JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Huazhong University of Science & Technology

Turning a CLIP Model into a Scene Text Detector

Wenwen Yu · Yuliang Liu · Wei Hua · Deqiang Jiang · Bo Ren · Xiang Bai

Speaker: Wenwen Yu

June 20, 2023

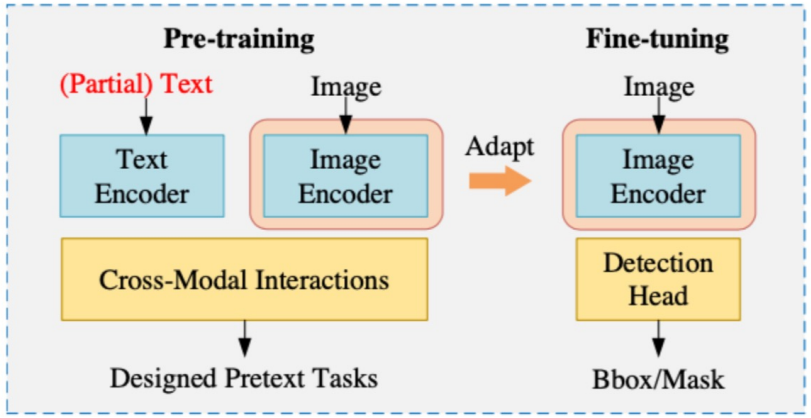
TUE-PM-272

Outline

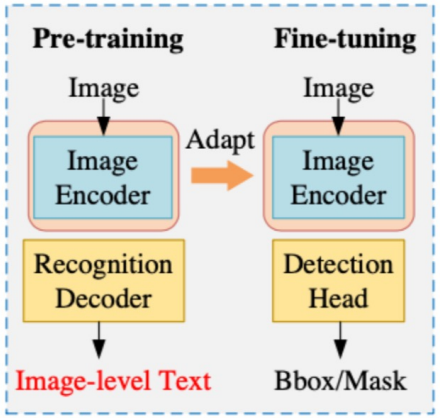
- Outline
  - Introduction
  - Method
  - Experiments
  - Conclusion

Introduction

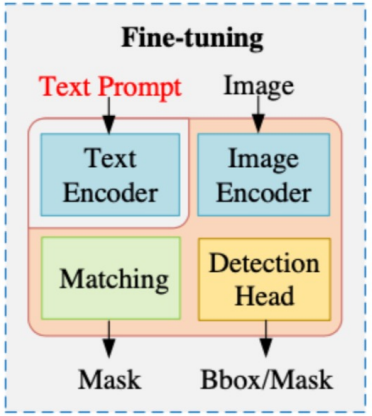
- Dfferent from general image, text and image naturally contain information in two modalities: visual and textual.
- It is a natural approach to apply the visual language pre-training model CLIP to text detection methods
- The existing methods that utilize visual language models generally follow a two-stage approach, which involves pre-training followed by fine-tuning.



(a) Two stage: design cross-modal interations pretext task



(b) Two stage: design text recognitionpretext task



(c) One stage: Ours

We propose a one-stage approach that directly integrates CLIP into the text detector, eliminating the need for designing pre-training tasks.

[1] Yu et al. Turning a CLIP Model into a Scene Text Detector. CVPR 2023.

Method

CLIP(Contrastive Language-Image Pretraining**)** [1]**:**

- ❑ CLIP is pretrained on a dataset of 400 million image-text pairs through contrastive learning.
- ❑ CLIP has the ability to understand visual concepts in open-ended scenes.

"heart" "summer" "self+relief" "child's drawing"

CLIP can associate relevant images based on textual content.

[1] Radford et al. Learning transferable visual models from natural language supervision. ICML 2021
[2] Goh et al. Multimodal neurons in artificial neural networks. Distill, 6(3):e30, 2021

Method

Solution：

Designing a pluggable module named TCM (**T**urning a **C**LIP **M**ode) to extract language knowledge from CLIP and guide the text detector.

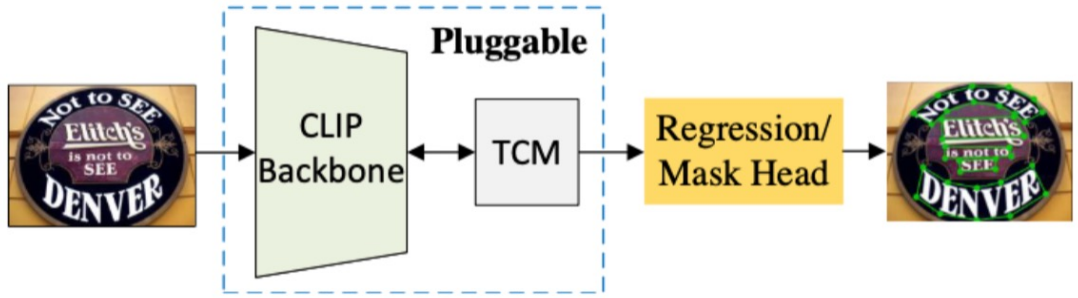
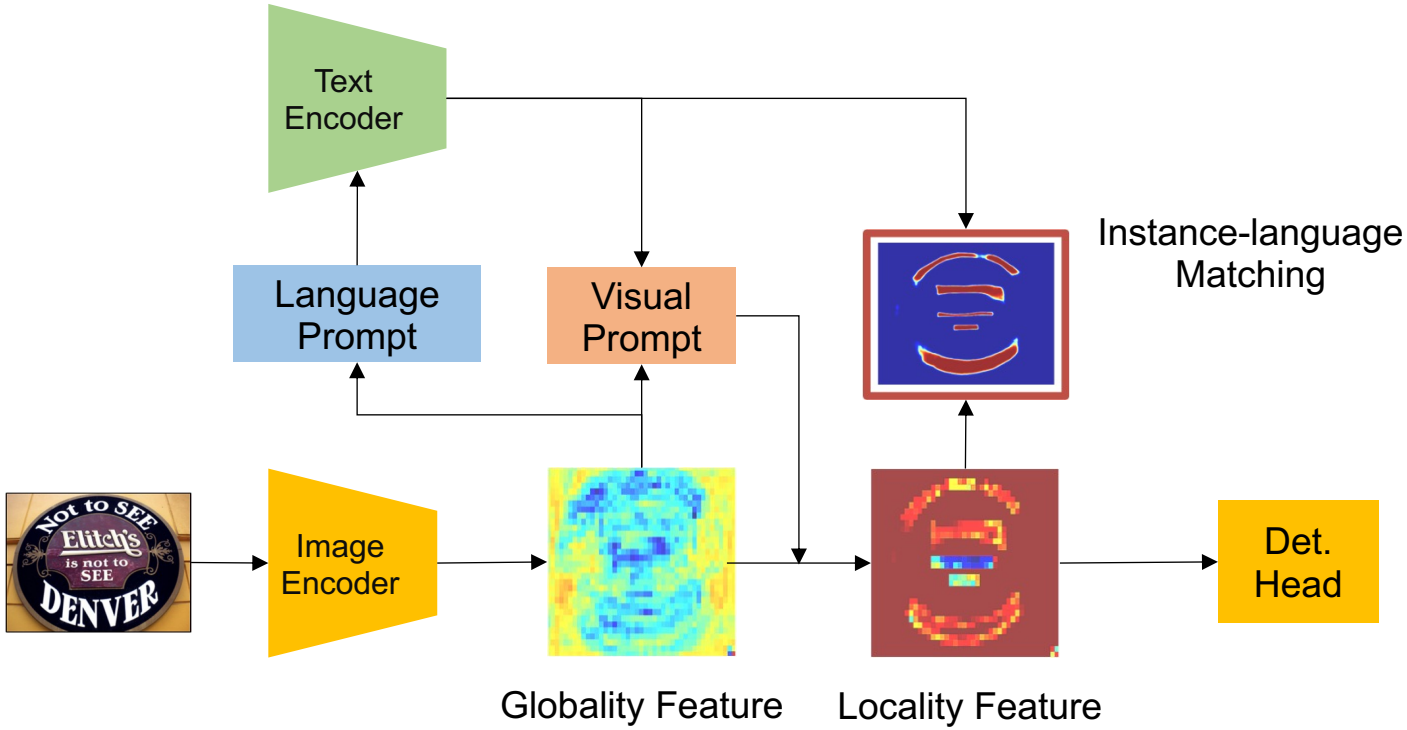


Figure 2. The overall framework of our approach.

Goals achieved:

- Designing an effective method to efficiently extract language knowledge from CLIP.
- Addressing the issue of inconsistent granularity between visual and language modal feature.

Method



The overall of pluggable TCM

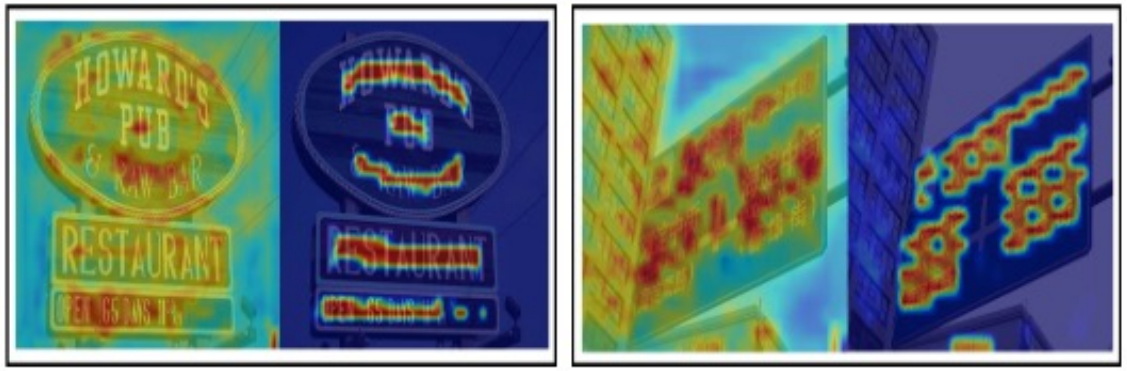
- Designed Language Prompt to extract textual knowledge from CLIP and utilize it to guide the text detector.
- Designed Visual Prompt to obtain fine-grained visual features.
- Designed Instance-language Matching to achieve feature alignment between language knowledge and multiple text instances.

Experiments

- Applying the TCM module to both segmentation-based and regression-based text detectors.
- TCM can improve the performance of text detectors.
- Visual Prompts can associate with text regions.

	Method	IC15		TD		CTW		FPS
		F	Δ	F	Δ	F	Δ	
Reg.	FCENet [60]	86.2	-	85.4†	-	85.5	-	11.5
	TCM-FCENet	**87.1**	**+0.9**	**86.9**	**+1.5**	**85.9**	**+0.4**	8.4
Seg.	PAN [39]	82.9	-	84.1	-	83.7	-	36
	TCM-PAN	**84.6**	**+1.7**	**85.3**	**+1.2**	**84.3**	**+0.6**	18
	DBNet [18]	87.3	-	84.9	-	83.4	-	14.5
	TCM-DBNet	**89.2**	**+1.9**	**88.8**	**+3.9**	**84.9**	**+1.5**	10

Table 1. Text detection results of cooperating with existing methods on IC15, TD, and CTW. † indicates the results from [52]. Reg. and Seg. short for regression and segmentation methods, respectively. FPS are reported with ResNet50 backbone on a single V100.

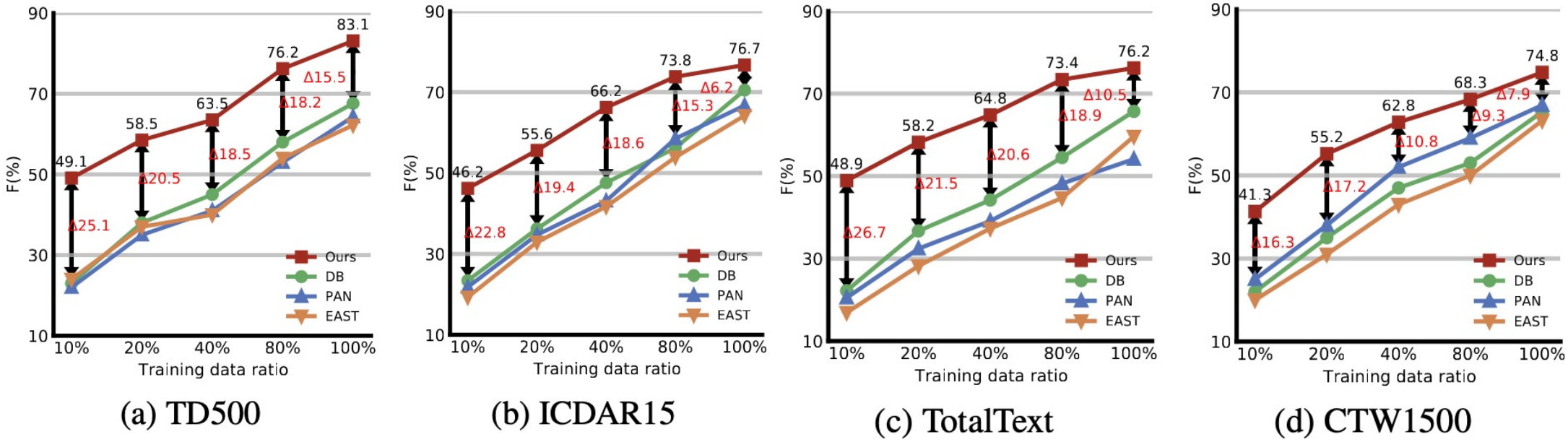


(a) CTW1500 (b) MSRA-TD500

Figure 7. Visualization results of our method. For each pair, the left is the image embedding $\boldsymbol{I}$, and the right is the generated visual prompt $\tilde{\boldsymbol{I}}$. Best view in screen. More results can be found in appendix.

Experiments

- TCM can enhance the few-shot learning capability of text detectors.



(a) TD500 (b) ICDAR15 (c) TotalText (d) CTW1500

Figure 5. Few-shot training ability with varying training data ratio. "F" represents F-measure.

Experiments

- Conducting experiments to test the model transferability across different datasets.
- TCM can improve the generalization performance of text detectors.

Method	ST → IC13	ST → IC15
EAST† [58]	67.1	60.5
PAN [39]	-	54.8
CCN [44]	-	65.1
ST3D [16]	73.8	67.6
DBNet [18]	71.7	64.0
TCM-DBNet	**79.6**	**76.7**

Table 2. Synthtext-to-real adaptation. † indicates the results from [42]. ST indicates SynthText. F-measure (%) is reported.

Method	IC13 → IC15	IC13 → TD
EAST† [58]	53.3	46.8
GD(AD) [52]	64.4	58.5
GD(10-AD) [52]	69.4	62.1
CycleGAN [59]	57.2	-
ST-GAN [19]	57.6	-
CycleGAN+ST-GAN	60.8	-
TST [42]	52.4	-
DBNet [18]	63.9	53.8
TCM-DBNet	**71.9**	**65.1**

Table 3. Real-to-real adaptation. † indicates that the results are from [52]. Note that the proposed method outperforms other methods. F-measure (%) is reported.

Experiments

- Comparing with two-stage pretraining methods based on visual language models.
- Overall, TCM outperforms two-stage pretraining methods in terms of performance.

	Methods	Pretext task	IC15	TT	TD	CTW
Convention	SegLink [29]	×	-	-	77.0	-
	PSENet-1s [14]	×	85.7	80.9	-	82.2
	LOMO [53]	×	87.2	81.6	-	78.4
	MOST [8]	×	88.2	-	86.4	-
	Tang *et al.* [33]	×	89.1	-	88.1	-
VLP	DB+ST†	×	85.4	84.7	84.9	-
	DB+STKM† [37]	✓	86.1	85.5	85.9	-
	DB+VLPT† [31]	✓	86.5	86.3	88.5	-
	DB+oCLIP* [48]	✓	-	-	-	84.4
	DB+TCM(Ours)	×	**89.4**	85.9	**88.8**	**85.1**

Table 4. Comparison with existing scene text pretraining techniques on DBNet (DB). † indicates the results from [31]. ST and VLP denote SynthText pretraining and visual-language pretraining methods, respectively. * stand for our reimplementation results. F-measure (%) is reported.

Experiments

- Ablation study: Investigating the impact of individual modules in TCM on the model's performance.
- Demonstrate the effectiveness of the designed modules on the model.

Method	PP	LP	LG	VG	IC15	TD	TT	CTW
					F	F	F	F
BSL	×	×	×	×	87.7	86.8	84.7	83.4
BSL+	✓	×	×	×	87.75	87.0	84.74	83.5
BSL+	✓	4	×	×	88.0	87.1	84.8	83.6
BSL+	×	4	×	×	87.8	87.7	85.1	83.9
BSL+	×	18	×	×	88.1	87.8	85.3	83.9
BSL+	×	32	×	×	88.4	88.2	85.4	84.5
BSL+	✓	4	✓	×	88.6	88.4	85.5	84.6
TCM	✓	4	✓	✓	89.2	88.8	85.6	84.9
TCM	✓	32	✓	✓	**89.4**	**88.8**	**85.9**	**85.1**
Δ					+1.7	+2.0	+1.2	+1.7

Table 6. Ablation study of our proposed components on IC15, TD, TT and CTW. "BSL", "PP", "LP", "LG", and "VG" represent the baseline method DBNet, the predefined prompt, the learnable prompt, the language prompt generator, and the visual prompt generator, respectively. F (%) represents F-measure. Δ represents the variance.

Experiments

- Investigating the impact of the Image Encoder and Text Encoder on performance.
- The involvement of the Text Encoder during training may disrupt the intrinsic language knowledge of the pretrained language model.

	Image encoder	Text encoder	F (%)
LR Factor	0.1	0.0	**88.7**
	0.1	0.1	87.8
	0.1	1.0	87.1
	1.0	1.0	86.3

Table 8. Ablation study of exploration on image encoder and text encoder. "LR" represents the learning rate.

Experiments

- Ablation study: Evaluating the performance when increasing the training dataset size and simultaneously increasing the difficulty of the test set.
- Training set: Merged public datasets with a total of 13,784 images.
- Test set: Collected dataset specifically focused on nighttime scenes.
- When increasing the training dataset size, TCM still demonstrates advantages over other methods in complex scenes.

Method	Training Data	Testing Data	F (%)
FCENet	Joint data	NightTime-ArT	55.2
DBNet	Joint data	NightTime-ArT	52.8
TCM-DBNet	Joint data	NightTime-ArT	**70.2**

Table 9. Ablation study of exploration on large amounts of training data. Joint data including IC13, IC15, TD, CTW, TT, and MLT17, with total 13,784 image, and testing on a NightTime-ArT data (326 images)

Figure 6. The examples of our constructed NightTime-ArT.

Experiments

Ablation study

- The impact of model parameter size on performance.
- When increasing the parameter size of the model, TCM still demonstrates advantages over other methods.

Method	Backbone	Params	FLOPs	F (%)
DBNet	R50	26 (M)	98 (G)	84.9
DBNet	R101	46 (M)	139 (G)	85.9
DBNet	R152	62 (M)	180 (G)	87.3
TCM-DBNet	R50	50 (M)	156 (G)	**88.7**

Table 10. Ablation study of the parameters comparison with DB-Net.

Experiments

Failure Cases

- Regions that are similar to text features may lead to false detections.

Figure 8. Failure cases. Red circle means false positive region.

Conclusion

- The paper introduces a pluggable module called TCM, which does not require designing pre-training tasks and can align visual and textual information features.
- The TCM framework can enhance existing scene text detectors.
- The TCM framework can improve the few-shot training capability of the detector.
- The TCM framework can enhance the model's generalization ability.
- The paper collected a nighttime scene dataset called NightTime-ArT.

Thanks

wenwenyu@hust.edu.cn

github.com/wenwenyu/TCM