OTE: Exploring Accurate Scene Text Recognition Using One Token (Supplementary Material)

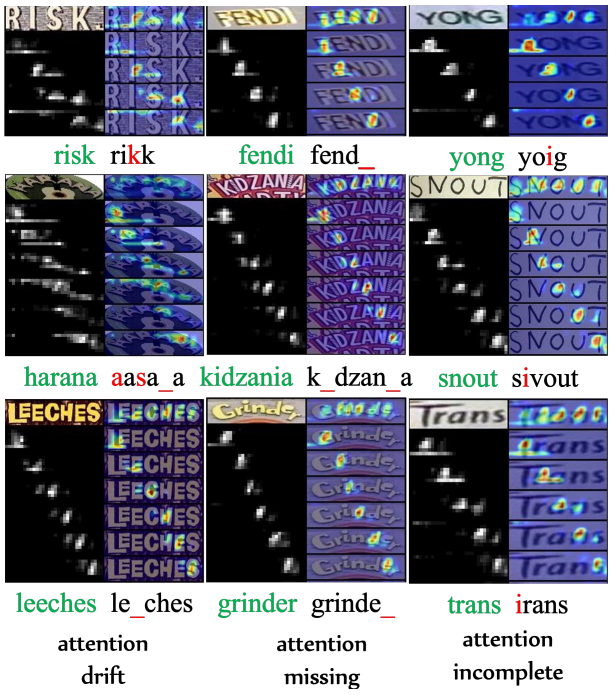

Figure 1. Visualization of inaccurate attention maps, including attention drift, attention missing, and attention incomplete.

1. Inaccurate attention maps in attention-based S2S decoder

The attention-based sequence-to-sequence decoder necessitates accurate attention maps for aligning visual features with text sequences. However, in complex scenarios characterized by blurriness, curvature, or variable lighting, decoders often struggle to generate accurate attention maps, which adversely affects text recognition performance. Inaccuracies in attention can be classified into three types: attention missing, attention drift, and attention incomplete, as depicted in Fig. 1. Such misalignments typically hinder the decoder's capability to pinpoint specific parts of visual features when decoding certain characters, leading to recognition errors. Fundamentally, this inaccuracy in attention is rooted in the insufficient robustness and distinctiveness of visual feature extraction. The decoding process is easily disrupted by similarities among visual features or by background noise, consequently underutilizing the visual information crucial for guiding the prediction of character sequences.

Table 1. The Effectiveness of Channel-wise Parallel Attention. ViT-small and auto-regressive decoding are used in this experiment.

Strategy	Regular Text			Irregular Text			Avg
	IIIT	SVT	IC13	IC15	SVTP	CUTE	
FC	95.7	93.1	97.5	85.7	88.4	89.9	92.3
CPA	**96.2**	**93.5**	**97.6**	**85.9**	**89.6**	**91.7**	**92.8**

Table 2. Evaluation of ResNet as backbone. ResNet represents using ResNet45 as the encoder to extract global semantics, and ViT represents using vit-small

Backbone	Regular Text			Irregular Text			Avg
	IIIT	SVT	IC13	IC15	SVTP	CUTE	
ResNet	93.8	92.1	95.8	82.6	83.7	88.2	90.0
ViT	**96.2**	**93.5**	**97.6**	**85.9**	**89.6**	**91.7**	**92.8**

2. Extensive experiments

2.1. The Effectiveness of Channel-wise Parallel Attention.

As our image-to-vector encoder employs a singular token for sequential decoding in scene text recognition, we implemented a parallel vector-to-sequence strategy to underscore the efficacy of our Channel-wise Parallel Attention (CPA) method. To demonstrate this, we replaced CPA with three fully connected layers of a similar parameter scale, combined with activation layers, for the transformation of global semantics to sequence embeddings, while keeping all other structures unchanged. As shown in Tab. 1, the results indicate that our CPA outperforms the analogous fully connected layers in all datasets, particularly in irregular text collections. This underscores the superior diversity and distinctiveness of text embeddings generated by our CPA.

Table 3. Comparison with SOTA methods on challenging datasets. OTE uses autoregressive decoding by default, all methods are trained on MJ and ST.

Method	ArT	COCO	Uber	Avg
CRNN	57.3	49.3	33.1	46.6
ViTSTR	66.1	56.4	37.6	53.4
TRBA	68.2	61.4	38.0	55.9
ABINet	65.4	57.1	34.9	52.5
PARSeq$_A$	**70.7**	64.0	42.0	58.9
OTE$_{\text{ViT-S}}$	68.8	63.0	46.8	59.5
OTE$_{\text{ViT-B}}$	70.1	64.3	**48.2**	**60.9**
OTE$_{\text{SVTR}}$	69.1	**64.5**	47.8	60.5

2.2. Evaluation of ResNet as backbone

We further experimented with using ResNet as the image-to-vector encoder to extract global semantics. Specifically, we employed ResNet45 as our backbone, applying 2-D global average pooling to the output features of the final layer to obtain global semantics, while maintaining all other structures unchanged. The results, presented in Tab. 2, demonstrate a notable performance advantage of ViT over ResNet. Thanks to ViT's dynamic attention mechanism and long-range modeling capabilities, it significantly outperforms ResNet, showing an average performance gain of 2.8% across six datasets. This finding confirms that our ViT-based image-to-vector encoder effectively captures comprehensive multi-grained global semantics, leading to precise text recognition.

2.3. More experiments on more challenging benchmarks

We conduct additional experiments on more challenging benchmarks, maintaining consistency with the experimental setup detailed in the main text. We select ArT, COCO, and Uber as our test datasets to evaluate the performance of our method's ensemble on these datasets, as illustrated in Tab. 3. The results indicate that our models, varying in size and structure, consistently achieve leading performance on these three challenging datasets. This further validates the effectiveness of our approach.

3. Discussion on Scene Text Retrieval and Scene Text Recognition

To a certain extent, retrieval and classification tasks are feature extraction and comprehension tasks at different levels. Scene text retrieval needs to match the visual representations of a large number of candidate scene texts with the text representations of a specific query, while scene text recognition maps the visual representations of specific text images to a specified label space, both of which require accurate and robust representations. The OTE, by aggregating the visual representation into a single token, naturally provides a unified interface for these two tasks. Additionally, by introducing character-wise fine-grained information, such global tokens also enhance the performance of both scene text recognition and scene text retrieval tasks. To further explore the robustness of the visual features extracted by our OTE, we explored the potential of directly using retrieval models for zero shot classification, as visualized in Fig. 2. The results indicate that even under the interference of blur and lighting, the retrieval model can correctly match text and images.

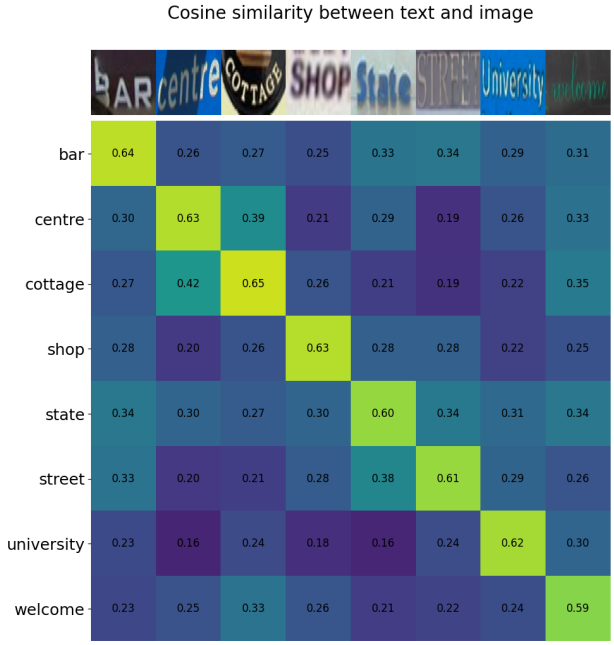

Figure 2. Visualization of inaccurate attention maps, including attention drift, attention missing, and attention incomplete.