models on both synthetic [11, 14] and real-world datasets [15]. Additionally, we compare them against current state-of-the-art (SOTA) models across various benchmarks.

4.5.1 Evaluation on synthetic dataset

We classify existing methods into three categories: CTC-based, attention-based, and language-aware, and evaluated our models on six benchmarks, as shown in Tab. 1. By using a plain ViT-S as the backbone, OTE achieves the second-best accuracy among language-free models using only 24M parameters, trailing only behind SVTR [8] and surpassing ABINet-Vision [9] (by a 3.4% boost) and Parseq [5]$_N$ (by a 0.2% boost). By increasing parameters and using ViT-B as the backbone, our model further enhances its performance and reaches an average accuracy of 92.5%, outperforming all other language-free models. Switching to the hierarchically structured vision transformer model SVTR [8] as the backbone, our model's performance jumps to 92.6%.

Compared to language-aware models, our OTE demonstrates significant advantages. With only 24M parameters, OTE/ViT-S achieves a 92.8% accuracy in autoregressive decoding. When using a larger (OTE/ViT-B) or stronger (OTE_A/SVTR) version, our methods set new state-of-the-art records, achieving 93.5% and 93.4%, respectively. Notably, our approach accomplishes this without needing

Table 3. Performance of models trained on the training set of **Union14M-L**. A and N represent the use of autoregressive and non-autoregressive as the backbone, respectively. PT denotes pre-training. **Bold** and underlined values denote the 1st and 2nd results in each column.For a fair comparison, following [15], IC13, and IC15 are larger versions, with Avg representing the average of all benchmarks.

Type	Method	Common Benchmarks							Union14M-Benchmark								
		IIIT 3000	IC13 1015	SVT 647	IC15 2077	SVTP 645	CUTE 288	Avg	Curve	Multi-Oriented	Artistic	Contextless	Salient	Multi-Words	General	Avg	Paramter(M)
CTC	CRNN [31]	90.8	91.8	83.8	71.8	70.4	80.9	81.6	19.4	4.5	34.2	44.0	16.7	35.7	60.4	30.7	8.3
	SVTR [8]	95.9	95.5	92.4	83.9	85.7	93.1	91.1	72.4	68.2	54.1	68.0	71.4	67.7	77.0	68.4	24.6
Attention	MORAN [21]	94.7	94.3	89.0	78.8	83.4	87.2	87.9	43.8	12.8	47.3	55.1	45.7	54.6	44.7	43.4	-
	ASTER [32]	94.3	92.6	88.9	77.7	80.5	86.5	86.7	38.4	13.0	41.8	52.9	31.9	49.8	66.7	42.1	-
	DAN [42]	95.5	95.2	88.6	78.3	79.9	86.1	87.3	46.0	22.8	49.3	61.6	44.6	61.2	67.0	50.4	-
	SATRN [18]	97.0	97.9	95.2	87.1	91.0	96.2	93.9	74.8	64.7	67.1	76.1	72.2	74.1	75.8	72.1	-
	RobustScanner [46]	96.8	95.7	92.4	86.4	83.9	93.8	91.2	66.2	54.2	61.4	72.7	60.1	74.2	75.7	66.4	-
LM	SRN [45]	95.5	94.7	89.5	79.1	83.9	91.3	89.0	49.7	20.0	50.7	61.0	43.9	51.5	62.7	48.5	55
	ABINet [9]	97.2	97.2	95.7	87.6	92.1	94.4	94.0	75.0	61.5	65.3	71.1	72.9	59.1	79.4	69.2	36.7
	VisionLAN [43]	96.3	95.1	91.3	83.6	85.4	92.4	91.3	70.7	57.2	56.7	63.8	67.6	47.3	74.2	62.5	32.8
	MATRN [25]	98.2	97.9	96.9	88.2	94.1	**97.9**	95.5	80.5	64.7	71.1	74.8	79.4	67.6	77.9	74.6	44.2
	MAERec-S[15]	86.8	96.9	93.7	84.9	89.6	93.8	91.0	73.7	64.4	62.1	71.5	69.5	49.3	78.7	67.0	35.8
	MAERec-S[15] with PT	98.0	97.6	96.8	87.1	93.2	**97.9**	95.1	81.4	71.4	72.0	**82.0**	78.5	**82.4**	82.5	**78.6**	35.8
Ours	OTE$_N$ / ViT-S	97.1	97.4	96.8	86.5	92.6	94.1	94.0	78.0	74.6	66.4	68.0	73.7	59.5	79.6	71.4	24.0
	OTE$_A$ / ViT-S	98.1	97.5	96.9	88.2	93.6	96.5	95.1	84.0	81.5	71.5	73.6	79.2	64.0	81.8	76.5	24.0
	OTE$_N$ / ViT-B	97.6	97.1	96.1	86.1	92.6	95.5	94.1	77.8	78.5	65.4	65.1	74.4	53.2	79.9	70.6	94.2
	OTE$_A$ / ViT-B	**98.4**	97.6	96.8	88.2	93.8	96.2	95.2	**85.6**	**88.4**	71.5	73.4	**81.8**	65.9	**82.9**	78.5	94.2
	OTE$_N$ / SVTR	98.1	97.5	96.6	86.7	91.2	96.2	93.4	79.2	76.0	70.0	74.3	76.0	64.2	80.1	74.3	25.2
	OTE$_A$ / SVTR	98.1	**98.0**	**98.0**	**89.1**	**95.5**	97.6	**96.1**	83.1	82.8	**73.5**	73.7	79.7	70.3	82.2	77.9	25.2

Table 4. Comparison with SOTA methods on challenging datasets. We chose OTE / SVTR for evaluation.

Method	Lang.	ArT	COCO	Uber
CRNN [31]	×	57.3	49.3	33.1
ViTSTR [3]	×	66.1	56.4	37.6
TRBA [4]	×	68.2	61.4	38.0
ABINet [9]	✓	65.4	57.1	34.9
PARSeq$_A$ [5]	✓	**70.7**	64.0	42.0
OTE$_N$ / SVTR	×	67.2	62.9	45.9
OTE$_A$ / SVTR	✓	69.1	**64.5**	**47.8**

Table 5. The comparison between OTE and recent methods in scene text retrieval. In particular, mAP is used to evaluate retrieval accuracy.

Method	mAP
Mishra *et al.* [24]	42.70
He *et al.* [12] (dictionary)	66.95
He *et al.* [12] (PHOC)	46.34
Gomez *et al.* [10]	69.83
Mafla *et al.* [22]	71.67
ABCNet [20]	67.25
Mask TextSpotter v3 [19]	74.48
Wang *et al.* [38]	77.09
OTE	**80.90**

complex, manually defined language modeling and post-processing, which is common in other methods. Across every sub-category of datasets, our model achieves the best (SVT [39], IC13 [16], IC15[17], SVTP [27], CUTE [30]) or second-best (IIIT [23]) performance.

In Tab. 4, to assess our model's performance on a broader challenging dataset, we further evaluate our method on three additional challenging datasets: ArT [6], COCO [36], and Uber [49]. The results indicate that our model excels on these challenging benchmarks, particularly standing out on COCO [36] and Uber [49].

4.5.2 Evaluation on real-world dataset

We further conduct experiments on real-world datasets [15], demonstrating our model's robustness with results shown in Tab. 3. On six common benchmarks, our model exhibit similar trends to those on synthetic datasets and achieved state-of-the-art (SOTA) performance, demonstrating our approach's effectiveness and stability. To comprehensively analyze our method's performance, we evaluate six challenging datasets from [15]: curve text, multi-oriented text, artistic text, contextless text, salient text, and multi-word text. Notably, our model's performance varies across these types. Specifically, OTE excels in recognizing curve text, multi-oriented text, and salient text, significantly outperforming current SOTA methods, including the large-scale pre-trained MAERec, and slightly leads in artistic text and contextless text datasets. We attribute this to most existing methods utilizing cross-attention to query specific visual features through positional information, which can lead to attention drift in diverse text shapes, making it challenging to locate corresponding visual features accurately. In contrast, our paradigm of directly extracting fine-grained global information adapts better to various shapes, positions, and styles of fonts. We further visualize qualitative recognition results on six challenging benchmarks, as

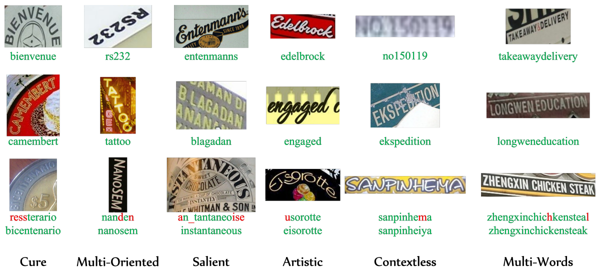


Figure 5. Qualitative Recognition Results on six challenging benchmarks, with red indicating errors predicted characters.

shown in Fig. 5. However, our performance on multi-word text needs to catch up to SOTA, which can be attributed to the backbone's limited capability to capture fine-grained features. As text length increases, the backbone's difficulty in compressing and summarizing the input image also rises, as evidenced by OTE/SVTR outperforming OTE/ViT-S and OTE/ViT-B in this aspect. Moreover, decoding all characters from the global token becomes more challenging with longer texts. Nonetheless, this issue is inherent to the STR task, and our model still surpasses most language model-based methods for long-length text.

4.6. Evaluation on Scene Text Retrieval

Due to introducing the character-wise fine-grained semantics into global tokens, we further evaluate the effectiveness of OTE on the retrieval dataset to show our significance. To conduct a fair comparison with previous methods, we directly use the scene text detector provided by Wang *et al.* [38] without fine-tuning to crop the text patches. As shown in Tab. 5, our method obtains a new state-of-the-art result (80.9 mAP) compared with all the existing retrieval methods. Compared with the best method Wang *et al.* [38], our OTE obtains a 3.81% improvement in mAP.

5. Discussion

5.1. Efficacy

The efficacy of our method can be attributed to two fundamental aspects: 1) **Robust Global Token Extraction**, the OTE uses only one token to capture the most crucial and distinctive features while eliminating unnecessary noise and redundancy. We employed t-SNE to visualize global tokens from the 20 most common words across six standard benchmarks, as shown in Fig. 6. The results revealed these tokens effectively capture global semantics, showing remarkable resilience to text image quality and style variations. The impressive performance in retrieval tasks also indicates that such features align well with textual labels. 2) **Vector-to-Sequence Decoding Paradigm.** Different from traditional sequence-to-sequence (S2S) attention-based methods, which often grapple with challenges like attention drift or missing in aligning low-level visual tokens for character embedding, our V2S approach capitalizes on the characteristic of reusability. It is akin to easily deconstructing the corresponding character sequence when the word is known.

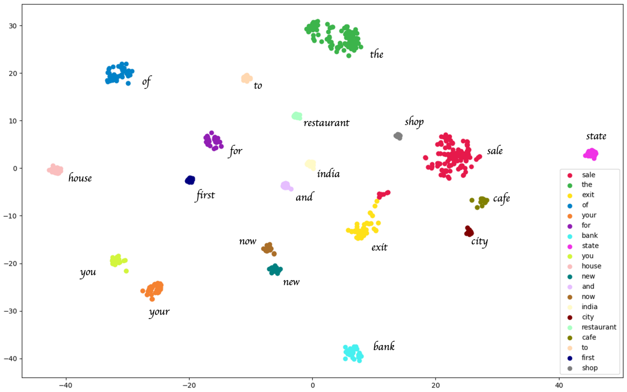


Figure 6. The t-SNE [35] visualization of global tokens for the most common 20 words in the 6 common benchmarks

5.2. Limitation

OTE encounters limitations in processing words of extremely long lengths. This challenge can be attributed to the encoder's difficulty in efficiently compressing features and capturing fine-grained global semantics. This issue is not unique to OTE but is a common challenge faced by current STR algorithms, particularly language-aware methods. In our experiments, scaling up the encoder (from ViT-S to ViT-B) or employing a more robust backbone (switching from ViT-S to SVTR) can alleviate this problem.

6. Conclusion

In this paper, we have explored a new paradigm for scene text recognition, where precise recognition can be achieved with just one token. Through constructing a ViT-based image-to-vector encoder, our one token recognizer successfully eliminates the requirement of sequential tokens in scene text recognition and proves that **One token** is sufficient for sequential character-wise prediction. In addition, the character-level fine-grained information is also proven to enhance the image-text retrieval. The extensive experiments demonstrate the effectiveness of our method. Our method provides a new perspective on OCR tasks, and we hope that this simple and effective method can inspire more community researchers.

7. Acknowledgments

This work is supported by the National Key Research and Development Program of China (2022YFB3104700), the National Nature Science Foundation of China (U23B2028, 62121002, 62102384).

References

[1] David Aldavert, Marçal Rusinol, Ricardo Toledo, and Josep Lladós. Integrating visual and textual cues for query-by-string word spotting. In *2013 12th International conference on document analysis and recognition*, pages 511–515. IEEE, 2013. 3

[2] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2552–2566, 2014. 3

[3] Rowel Atienza. Vision transformer for fast and efficient scene text recognition. In *International Conference on Document Analysis and Recognition*, pages 319–334. Springer, 2021. 6, 7

[4] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4715–4723, 2019. 6, 7

[5] Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. In *European Conference on Computer Vision*, pages 178–196. Springer, 2022. 1, 2, 5, 6, 7

[6] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1571–1576. IEEE, 2019. 5, 7

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 5

[8] Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoting Yin, Tianlun Zheng, Chenxia Li, Yuning Du, and Yu-Gang Jiang. Svtr: Scene text recognition with a single visual model. *arXiv preprint arXiv:2205.00159*, 2022. 1, 2, 5, 6, 7

[9] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. 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*, pages 7098–7107, 2021. 1, 2, 5, 6, 7

[10] Lluís Gómez, Andrés Mafla, Marçal Rusinol, and Dimosthenis Karatzas. Single shot scene text retrieval. In *Proceedings of the European conference on computer vision (ECCV)*, pages 700–715, 2018. 3, 7

[11] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *CVPR*, pages 2315–2324, 2016. 5, 6

[12] Tong He, Zhi Tian, Weilin Huang, Chunhua Shen, Yu Qiao, and Changming Sun. An end-to-end textspotter with explicit alignment and attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5020–5029, 2018. 1, 7

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[14] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, 2014. 5, 6

[15] Qing Jiang, Jiapeng Wang, Dezhi Peng, Chongyu Liu, and Lianwen Jin. Revisiting scene text recognition: A data perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20543–20554, 2023. 5, 6, 7

[16] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernández Mota, Jon Almazán, and Lluís-Pere de las Heras. Icdar. pages 1484–1493, 2013. 5, 7

[17] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman K. Ghosh, Andrew D. Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. ICDAR 2015 competition on robust reading. In *ICDAR*, pages 1156–1160, 2015. 5, 7

[18] Junyeop Lee, Sungrae Park, Jeonghun Baek, Seong Joon Oh, Seonghyeon Kim, and Hwalsuk Lee. On recognizing texts of arbitrary shapes with 2d self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 546–547, 2020. 7

[19] Minghui Liao, Guan Pang, Jing Huang, Tal Hassner, and Xiang Bai. Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In *European Conference on Computer Vision*, pages 706–722. Springer, 2020. 7

[20] Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9809–9818, 2020. 7

[21] Canjie Luo, Lianwen Jin, and Zenghui Sun. Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, 90:109–118, 2019. 7

[22] Andrés Mafla, Ruben Tito, Sounak Dey, Lluís Gómez, Marçal Rusiñol, Ernest Valveny, and Dimosthenis Karatzas. Real-time lexicon-free scene text retrieval. *Pattern Recognition*, 110:107656, 2021. 7

[23] Anand Mishra, Karteek Alahari, and C. V. Jawahar. Scene text recognition using higher order language priors. In *BMVC*, pages 1–11, 2012. 5, 7

[24] Anand Mishra, Karteek Alahari, and CV Jawahar. Image retrieval using textual cues. In *Proceedings of the IEEE international conference on computer vision*, pages 3040–3047, 2013. 3, 5, 7

[25] Byeonghu Na, Yoonsik Kim, and Sungrae Park. Multi-modal text recognition networks: Interactive enhancements between visual and semantic features. In *European Conference on Computer Vision*, pages 446–463. Springer, 2022. 7

[26] Nibal Nayef, Yash Patel, Michal Busta, Pinaki Nath Chowdhury, Dimosthenis Karatzas, Wafa Khlif, Jiri Matas, Umapada Pal, Jean-Christophe Burie, Cheng-lin Liu, et al. Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 1582–1587. IEEE, 2019. 5

[27] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *ICCV*, pages 569–576, 2013. 5, 7

[28] Zhi Qiao, Yu Zhou, Dongbao Yang, Yucan Zhou, and Weiping Wang. Seed: Semantics enhanced encoder-decoder framework for scene text recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13528–13537, 2020. 6

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4

[30] Anhar Risnumawan, Palaiahnakote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. *Expert Syst. Appl.*, pages 8027–8048, 2014. 5, 7

[31] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2016. 1, 2, 6, 7

[32] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2035–2048, 2018. 7

[33] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 3

[34] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE, 2015. 2

[35] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 8

[36] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016. 5, 7

[37] Zhaoyi Wan, Minghang He, Haoran Chen, Xiang Bai, and Cong Yao. Textscanner: Reading characters in order for robust scene text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12120–12127, 2020. 1, 2, 6

[38] Hao Wang, Xiang Bai, Mingkun Yang, Shenggao Zhu, Jing Wang, and Wenyu Liu. Scene text retrieval via joint text detection and similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4558–4567, 2021. 3, 4, 5, 7, 8

[39] Kai Wang, Boris Babenko, and Serge J. Belongie. End-to-end scene text recognition. In *ICCV*, pages 1457–1464, 2011. 5, 7

[40] Peng Wang, Cheng Da, and Cong Yao. Multi-granularity prediction for scene text recognition. In *European Conference on Computer Vision*, pages 339–355. Springer, 2022. 6

[41] Peng Wang, Cheng Da, and Cong Yao. Multi-granularity prediction for scene text recognition. In *European Conference on Computer Vision*, pages 339–355. Springer, 2022. 1, 2

[42] Tianwei Wang, Yuanzhi Zhu, Lianwen Jin, Canjie Luo, Xiaoxue Chen, Yaqiang Wu, Qianying Wang, and Mingxiang Cai. Decoupled attention network for text recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12216–12224, 2020. 6, 7

[43] Yuxin Wang, Hongtao Xie, Shancheng Fang, Jing Wang, Shenggao Zhu, and Yongdong Zhang. 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 (ICCV)*, pages 14194–14203, 2021. 2, 6, 7

[44] Zixiao Wang, Hongtao Xie, Yuxin Wang, Jianjun Xu, Boqiang Zhang, and Yongdong Zhang. Symmetrical linguistic feature distillation with clip for scene text recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 509–518, 2023. 1

[45] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Towards accurate scene text recognition with semantic reasoning networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12113–12122, 2020. 6, 7

[46] Xiaoyu Yue, Zhanghui Kuang, Chenhao Lin, Hongbin Sun, and Wayne Zhang. Robustscanner: Dynamically enhancing positional clues for robust text recognition. In *European Conference on Computer Vision*, pages 135–151. Springer, 2020. 6, 7

[47] Boqiang Zhang, Hongtao Xie, Yuxin Wang, Jianjun Xu, and Yongdong Zhang. Linguistic more: taking a further step toward efficient and accurate scene text recognition. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 1704–1712, 2023. 1

[48] Xinyun Zhang, Binwu Zhu, Xufeng Yao, Qi Sun, Ruiyu Li, and Bei Yu. Context-based contrastive learning for scene text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3353–3361, 2022. 6

[49] Ying Zhang, Lionel Gueguen, Ilya Zharkov, Peter Zhang, Keith Seifert, and Ben Kadlec. Uber-text: A large-scale dataset for optical character recognition from street-level imagery. In *SUNw: Scene Understanding Workshop-CVPR*, page 5, 2017. 5, 7