CLIP is Almost All You Need: Towards Parameter-Efficient Scene Text Retrieval without OCR

Xugong Qin[1], Peng Zhang[1,5,*], Jun Jie Ou Yang[3], Gangyan Zeng[1], Yubo Li[2], Yuanyuan Wang[1], Wanqian Zhang[2], Pengwen Dai[4]
[1]School of Cyber Science and Engineering, Nanjing University of Science and Technology
[2]Institute of Information Engineering, Chinese Academy of Sciences
[3]University of Southern California, [4]Sun Yat-sen University
[5]Laboratory for Advanced Computing and Intelligence Engineering
qinxugong@njust.edu.cn

Abstract

Scene Text Retrieval (STR) seeks to identify all images containing a given query string. Existing methods typically rely on an explicit Optical Character Recognition (OCR) process of text spotting or localization, which is susceptible to complex pipelines and accumulated errors. To settle this, we resort to the Contrastive Language-Image Pre-training (CLIP) models, which have demonstrated the capacity to perceive and understand scene text, making it possible to achieve strictly OCR-free STR. From the perspective of parameter-efficient transfer learning, a lightweight visual position adapter is proposed to provide a positional information complement for the visual encoder. Besides, we introduce a visual context dropout technique to improve the alignment of local visual features. A novel, parameter-free cross-attention mechanism transfers the contrastive relationship between images and text to that between visual tokens and text, producing a rich cross-modal representation, which can be utilized for efficient reranking with a linear classifier. The resulting model, ***CAYN****, which proves that* ***C****LIP is* ***A****lmost all* ***Y****ou* ***N****eed for STR with no more than 0.50M additional parameters required, achieves new state-of-the-art performance on the STR task, with 92.46%/89.49%/85.98% mAP on the SVT/IIIT-STR/TTR datasets. Our findings demonstrate that CLIP can serve as a reliable and efficient solution for OCR-free STR.*

1. Introduction

Scene text extraction and understanding [27–32, 38, 47, 48] have received significant attention in recent years due to the important role of text as an information carrier. Among these fundamental tasks, scene text retrieval (STR) [35, 43] aims to find all the images containing the given text query from a gallery, which receives great attention due to its practical applications in content retrieval and information security, e.g., product retrieval [1], electronic book archives management [42], and video key frame extraction [36].

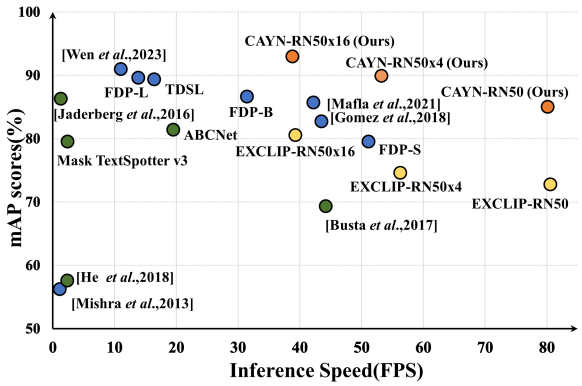


Figure 1. Illustration of the trade-off between accuracy (mAP) and inference speed (FPS) on the SVT dataset. The proposed CAYN achieves a better balance than existing methods.

As a pioneer work, Mishra *et al.* [25] define STR as a text-to-image retrieval problem and point out the inefficiency of imposing an exact localization-and-recognition (spotting) pipeline. To solve it, they propose a query-driven search approach where approximate locations of characters in the text query are first found, then the images are ranked with the likelihood of containing characters from the query text for matching, and finally, spatial constraints are imposed for re-ranking to generate a ranked list of images. The method sets up a standard localization-based pipeline for STR, which outperforms spotting-based methods [8, 12, 20, 21] in terms of both speed and accuracy [11]. For this reason, a detection or segmentation head has been a common component in subsequent research [7, 39, 41, 49].

*Corresponding author.

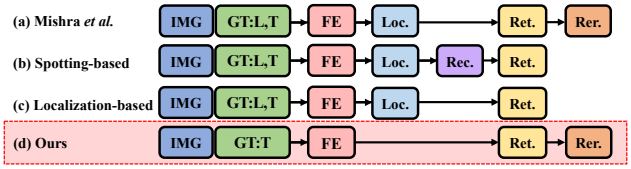


Figure 2. Comparison of pipelines between existing STR methods and ours (the text branch is omitted for simplicity). "L" and "T" represent the supervision of localization and transcription levels. "FE" denotes the feature extractor. "Loc": Localization; 'Rec": Recognition; "Ret": Retrieval; "Rer": Reranking.

Benefiting from the sharing of feature extractors and joint optimization in localization and matching, the localization-based approach has achieved remarkable progress in recent years as shown in Fig. 1. However, localization, as a necessary process in existing works, introduces inevitable accumulated errors and complicates the whole STR pipeline. The other issue lies in the inefficiency and inflexibility of training and maintaining a scene text retriever, which requires a long pre-training period on large-scale synthetic data [7] and a full fine-tuning of a large number of parameters on a real STR dataset [26].

Recently, the contrastive language-image pre-training (CLIP) [33] models are proposed, which bridge the gap between the modality of vision and language, have demonstrated the ability to perceive and understand scene text, bringing new paradigms to the document analysis and understanding community [40, 44, 45]. FDP [49] first explores the leverage of the CLIP models for STR, which first performs a tough text localization at the feature level and then utilizes the localization information and "scene text" prompt to guide the CLIP models to focus on scene text regions. Queries are distinguished as content and function words combined with class-aware learnable prompts to tune the CLIP models. Despite impressive performance, FDP requires an explicit localization process, resulting in a complex pipeline. Besides, FDP introduces more than 20M extra parameters, resulting in a long training process and substantial parameters to be maintained. In this regard, a natural question is raised: Is it possible to realize strictly OCR-free but accurate STR while keeping efficient at both training and inference stages?

To answer the question, we first carefully examine the potential of CLIP on STR by exploring hand-crafted textual prompt design and visual position embedding interpolation. As a result, we find that interpolating visual position embeddings can already serve as a strong baseline without introducing extra annotations or fine-tuning. However, directly interpolating visual position embeddings may introduce positional information loss. As compensation, we propose a lightweight visual position adapter (VPA) to provide positional information for the visual encoder. Inspired by Mishra *et al.* [25], we seek to rerank the top images ranked by the contrastive cross-modal similarity. A visual context dropout (VCD) technique is introduced to generate locally aligned visual features. Besides, we propose a novel parameter-free cross-attention mechanism that transfers the contrastive relationship between images and text to that between tokens and text, resulting in a rich cross-modal representation. We employ a linear classifier on the representation to predict the degree of image-text matching as the reranking scores. Extensive experiments on three benchmarks show that the proposed method, CAYN, can achieve SOTA performance with fast speed. CAYN achieves strictly OCR-free STR with no more than 0.50M additional parameters required. The comparison between CAYN and existing methods is illustrated in Fig. 2.

The overall contributions can be summarized as follows:

- We realize the first strictly OCR-free STR without explicit localization or recognition processes, eliminating the intermediate errors and simplifying the STR pipeline.
- A comprehensive study is performed to evaluate the ability of CLIP models on STR, providing a strong zero-shot baseline. From the perspective of parameter-efficient transfer learning (PEFT), a visual position adapter with no more than 0.50M is introduced to adapt the CLIP models to a large input resolution efficiently.
- The proposed method aims to unleash the potential of CLIP models to a maximum extent. Novel parameter-free attention is proposed for multimodal interaction based on fine-grained representation; only a light linear classification head is required for reranking.
- Experiments on three public datasets across ResNet- and ViT-based CLIP models demonstrate that CAYN outperforms existing methods in terms of speed and accuracy.

2. Related Work

2.1. Scene Text Retrieval

Existing STR methods can be categorized into localization-based [7, 24, 39, 41, 49] and spotting-based [8, 12, 20, 21, 25] in which explicit localization (and recognition) process is/are utilized to perceive scene text.

Spotting-Based STR firstly utilizes an OCR engine for text extraction and then matches results with the given query. Normalized Levenshtein distance, widely used in string matching, is adopted as the similarity measure for ranking all image candidates. Jaderberg *et al.* [12] leverage a deep convolutional network for text spotting, in which the matching score is computed by averaging the word probability distributions across all detections in an image. With the rapid development of scene text spotting, state-of-the-art text spotting methods [2, 20, 21] are adopted to improve the performance. However, the approach suffers from both the accumulative errors from the localization and recognition, making it hard to meet the high speed and accuracy requirement in real applications.

Localization-Based STR formulates STR as two combined tasks: text localization and word-spotting, which makes it easy to borrow from the two well-developed subfields. Considering the requirement of fast inference speed, fast one-stage detection frameworks [34, 37] are utilized for constructing efficient STR system [24, 34, 39, 41]. According to the measurement of the manner of similarity between the query and the document candidates, existing methods can be mainly divided into PHOC-based [7, 24], cross-modal similarity-based [39, 49], and prototype-based [41]. Upon the CLIP visual features, FDP [49] adds a segmentation head for scene text localization, and then the obtained segmentation map is binarized and taken as the attention mask to generate the global image representation, which makes it not fully OCR-free. Although localization-based STR has achieved great progress, whether an explicit localization process is required for STR is still an open problem.

2.2. OCR-Free Document Understanding

Recently, a series of OCR-free document understanding methods have been proposed, benefiting greatly from large-scale pre-training. For instance, Donut [15] proposes the first end-to-end training OCR-free document transformer, pre-trained with large-scale synthetic documents. Pix2Struct [16] is pre-trained by learning to parse masked screenshots of web pages into structured format i.e., simplified HTML, followed by a variable-resolution input document representation. StrucTexTv2 [46] proposes a self-supervised pre-training framework, in which segment-level document image masking is introduced to learn joint visual-textual representation. Compared to common document understanding tasks like DocVQA, the STR task can be viewed as a relatively low-level document understanding task that predicts whether the documents contain a provided query. Through large-scale pre-training, the connection between visual text and text is established, bringing new possibilities for document understanding as well as STR.

2.3. Parameter-Efficient Tuning

With the prevalence of pre-training models, e.g., BERT [4] and CLIP [33], parameter-efficient tuning (PET) becomes a practical technique that maintains the ability of pre-training models and can efficiently transfer to downstream tasks [3] with few parameters and computations. Among them, adapters [3, 9, 10] and prompt learning [13, 17, 51] are two dominant and active branches in PET.

Adapters [9] are lightweight learnable modules inserted in pre-trained models. Specifically, the weights from the pre-training stage are frozen to preserve general knowledge and the parameters of adapters are tuned to learn task-specific knowledge. Hu *et al.* [10] propose low-rank adaptation (LoRA), which freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the transformer architecture, greatly reducing the number of trainable parameters for downstream tasks. Chen *et al.* [3] introduce an effective adaptation approach for visual transformer, which can adapt the pre-trained ViTs into many different image and video tasks. CLIP-Adapter [5] proposes to conduct fine-tuning with feature adapters on either the visual or language branch, which can achieve competitive performance while maintaining a simple design.

Prompt Learning [17] aims to unleash the potential of the pre-trained language/vision-language models with learnable prompt input. Li *et al.* [19] propose a lightweight alternative to fine-tuning for natural language generation tasks, which keeps language model parameters frozen and instead optimizes a sequence of continuous task-specific vectors, allowing subsequent tokens to attend to this prefix. CoOP [52] proposes to introduce a learnable context in text prompt for vision-language models like CLIP, which is improved by CoCoOp [51] via introducing context conditioned on visual inputs. Comparatively, VPT [13] proposes visual prompt learning which involves learnable context in the visual backbone. For STR, FDP [49] introduces class-aware prompt learning, which learns individual contexts for content and function words separately. Despite achieving great progress, PET remains less explored for STR.

3. Methodology

Given a set of text queries $Q = \{Q_1, ..., Q_{N_Q}\}$, STR aims to find all the images that contain the text query from the document gallery $D = \{D_1, ..., D_{N_D}\}$. The overall framework of the proposed framework is shown in Fig. 3, which includes retrieval in a contrastive manner and reranking according to the matching degree of image-text pairs.

The section is organized as follows. We first attempt to adapt CLIP models for the STR task with different input image resolutions and text prompts, followed by the retrieval and reranking parts. The label generation and optimization processes are presented at last.

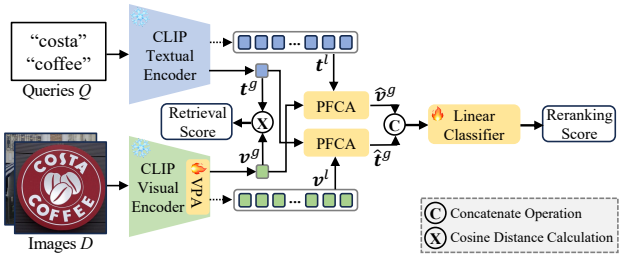


Figure 3. The overall framework of the proposed method. It mainly comprises the retrieval and reranking stages, in which only the VPA and the linear classifier are tunable.

3.1. Exploring CLIP's Ability on STR

Despite the great success of adapting CLIP to visual [52] and cross-modal tasks [14], the adaption of CLIP for the

STR task remains less explored. To achieve this, we perform pre-experiments from two aspects i.e.: exploring predefined prompts and extending to a large-resolution image input. We denote the extension of the original CLIP as "EXCLIP", which is demonstrated to be a strong baseline.

Retrieval with CLIP. The queries and documents are encoded by the language and vision encoders of CLIP respectively to obtain the global representation $F_Q \in \mathbb{R}^{N_Q \times E}$ and $F_D \in \mathbb{R}^{N_D \times E}$ (E denotes the embedding dimension), and the cosine similarity between each query-document pair from F_Q and F_D is computed to rank all the documents in a paradigm of dense retrieval, resulting a similarity matrix $S \in \mathbb{R}^{N_Q \times N_D}$ for evaluation.

Table 1. Zero-shot performance of EXCLIP-RN50 at a resolution of 512 with different predefined prompts on the SVT and TTR datasets. {query} denotes the content of a query string.

Predefined prompt (string content)	SVT	TTR
{query}	65.56	33.63
scene text {query}	49.06	23.86
the text {query}	53.55	25.91
the word {query}	67.08	36.16
"{query}"	**72.81**	<u>41.87</u>
the word "{query}"	69.54	**42.01**
a photo contains the word "{query}"	<u>69.86</u>	41.74
the word "{query}" in a photo	68.57	38.57

Hand-Craft Textual Prompt Design. Following the prompt engineering in [33], we attempt to endow the input query with more specific semantics for the downstream STR task. As shown in Tab. 1, interestingly, we find that a simple prompt with a pair of quotation marks shows a strong semantic relation with the text query, which performs best with the RN50 backbone among all predefined prompts and is chosen as the default prompt template for the work.

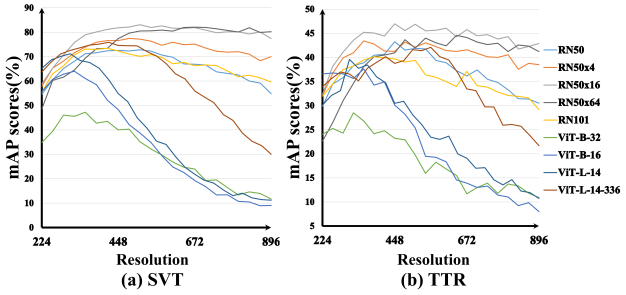


Figure 4. Zero-shot performance of EXCLIP models at different input resolutions on the SVT and TTR datasets.

Visual Position Embedding Interpolation. One major limitation of extending CLIP to large input is the fixed-length position encoding in attention layers. Since the relative 2D position relationship is kept in the visual backbone, one direct solution is directly interpolating the embedding to different sizes. In particular, "bicubic" interpolation is adopted to transform pre-trained visual position embedding, which allows the CLIP models to deal with other sizes of input. As shown in Fig. 4, several points can be concluded 1) Interpolating position embeddings in ResNet backbones generalizes much better than that in ViT backbones (can only work well in a relatively small scope) on larger input sizes. 2) Performance improves with the increase of parameters, demonstrating the scaling ability on the STR task. 3) Surprisingly, we find the RN50x16 model can achieve a performance of 83.08% at most, which is already comparable to single-shot-str [7], demonstrating the potential of CLIP models being a strong scene text retriever. Based on the above results, we mainly choose the RN50, RN50x4, RN50x16, and ViT-B-16 in experiments and set the image size of adapted models as shown in Tab. 2, considering the trade-off between speed and accuracy.

Table 2. The setting of models in this work. We expand the original size in CLIP to better adapt to the STR task.

Backbone	Org_Size	Exp_Size	E
RN50	224	512	1024
RN50x4	288	576	640
RN50x16	384	640	768
ViT-B-16	224	512	512

For the CLIP-ViT models, we find that the generalization of interpolate visual position embedding is relatively weak, which may be attributed to the shallow encoding of position information in the transformer architecture. We follow existing works to combine a splitting strategy [22] to ensure each split is within the working scope of the CLIP models. For an image with 512 $\times$ 512 input size, we can split the input image with 2 $\times$ 2 subdivisions for the ViT-B-16 model. The similarity scores are computed between the query and each sub-image, in which the highest score serves as the global matching score. ResNet models demonstrate a higher upper limit even though the split is introduced in ViT models. However, PET adaption of ViT models remains deserved for research, which is absent in existing work [49].

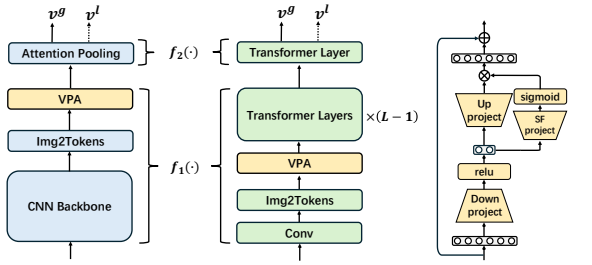
3.2. Retrieval with Rich Visual Representation

Besides the improvement by tuning text prompts and enlarging image resolutions, a novel lightweight visual position adapter (VPA) is proposed for efficient adaption. The proposed VPA aims to solve the challenge of the large variation in resolution from the perspective of PET, which generates rich visual features by providing the positional information supplement with a high parameter compression rate.

Visual Position Adapter. Unlike existing works that introduce layer-wise adapters [6, 10] in encoders or adapters at the end of encoders [5], the proposed VPA is inserted after convolutional features are flattened into tokens before the attention layer, as shown in Fig. 5. Given an input of $\boldsymbol{x} \in \mathbb{R}^{1 \times d}$, the feed-forward process of the adapter module can be written as follows:

$$\boldsymbol{x}_r = \sigma(\boldsymbol{x}\mathbf{W}_{down}), \tag{1}$$

$$\boldsymbol{x}_{out} = \boldsymbol{x} + \text{sigmoid}(\boldsymbol{x}_r\mathbf{W}_{sf}) \cdot (\boldsymbol{x}_r\mathbf{W}_{up}), \tag{2}$$



(a) CNN-based Image Encoder (b) ViT-based Image Encoder (c) Visual Position Adapter

Figure 5. Illustration of the VPA and VCD in CLIP visual encoders based on (a): ResNet and (b): ViT. The detailed architecture of VPA is shown in (c). The dotted line denotes that local features are generated with VCD.

where $W_{down} \in \mathbb{R}^{d\times\frac{d}{r}}$, $W_{up}, W_{sf} \in \mathbb{R}^{\frac{d}{r}\times d}$ ($\frac{d}{r} \ll d, r$ denotes the reduction ratio) are down-projection, up-projection, and a learnable scale factor projection respectively, and σ is a non-linear layer which is implemented with a ReLU in the work.

Considering a batch of N matched image-text pairs, the similarities across the global visual features $\boldsymbol{v}^g \in \mathbb{R}^{N\times E}$ and textual features $\boldsymbol{t}^g \in \mathbb{R}^{N\times E}$ are computed to obtain a $N \times N$ matrix. The symmetric cross-entropy loss [33] is adapted for optimization, which can be defined as follows:

$$\mathcal{L}_{retrieval} = -\frac{1}{2N}\Big(\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(\boldsymbol{t}_i^g, \boldsymbol{v}_i^g)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\boldsymbol{t}_i^g, \boldsymbol{v}_j^g)/\tau)} + \sum_{i=1}^{N}\log\frac{\exp(\text{sim}(\boldsymbol{v}_i^g, \boldsymbol{t}_i^g)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\boldsymbol{v}_i^g, \boldsymbol{t}_j^g)/\tau)}\Big), \tag{3}$$

where $\text{sim}(.,.)$ and τ denote the cosine similarity computation and the learnable temperature factor, respectively.

3.3. Reranking with Fine-grained Alignment

In the reranking stage, a classical image-text matching task is performed with features from the CLIP encoders.

Visual Context Dropout. The CLIP visual encoder can be formulated as $f = f_2 \circ f_1$ as shown in Fig. 5, where f_1 and f_2 denote the front and rear parts of the visual encoder, respectively. Specifically, f_2 denotes the module that includes the last attention layer, i.e., the attention pooling layer in ResNet and the last transformer layers in ViT. A global feature is pooled from the local features according to the attention mechanism, which can be seen as a weighted sum of transformed tokens by the value projection. From the perspective of each token, attention brings context as well as noise; the alignment with the cross-modal representation space learned by CLIP may be suboptimal. Due to the symmetrical relationship for all visual tokens, the value projection may already be good enough for fine-grained alignment, which has been demonstrated in zero-shot semantic segmentation [50]. So we drop out the visual context in attention layers in f_2, degenerate attention to a linear transform (parameterized by the value projection) to get locally aligned visual features $\boldsymbol{v}^l \in \mathbb{R}^{N\times N_V\times E}$.

Parameter-Free Cross-Attention. The cross-attention mechanism can efficiently enhance global features by incorporating local features, which are usually implemented through a multi-head attention layer [18]. However, it is challenging to achieve cross-modal fine-grained alignment with limited data in downstream tasks. Alternatively, we propose to directly re-utilize aligned local visual features $\boldsymbol{v}^l$ to perform a parameter-free cross-attention queried by global textual features $\boldsymbol{t}^g$, based on the assumption that the contrastive relationship between batched texts and images can be well transferred to that between texts and tokens from the same image if the local token-level features $\boldsymbol{v}^l$ are in approximately the same presentation space with the global image-level features $\boldsymbol{v}^g$ (illustrated as $\boldsymbol{t}^g \leftrightarrows \boldsymbol{v}^g \leftrightarrows \boldsymbol{v}^l$). We use a residual connection to combine the aggregated features. The process can be represented as:

$$\hat{\boldsymbol{t}}_i^g = \boldsymbol{t}_i^g + \frac{\exp(\text{sim}(\boldsymbol{t}_i^g, \boldsymbol{v}_{i,j}^l)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\boldsymbol{t}_i^g, \boldsymbol{v}_{i,j}^l)/\tau)}\boldsymbol{v}_{i,j}^l, \tag{4}$$

Symmetrically, we can aggregate local text features $\boldsymbol{t}^l \in \mathbb{R}^{N\times N_T\times E}$ queried by the global visual features $\boldsymbol{v}^g$:

$$\hat{\boldsymbol{v}}_i^g = \boldsymbol{v}_i^g + \frac{\exp(\text{sim}(\boldsymbol{v}_i^g, \boldsymbol{t}_{i,j}^l)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\boldsymbol{v}_i^g, \boldsymbol{t}_{i,j}^l)/\tau)}\boldsymbol{t}_{i,j}^l. \tag{5}$$

The enhanced features are then concatenated along the embedding dimension to predict the matching degree of the image-text pair $\boldsymbol{p} \in \mathbb{R}^{M\times 2}$ with a linear classifier parameterized with $\mathbf{W}_{itm} \in \mathbb{R}^{2E\times 2}$:

$$\boldsymbol{p} = \text{CAT}(\hat{\boldsymbol{t}}^g, \hat{\boldsymbol{v}}^g)\cdot\mathbf{W}_{itm}. \tag{6}$$

The predicted logits $\boldsymbol{p}$ are then normalized and supervised with a cross-entropy loss:

$$\mathcal{L}_{reranking} = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{\exp(\boldsymbol{p}_{i,\boldsymbol{y}_i})}{\sum_{j=1}^{2}\exp(\boldsymbol{p}_{i,j})}, \tag{7}$$

where $\boldsymbol{y}_i$ denotes the matching label of the i-th image-text pair within the M pairs.

For the ViT models, the split sub-images are batched encoded, resulting in global features $\boldsymbol{v}^g \in \mathbb{R}^{N\times N_s\times E}$ and local features $\boldsymbol{v}^l \in \mathbb{R}^{(N\times N_s)\times L\times E}$. We reshape the local features to concatenate the splits along the context length dimension to $\boldsymbol{v}^l \in \mathbb{R}^{N\times(N_s\times L)\times E}$; the global feature $\boldsymbol{v}^g$ is aggregated with a PFCA queried by the global textual features $\boldsymbol{t}^g$, similarly with attention in Eq. (4) but the attention dimension changes from the contexts to splits.

During inference, only images with top rank from the retrieval stage participate in the reranking process. For top K ($K \ll N_D$) images per query, we update the matching scores with the sum of the original scores and the image-text matching scores, making the reranking quite efficient.

Table 3. Comparison with existing methods on the SVT, IIIT-STR, and TTR datasets. **Bold** indicates the best performance, and underline indicates the second-best performance. "Params" means the number of fine-tuned parameters. "L" and "T" denote the supervision of localization and transcription labels, respectively.

Method	Venue	Params	Supervision	SVT	IIIT-STR	TTR	FPS
Jaderberg *et al.* [12]	IJCV'16	≈500.00M	L+T	86.30	66.50	-	0.30
Busta *et al.* [2]	ICCV'17	-	L+T	69.37	62.94	-	44.21
He *et al.* [8]	CVPR'18	-	L+T	80.54	66.95	-	2.35
Mask TextSpotter v3 [20]	ECCV'20	45.47M	L+T	84.54	74.48	72.42	2.40
ABCNet [21]	CVPR'21	36.88M	L+T	82.43	67.25	69.30	17.50
Mishra *et al.* [25]	ICCV'13	-	L+T	56.24	42.70	-	0.10
Gomez *et al.* [7]	ECCV'18	58.64M	L+T	83.73	69.83	66.02	43.50
Mafla *et al.* [24]	PR'21	58.64M	L+T	85.74	71.67	-	42.20
TDSL [39]	CVPR'21	33.85M	L+T	89.38	77.09	74.75	12.00
Wen *et al.* [41]	WSDM'23	-	L+T	90.95	77.40	80.09	11.00
FDP-S [49]	MM'24	40.68M	L+T	82.56	81.77	65.26	45.11
FDP-B [49]	MM'24	22.55M	L+T	86.64	86.65	73.63	31.43
FDP-L [49]	MM'24	33.45M	L+T	89.63	89.46	79.18	11.82
EXCLIP-RN50	ICML'21	-	-	72.81	63.70	41.87	82.54
EXCLIP-RN50x4	ICML'21	-	-	74.64	61.84	42.23	56.26
EXCLIP-RN50x16	ICML'21	-	-	81.60	64.04	44.26	39.27
CAYN-RN50 (Ours)	-	**0.20M**	T	85.05	82.88	74.04	**80.13**
CAYN-RN50x4 (Ours)	-	0.31M	T	88.94	85.77	81.90	53.22
CAYN-RN50x16 (Ours)	-	0.45M	T	**92.46**	**89.49**	**85.98**	38.79

3.4. Label Generation and Optimization

Label Generation. Unlike the noisy data in CLIP pre-training, accurate query-document matching labels are available in STR tasks. We follow the label generation in CLIP [33] to provide one-hot labels in the retrieval stage, which works well in the downstream STR task.

In the reranking stage, the use of accurate matching labels and hard example sampling for optimization is proven to be crucial. To be specific, for image-text pairs, we adopt a hardness-aware sampling, in which the negatives are sampled according to the normalized softmax probability within a batch. The ratio of positive and negative pairs is set to 1:2.

Optimization. The retrieval and reranking stages are jointly optimized, which can be defined as:

$$\mathcal{L} = \mathcal{L}_{retrieval} + \lambda \mathcal{L}_{reranking}, \tag{8}$$

where λ is a balance factor for the retrieval and reranking losses and is set to 1.0 in our experiment.

4. Experiments

4.1. Datasets

The main experiments are conducted on common STR benchmarks including Street View Text (SVT), IIIT Scene Text Retrieval (IIIT-STR), and Total-Text Retrieval (TTR). For more details, please refer to the Appendix.

Following the existing methods [39, 41, 49], the model is only trained on the MLT5k dataset and tested on other datasets, including SVT, IIIT-STR, and TTR. The retrieval performance is evaluated with the mean average precision (mAP) measure following existing works [25, 39], which is the mean of the average precision for all the queries.

4.2. Implementation Details

Training. The models are trained for a total epoch of 10 on the MLT5k dataset with an AdamW [23] optimizer. The initial learning rate is set to 0.0005 with a batch size of 64 for CLIP-ResNet and 28 for CLIP-ViT, following a "constant" learning rate policy. The pre-processing includes resizing the long side of the image to an integer multiple of patch size, followed by zero-padding the image to a squared shape, which can avoid the information loss resulting from the cropping operation [33] in the pre-processing stage.

Inference. During inference, the pre-processing pipeline and the test scale are identical to those in the training stage. The models are evaluated following [49] with a batch size of 1. The number of reranked images K is set to 32 on all three datasets.

4.3. Comparisons with State-of-the-Art Methods

We validate the proposed method on several public datasets of different types, including SVT, IIIT-STR, and TTR, to demonstrate the effectiveness of CAYN on efficient and accurate scene text retrieval. As shown in Tab. 3, in general, the proposed method CAYN outperforms existing methods, including OCR-based and localization-based methods, and maintains a fast inference speed.

Accuracy-Speed Trade-Off. CAYN-RN50x16 achieves the SOTA performance of 92.46%, 89.49%, and 85.98% at 38.79 FPS on the SVT, IIIT-STR, and TTR datasets, respectively, and outperforms recently proposed FDP-L [49] by a large margin of 2.83% and 6.80% on the SVT and TTR datasets in terms of mAP. CAYN-RN50x16 runs at 38.79 FPS and is 3 times faster than FDP-L. CAYN-RN50

Table 4. Comparison of variants of CAYN with different model configurations on the SVT and TTR datasets.

#	Retrieval	Reranking		RN50			ViT-B-16		
	VPA	VCD	PFCA	Params	SVT	TTR	Params	SVT	TTR
0	-	-	-	-	72.81	41.87	-	68.43	39.23
1	✓	×	×	0.20M	81.98	66.54	0.22M	75.11	62.96
2	×	✓	✓	**4098**	80.19	55.67	**2050**	74.27	54.37
3	✓	×	✓	0.20M	82.83	68.02	0.22M	69.70	59.65
4	✓	✓	✓	0.20M	**85.05**	**74.04**	0.22M	**77.44**	**68.45**
5	✓	✓	MHCA	8.60M	79.09	68.10	2.30M	72.17	61.07

runs fastest at 80.13 FPS and achieves the performance of 85.05%/82.88%/74.04% mAP on the three datasets. CAYN-RN50x4 can achieve a good balance between speed and accuracy with 88.94%/85.77%/81.90% mAP at 53.22 FPS. Compared to mainstream localization-based methods, CAYN benefits from learning global-level similarity, enabling efficient image ranking without the errors associated with explicit localization.

Tuned Parameters & Reliance on Labels. From the perspective of PET, the tuned parameters of CAYN are 0.20M, 0.31M, and 0.45M in RN50, RNx4, and RN50x16 respectively, which are much less than existing methods (no more than 2% parameters of FDP [49]) that require 20M+ parameters to be tuned. Considering the requirement for annotations, due to the nature of localization-free, CAYN does not need localization labels. Only weak transcription-level annotations of image-text matching are required, without the need to know the number of instances.

4.4. Ablation Study

The ablation study is conducted on a regular text dataset SVT and a curved text dataset TTR to demonstrate the effectiveness of the proposed modules in Tab. 4, which include both ResNet-based and ViT-based CLIP models.

Effectiveness of VPA. As shown in Tab. 4, we can see that the proposed VPA improves the performance significantly for both RN50 and ViT-B-16 on the two datasets. For the RN50 backbone, VPA achieves 9.17% and 24.67% performance gain in terms of mAP on the SVT and TTR datasets. For the ViT-B-16 backbone, VPA brings 6.68% (on the SVT dataset) and 23.73% (on the TTR dataset) improvements. Moreover, compared to the total parameters of 102.01M and 149.62M for the RN50-based and ViT16-based CLIP models, VPA only brings 0.20M and 0.22M additional parameters to the two backbones.

Effectiveness of reranking with VCD and PFCA. As shown in Tab. 4 and Fig. 6, reranking can consistently improve the performance upon both the EXCLIP and VPA retrieval baselines with minimum parameters of a linear classifier and is also efficient in training. On the SVT and TTR datasets, reranking brings 7.38% and 13.80% (RN50), and 5.84% and 15.14% (ViT16) enhancement, respectively, compared with the EXCLIP baseline. Compared with the VPA baseline, improvements of 3.07%/7.50% (RN50) and

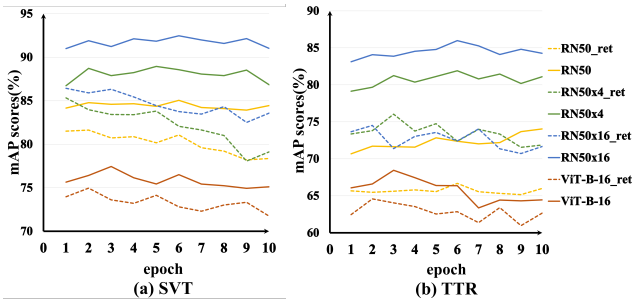


Figure 6. Comparison between CAYN and the retrieval-only (_ret) version with different backbone models on the SVT and TTR datasets. Reranking consistently improves performance.

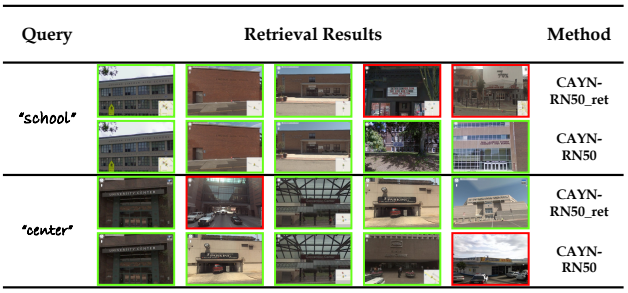
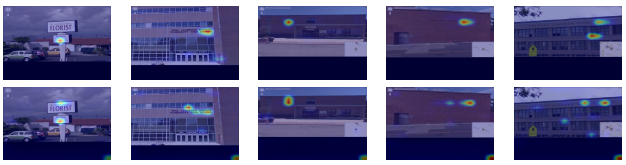
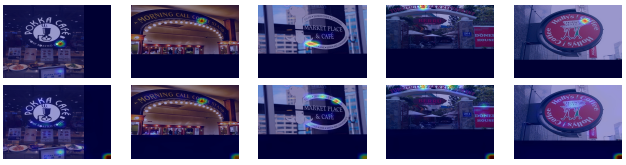


Figure 7. Visualization of Top-5 retrieval results. The correct and incorrect results are highlighted in green and red, respectively.



(a) Query: "school"



(b) Query: "coffee"

Figure 8. Visualization of attention maps of PFCA (upper rows) and MHSA (bottom rows) based on RN50.

2.33%/5.49% (ViT16) are obtained respectively. Besides the quantitative results, as shown in Fig. 7, the retrieval quality is also improved significantly with reranking.

When removing VCD, the performance on the SVT and TTR datasets drop 2.22% and 6.02% with RN50 and 7.74% and 8.80% with ViT16, demonstrating that the dropout of

visual context in the last attention layers brings better local alignment and is helpful in PFCA. To demonstrate the effectiveness of PFCA, we replace PFCA with a multi-head cross-attention (MHCA). Decreases of 5.96% and 5.94% (RN50) are found on SVT and TTR, as well as an mAP drop of 5.27% and 7.38% upon ViT16, which illustrates that CLIP models have been well aligned without the requirement of extra alignment layers. As shown in Fig. 8, PFCA can produce more accurate and concentrated attention maps than MHCA without introducing any parameters, demonstrating the superiority of the proposed PFCA.

Comparing VPA with Other PET Methods. To further demonstrate the effectiveness of the proposed VPA, we compare it with other PET techniques on the SVT and TTR datasets, including CLIP-adapter [5], CoOP [52], and Extra VPE used in FDP [49]. As shown in Tab. 5, compared with the zero-shot EXCLIP baseline, PET on the STR task generally improves performance on both SVT and TTR. Among all methods, CoOP [52] brings the minimum number of parameters. However, due to the capacity and limited representation, context optimization falls behind other methods by an obvious margin, especially on the TTR, which explains the relatively poor performance of the prompt learning-based FDP method on the TTR dataset.

Table 5. Comparison of VPA variants and other PET methods based on CLIP-RN50 on the SVT and TTR datasets.

Method	Params	SVT	TTR
EXCLIP	-	72.81	41.87
CoOP [52]	**2048**	78.70	51.01
CLIP-adapter [5]	0.52M	77.88	56.01
Extra VPE [49]	0.53M	<u>79.35</u>	<u>64.21</u>
Textual adapter	0.20M	76.86	54.06
Visual adapter	0.20M	77.28	54.43
VPA (Ours)	0.20M	**81.98**	**66.54**

Compared with the Extra VPE learning used in FDP [49], VPA outperforms it by 2.63% and 2.33% with fewer parameters. Besides, the parameters of VPA are not dependent on the resolution, whereas the parameters of Extra VPE increase quadratically. At last, we introduce two variants of textual and visual adapters for comparison, which are inserted at the bottom of the textual and visual encoders. VPA outperforms both variants clearly on both SVT and TTR, demonstrating the superiority of VPA that can supply the visual position information to adapt to higher resolutions, which matters for scene text perception.

Table 6. Comparison of different reduction ratios of VPA based on the RN50 backbone on different datasets.

Reduction	4	8	16	32	64	128	256
Params	3.15M	1.58M	0.79M	0.40M	0.20M	0.10M	0.05M
SVT	**82.07**	80.88	80.83	81.12	<u>81.98</u>	81.44	80.69
TTR	66.76	67.56	**68.17**	<u>67.06</u>	66.54	66.34	66.98

Function of Components in VPA. As shown in Tab. 6, VPA works well with different reduction ratios. Considering the trade-off between parameters and performance, we set the reduction ratio to 64 for ResNet models. For ViT models, the reduction ratio is set to 8 to ensure the parameters are close. As shown in Tab. 7, without a residual connection, VPA degenerates into a direct MLP network, making it ineffective in the PET manner. With the constant scaling factor, the project can learn the channel-wise importance of the adapted features, thus improving the overall performance with few parameters.

Table 7. Comparison of different components in the proposed VPA on different datasets.

Method	Params	SVT	TTR
VPA	0.20M	**81.98**	**66.54**
w/o residual connection	0.20M	50.54	38.15
w/o scaling factor projection	0.13M	80.52	65.96

Table 8. Different modality of queries used in PFCA.

Text	Vision	SVT	TTR	IIIT-STR
-	-	81.98	66.54	81.79
✓	×	**85.84**	73.33	82.71
×	✓	72.80	61.03	80.12
✓	✓	85.05	**74.04**	**82.88**

Function of Modalities in PFCA. As reported in Tab. 8, we find that using text query only (refer to Eq. (4)) has already achieved remarkable performance. The reason lies in the gap in information density between the textual and visual modalities, which is also directly related to the STR task, in which the text modality acts as the query. Comparatively, the global representation of an image contains much noise for textual context as described in Eq. (5), leading to ineffective cross-modal interaction. However, the reranking performance on IIIT-STR, which contains large-scale background images without text, is somewhat good, demonstrating the function of query-by-vision features on distinguishing text/non-text images. As a result, we keep the symmetric cross-modal interaction for reranking.

5. Conclusion

In this paper, following the spirit of PET, we propose a novel strictly OCR-free method termed "CAYN" for efficient and accurate STR based on the CLIP models. A comprehensive study is conducted to explore the potential of CLIP for STR, proving that CLIP can serve as a strong baseline by interpolating position embeddings. We propose a visual position adapter to complement the visual encoder's positional information. A visual context dropout technique is introduced to generate locally aligned visual features. Besides, we propose a novel parameter-free cross-attention mechanism for cross-modal feature aggregation, which is used to perform image-text matching with a linear classifier. The experiments on three public benchmarks demonstrate that the CAYN can outperform existing methods and be efficient in training and inference.

Acknowledgement

This work is supported in part by National Key R&D Program of China (No.2022YFB3103800) and National Natural Science Foundation of China (No.U2336205). Also supported by the fund of the Laboratory for Advanced Computing and Intelligence Engineering and the Collaborative Education Project of the Ministry of Education: Construction of Cyberspace Security Experimental Teaching and Training Platform (2408131129).

References

[1] Xiang Bai, Mingkun Yang, Pengyuan Lyu, Yongchao Xu, and Jiebo Luo. Integrating scene text and visual appearance for fine-grained image classification. *IEEE Access*, 6:66322–66335, 2018. 1

[2] Michal Busta, Lukas Neumann, and Jiri Matas. Deep textspotter: An end-to-end trainable scene text localization and recognition framework. In *Proceedings of the IEEE international conference on computer vision*, pages 2204–2212, 2017. 2, 6

[3] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. 3

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 3

[5] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 3, 4, 8

[6] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*, 35:35959–35970, 2022. 4

[7] 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. 1, 2, 3, 4, 6

[8] 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, 2, 6

[9] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 3

[10] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 3, 4

[11] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7310–7311, 2017. 1

[12] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *International journal of computer vision*, 116:1–20, 2016. 1, 2, 6

[13] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 3

[14] Haojun Jiang, Jianke Zhang, Rui Huang, Chunjiang Ge, Zanlin Ni, Shiji Song, and Gao Huang. Cross-modal adapter for vision-language retrieval. *Pattern Recognition*, page 111144, 2024. 3

[15] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022. 3

[16] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023. 3

[17] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. 3

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 5

[19] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 2021. 3

[20] Minghui Liao, Guan Pang, Jing Huang, Tal Hassner, and Xiang Bai. Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 706–722. Springer, 2020. 1, 2, 6

[21] 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. 1, 2, 6

[22] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024. 4

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. 6

[24] Andrés Mafla, Ruben Tito, Sounak Dey, Lluís Gómez, Marçal Rusinol, Ernest Valveny, and Dimosthenis Karatzas. Real-time lexicon-free scene text retrieval. *Pattern Recognition*, 110:107656, 2021. 2, 3, 6

[25] 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. 1, 2, 6

[26] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, pages 1454–1459. IEEE, 2017. 2

[27] 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. 1

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

[29] Xugong Qin, Yu Zhou, Dongbao Yang, and Weiping Wang. Curved text detection in natural scene images with semi-and weakly-supervised learning. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 559–564. IEEE, 2019.

[30] Xugong Qin, Yu Zhou, Youhui Guo, Dayan Wu, Zhihong Tian, Ning Jiang, Hongbin Wang, and Weiping Wang. Mask is all you need: Rethinking mask r-cnn for dense and arbitrary-shaped scene text detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 414–423, 2021.

[31] Xugong Qin, Yu Zhou, Youhui Guo, Dayan Wu, and Weiping Wang. Fc^{2}rn: A fully convolutional corner refinement network for accurate multi-oriented scene text detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4350–4354. IEEE, 2021.

[32] Xugong Qin, Pengyuan Lyu, Chengquan Zhang, Yu Zhou, Kun Yao, Peng Zhang, Hailun Lin, and Weiping Wang. Towards robust real-time scene text detection: From semantic to instance representation learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2025–2034, 2023. 1

[33] 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, 3, 4, 5, 6

[34] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 3

[35] Xuejian Rong, Chucai Yi, and Yingli Tian. Unambiguous text localization, retrieval, and recognition for cluttered scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1638–1652, 2022. 1

[36] Hao Song, Hongzhen Wang, Shan Huang, Pei Xu, Shen Huang, and Qi Ju. Text siamese network for video textual keyframe detection. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 442–447. IEEE, 2019. 1

[37] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Proc. Int. Conf. Computer Vision (ICCV)*, 2019. 3

[38] Xunquan Tong, Pengwen Dai, Xugong Qin, Rui Wang, and Wenqi Ren. Granularity-aware single-point scene text spotting with sequential recurrence self-attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1

[39] 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. 1, 2, 3, 6

[40] 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. 2

[41] Lilong Wen, Yingrong Wang, Dongxiang Zhang, and Gang Chen. Visual matching is enough for scene text retrieval. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 447–455, 2023. 1, 2, 3, 6

[42] Xiao Yang, Dafang He, Wenyi Huang, Alexander Ororbia, Zihan Zhou, Daniel Kifer, and C Lee Giles. Smart library: Identifying books on library shelves using supervised deep learning for scene text reading. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–4. IEEE, 2017. 1

[43] Xu-Cheng Yin, Xuwang Yin, Kaizhu Huang, and Hong-Wei Hao. Accurate and robust text detection: A step-in for text retrieval in natural scene images. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 1091–1092, 2013. 1

[44] Wenwen Yu, Yuliang Liu, Wei Hua, Deqiang Jiang, Bo Ren, and Xiang Bai. Turning a clip model into a scene text detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6978–6988, 2023. 2

[45] Wenwen Yu, Yuliang Liu, Xingkui Zhu, Haoyu Cao, Xing Sun, and Xiang Bai. Turning a clip model into a scene text

spotter. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[46] Yuechen Yu, Yulin Li, Chengquan Zhang, Xiaoqiang Zhang, Zengyuan Guo, Xiameng Qin, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. Structextv2: Masked visual-textual prediction for document image pre-training. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[47] Gangyan Zeng, Yuan Zhang, Yu Zhou, and Xiaomeng Yang. Beyond ocr+ vqa: Involving ocr into the flow for robust and accurate textvqa. In *Proceedings of the 29th ACM international conference on multimedia*, pages 376–385, 2021. 1

[48] Gangyan Zeng, Yuan Zhang, Yu Zhou, Xiaomeng Yang, Ning Jiang, Guoqing Zhao, Weiping Wang, and Xu-Cheng Yin. Beyond ocr+ vqa: Towards end-to-end reading and reasoning for robust and accurate textvqa. *Pattern Recognition*, 138:109337, 2023. 1

[49] Gangyan Zeng, Yuan Zhang, Jin Wei, Dongbao Yang, Peng Zhang, Yiwen Gao, Xugong Qin, and Yu Zhou. Focus, distinguish, and prompt: Unleashing clip for efficient and flexible scene text retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2525–2534, 2024. 1, 2, 3, 4, 6, 7, 8

[50] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 5

[51] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 3

[52] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3, 8