Supplemental Material to
SCOB: Universal Text Understanding via Character-wise Supervised Contrastive Learning with Online Text Rendering for Bridging Domain Gap

Content

This supplementary material provides details of pre-training experiments (Section 1), detailed settings of fine-tuning (Section 2), and more diverse samples of t-SNE [36] (Section 3).

1. Pre-trainings

1.1. Details of Pre-training

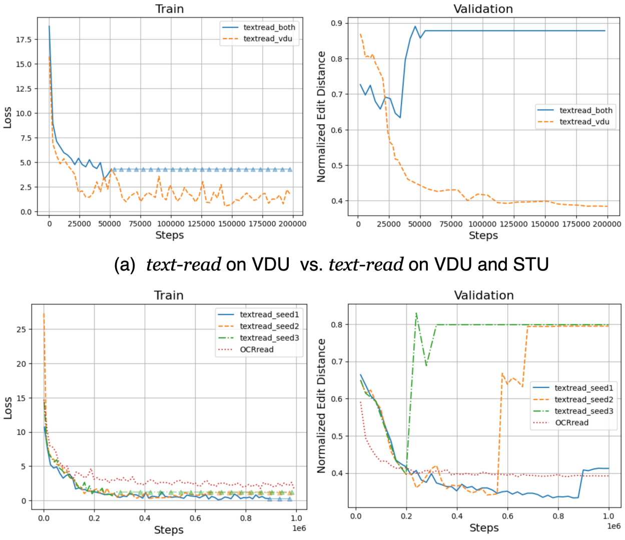
In Pix2Seq [4], sequence augmentation was applied to add a noise token to the sequence. On the other hand, we do not apply sequence augmentation to our model because we could not empirically observe the performance improvement with noise tokens. The learning rate is set to 1e-4 with a cosine decay scheduler [24] where a warm-up step is 100K steps. For the data augmentation, we adopt the widely used augmentation methods for OCR as follows: random crop following CRAFT [1], random rotation, random resize options (e.g., bilinear, nearest neighbor, bicubic, and Lanczos interpolation), and photometric distortion.

The detailed settings of the online text renderer are as follows. The range of resolution is from 400 to 768, the range of background RGB values is from 151 to 255, and the Gaussian blur radius value is 0 to 1.8 with a probability of 0.2. Figure 3 provides image samples generated in a one-to-one correspondence to the ICDAR2015 [13] test set using the renderer.

1.2. Instability of *Text-read*

In this section, we present experimental results about the instability of *text-read* when the model is pre-trained with both VDU and STU data. As significant performance differences are observed in the early stages of training, we conduct a precise experiment. Specifically, we pre-train the model using IIT-CDIP and scene text datasets for 200K steps with a batch size of 16. For fine-tuning OCR, we employ scene text datasets used in the main paper and fine-tune the models for 50K training steps with 8 batch size that consists of 1920 × 1920 resolution images.

Figure 1 shows the convergence of two distinct reading approaches: *text-read* and *OCR-read*. Specifically, Figure 1 (a) shows that the loss and validation score converge stably when only VDU data is used as a training set. However, the model fails to generate appropriate sequences during validation and experiences NaN (not a number) loss when both VDU and STU datasets are used for training. Figure 1 (b) depicts the *text-read* pre-training process for 1M steps, which has failed across various seed values. On the other hand, *OCR-read* displays more stable convergence patterns despite its loss and validation score converging at a slightly higher level due to sensitivity issues such as coordinate token output. The sole difference between *OCR-read* and *text-read* lies in the presence or absence of coordinate prediction. Also, the distinction between VDU and STU domains is based on whether text is present in the natural scene back-



(a) *text-read* on VDU vs. *text-read* on VDU and STU

(b) *text-read* on VDU and STU for 1M steps

Figure 1. Graphs of training loss and validation score during *text-read* pre-training. (a) shows instability of *text-read* using VDU and STU together. (b) also shows instability of *text-read* using VDU and STU together for 1M steps compared to *OCR-read*. Validation score is normalized edit distance. Upper triangle in loss graph means NaN (not a number).

Method	KIE	Document VQA	OCR			Scene Text VQA	
	CORD [31]	DocVQA [27]	IC13 [14]	IC15 [13]	TotalText [6]	TextVQA [33]	ST-VQA [2]
Text-read on both domains	80.7	45.7	66.7	20.3	43.4	42.1	52.2
Text-read on VDU	85.3	53.1	77.1	35.8	56.1	42.8	53.2

Table 1. The performance evaluation for *text-read* according to pre-training domains. We report the end-to-end recognition F-measure scores on IC13, IC15, and TotalText evaluated with strong, strong, and full lexicons, respectively

ground. We posit that a domain gap exists between VDU and STU, contributing to the instability of *text-read*.

To evaluate the performance of the pre-trained weights on downstream tasks, we fine-tuned *text-read* on four downstream tasks using pre-trained weights for 200K steps ((a) of Figure 1). Table 1 indicates that *text-read* pre-training on both domains fails to transfer learning on downstream tasks. The performance of *text-read* using both domains is relatively lower than that of *text-read* using only the VDU domain. This result further supports that *text-read* is unstable in both domains, deteriorating the achievement of a reliable pre-training.

1.3. Details of Ablation Study

As presented in Table 4 of the main paper, we conduct an ablation study with five models. All models are pre-trained for 1M steps. Table 2 provides detailed performance of benchmarks. Details are as follows:

- A: $\text{W}_{\text{OCR-read}}$.
- B: We introduce augmentation for generating multi-view images. The augmentation types are random rotation, image resize, color jittering, converting into the gray channel, and Gaussian blur. Geometric augmentations such as random rotation and image resize are applied with low intensity. The model is trained via Eq. 1 and Eq. 2.
- C: The real and rendered data are the input data of the model. The model is trained via Eq. 3, and the renderer is the same as SCOB.
- D: $\text{W}_{\text{OCR-read}}$ with SCOB.
- E: The model is trained under the fully supervised learning. Thus, the coordinate information of real data is employed for pre-training.

We fine-tune 4 downstream tasks such as KIE, document VQA, OCR, and scene text VQA. For the tasks of VQA, we evaluate the validation set, because multiple submissions of the test dataset on the leaderboard can be regarded as cheating. We fine-tuned models for 50K steps using 16 batch size on VQAs for efficiency. For the OCR tasks, we employ intermediate fine-tuned models, which are evaluated on the test dataset.

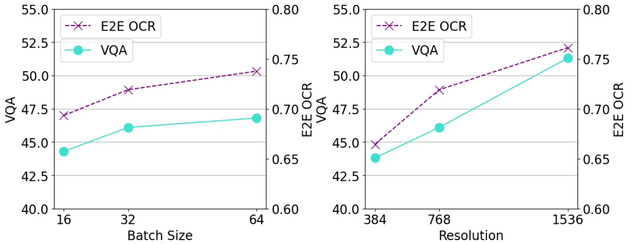


Figure 2. The results for VQA (average scores of DocVQA, TextVQA, STVQA) and E2E OCR (average F1 scores of IC13, IC15, TotalText) across batch sizes (16, 32, 64) and resolutions (384, 768, 1536). A fixed resolution of 768 and batch size of 32 were used respectively. The experiment involved 200K steps of *OCR-read* pre-training and 50K of fine-tuning.

1.4. Performance Trends with Batch Size and Resolution

We present the experimental results of the effect of batch size and resolution changes on downstream task performance under the *OCR-read* pre-training setting in Figure 2. The results confirm an enhancement in performance as both batch size and resolution increase. We identified a performance gap between SCOB and the current SoTA. However, in this paper, we delve into the composite domain of VDU and STU to compare the performance of pre-training tasks, verifying the superiority of SCOB under identical pre-training environments. Based on these results, we expect that there is latent potential within the model, likely to be uncovered through the careful tuning of factors such as scale, learning schedule, and resolution.

2. Fine-tunings

2.1. Details of Fine-tuning

In the fine-tuning stage, we use 1920×1920 image resolution to recognize text well, and intermediate training follows it. Moreover, we empirically confirmed that auxiliary loss [3] accelerates the convergence of the recognition performance of OCR. Thus we apply the auxiliary loss to all layers of the decoder in the fine-tuning.

Table 3 provides our detailed settings for fine-tuning. We fine-tune all models using Adam optimizer with a cosine decay scheduler where the warm-up step is set to 10% of training steps. We have not widely explored the hyperparameters, such as the batch ratio of the dataset, gradient clipping value, and learning rate. We believe more hyperparameter

Method	KIE	Document VQA	OCR			Scene Text VQA	
	CORD [31]	DocVQA [27]	IC13 [14]	IC15 [13]	TotalText [6]	TextVQA [33]	ST-VQA [2]
A. $\text{W}_{\text{OCR-read}}$	88.2	55.1	94.3	74.2	75.5	55.4	59.2
B. A w/ SupCon	87.7	50.0	93.4	76.5	**76.7**	53.0	60.6
C. A w/ rendering	88.0	47.8	94.1	76.4	75.4	50.8	57.7
D. A w/ SCOB	**88.5**	**55.5**	**95.0**	77.6	76.5	**56.2**	62.6
E. D w/ full annotation	86.8	55.1	94.7	**77.7**	75.5	56.1	**63.1**

Table 2. Ablation study on the proposed components. For the evaluation of OCR, we report the end-to-end recognition F-measure scores on IC13, IC15, and TotalText evaluated with strong, strong, and full lexicons, respectively.

Task	Steps	Batch Size	LR	GC	DPL	MSL	Resolution	Batch Ratio of Fine-tuning Dataset
Table Reconstruction*	400K	16	5e-5	1.0	6	3072	768	**1.0** PubTabNet [39]
KIE	200K	16	3e-5	1.0	0	512	1920	**1.0** CORD [31]
Document Classification	1M	16	2e-5	1.0	0	8	1920	**1.0** RVL-CDIP [10]
Document VQA	100K	16	3e-5	0.25	0	512	1920	**0.65** DocVQA [27], **0.07** TextVQA [33] , **0.07** ST-VQA [2], **0.07** OCRVQA [28], **0.07** VizWiz [9], **0.07** VQAv2 [7]
Infographics VQA	100K	16	3e-5	0.25	0	512	1920	**0.58** InfoVQA [26], **0.07** DocVQA [27], **0.07** TextVQA [33], **0.07** ST-VQA [2], **0.07** OCRVQA [28], **0.07** VizWiz [9], **0.07** VQAv2 [7]
Layout Analysis	100K	16	2e-4	1.0	0	512	1536	**1.0** PubLayNet [40]
Scene Text VQA	100K	16	3e-5	0.25	0	512	1920	**0.2** DocVQA [27], **0.2** TextVQA [33], **0.2** ST-VQA [2], **0.2** OCRVQA [28], **0.1** VizWiz [9], **0.1** VQAv2 [7]
Scene Text OCR*	100K	16	5e-5	1.0	0	512	2560	**0.3** HireText [23] , **0.3** TextOCR [34], **0.09** TotalText [6], **0.09** OpenImagesv6 [19], **0.01** IC13 [14], **0.01** IC15 [13]

Table 3. Detailed settings of downstream tasks. For the batch ratio, we represent the cell as '***batch ratio of dataset*** *dataset*'. *Abbr. LR*: learning rate, *GC*: gradient clipping, *DPL*: decoder prune layer, *MSL*: maximum sequence length of decoder. * denotes that not using intermediate training.

exploration will result in better performance. For the input image resolution, we determine the resolution considering the original image size and comparisons of the input size.

Here, we supplement more details of specific downstream tasks. To help understand downstream tasks, we provide the examples of image and its ground truth sequence of our model in Figures 4 and 5.

Table Reconstruction. In table reconstruction, the model should decode table structures and contents in the cell. Thus, the maximum sequence length of the decoder is set to 3072, which makes pruning 6 Transformer layers in the decoder to reserve batch size 16. In table reconstruction, the input can be categorized as image and PDF.

Layout Analysis. PubLayNet [40] is a dataset annotated with a bounding box format and 5 document layout categories: text, title, list, figure, and table. This includes 335,703 training images and 11,245 validation images. We fine-tune the train set and evaluate the validation set, following LayoutLMv3 [12].

Scene Text OCR. To evaluate the performance of our OCR model, we assess its ability to recognize text in different benchmarks, including ICDAR2013 [14], ICDAR2015 [13], and TotalText [6]. However, since each dataset has a distinct format for coordinate annotation, we conduct short annotation fine-tuning specifically for IC15 and TotalText to mitigate any discrepancies. Table 5 presents the performance of our model on IC13, IC15, and TotalText with additional lexicon conditions. Note that the OCR performance reported in Table 2 is obtained before annotation fine-tuning.

2.2. Comparison with More Models

Our methodologies are compared with more recent approaches such as UDOP [35], PaLI [5], and GIT2 [37] as presented in Table 4. UDOP and PaLI-17B exhibit superior performance across various benchmark criteria. These results show the potential of leveraging OCR to enhance the performance of generative models, albeit with concomitant resource implications. Based on PaLI-3B and PaLI-17B results, increasing the model size can also be a significant factor in the performance. Furthermore, it is noteworthy that PaLI employs a form of *text-read* task as a pre-training method. In this context, we believe that the integration of our SCOB approach holds the potential to further amplify the performance of PaLI.

2.3. Comparison on Layout Analysis

Table 7 shows the performance of PubLayNet [40]. Our models demonstrate comparable performance to previous methods. Importantly, our models can handle multiple downstream tasks using a single pipeline, namely the

Method	#GPUs	Universal Text Understanding Downstream Tasks										
		Table Reconstruction	KIE	Document Classification	Document VQA		Layout Analysis	Scene Text OCR			Scene Text VQA	
		PubTabNet [39]	CORD [31]	RVL-CDIP [10]	DocVQA [27]	InfoVQA [26]	PubLayNet [40]	IC13 [14]	IC15 [13]	TotalText [6]	TextVQA [33]	ST-VQA [2]
W$_{\text{OCR-read}}$	8×V100	96.0	88.2	94.2	56.1	22.7	93.8	95.8	89.6	84.9	55.4	**62.9**
W$_{\text{OCR-read}}$ w/ SCOB	8×V100	95.9 (-0.1)	88.5 (+0.3)	94.6 (+0.4)	60.2 (+4.1)	**28.5** (+5.8)	93.9 (+0.1)	**96.6** (+0.8)	**90.9** (+1.3)	**86.0** (+1.1)	**56.2** (+0.8)	62.6 (-0.3)
W$_{\text{text-read}}$	8×V100	**96.2**	85.5	94.4	57.0	25.2	93.6	96.0	87.2	83.7	49.3	57.2
W$_{\text{text-read}}$ w/ SCOB	8×V100	96.0 (-0.2)	87.4 (+1.9)	94.3 (-0.1)	59.6 (+2.6)	27.5 (+2.3)	93.9 (+0.3)	96.0 (+0.0)	90.2 (+3.0)	85.3 (+1.6)	54.4 (+5.1)	61.2 (+4.0)
TableFormer [29]	n/a	93.7	-	-	-	-	-	-	-	-	-	-
Donut$_{\text{proto}}$ [16]	8×V100	-	85.4	94.5	47.1	10.2*	-	-	-	-	-	-
Donut [16]	64×A100	-	**90.9**	95.3	67.5	24.4*	-	-	-	-	36.8*	61.5*
LayoutLMv3 [12]	32×V100	-	84.4*	**95.5**	**83.4**	-	**95.1**	-	-	-	-	-
SPTS [32]	32×V100	-	-	-	-	-	-	93.3	77.5	82.4	-	-
PreSTU [15]	n/a	-	-	-	-	-	-	-	-	-	54.5	62.6
UDOP [35]	n/a	-	-	**96.0**	**84.7**	**47.4**	-	-	-	-	-	
PaLI-3B [5]	n/a	-	-	-	-	-	-	-	-	-	60.1	67.5
PaLI-17B [5]	1024×TPUv4	-	-	-	-	-	-	-	-	-	**71.8**	**77.1**
GIT2 [37]	n/a	-	-	-	-	-	-	-	-	-	68.4	75.1

Table 4. Comparison including additional SoTA models: UDOP [35], PaLI [5], and GIT2 [37]. Note that UDOP and PaLIs employ the result of OCR as an input. The total number of training parameter of UDOP, PaLI-3B, PaLI-17B and GIT2 are 794M, 3B, 17B and 5.1B, respectively. The table shows the extensive benchmarks for text-related downstream tasks. "#GPUs" denotes the total number of employed GPUs for pre-training. The left section is VDU tasks, and the right section is STU tasks. The best performance is represented in **bold**. Note that Donut was pre-trained on IIT-CDIP and SynthDoG, while Donut$_{\text{proto}}$ was pre-trained on SynthDoG [16]. $*$ denotes the performance results of our fine-tuning, conducted following the author's guidelines.

Method	IC13 End-to-End				IC15 End-to-End				Total-Text	
	N	*G*	*W*	*S*	*N*	*G*	*W*	*S*	*N*	*F*
W$_{\text{OCR-read}}$	90.4	91.8	93.7	94.3	66.4	69.4	73.3	75.7	72.6	78.9
W$_{\text{OCR-read}}$ w/ SCOB	**92.2**	**93.1**	**94.6**	**95.0**	**71.9**	**74.2**	**76.6**	**78.8**	**75.1**	**79.8**
W$_{\text{text-read}}$	87.7	89.2	91.1	91.8	55.6	58.7	61.3	63.3	69.1	75.1
W$_{\text{text-read}}$ w/ SCOB	88.6	90.0	91.7	92.1	59.3	61.1	63.1	65.5	73.1	78.2

Table 5. The end-to-end recognition results on ICDAR2013, ICDAR2015, and Total-Text. *Abbr.* *N, G, W, S, F*: none, generic, weak, strong, and full lexicons, respectively.

Method	*G*	IC13 [14]		IC15 [13]		TotalText [6]	
		N	*S*	*N*	*S*	*N*	*F*
Box-based Localization							
MTSv2 [25]		80.3	93.3	-	83.0	65.3	77.4
MTSv3 [21]		80.2	-	-	83.3	71.2	78.4
DEER [17]		-	-	71.7	82.7	74.8	81.3
SwinTextSpotter [11]		-	-	-	83.9	51.8	77.0
TESTR [38]		-	-	65.3	85.2	73.3	83.9
TTS [18]		-	-	-	**85.2**	**75.6**	**84.4**
W$_{\text{OCR-read}}$	✓	90.4	94.3	66.4	75.7	72.6	78.9
W$_{\text{OCR-read}}$ w/ SCOB	✓	**92.2**	**95.0**	**71.9**	78.8	75.1	79.8
W$_{\text{text-read}}$	✓	87.7	91.8	55.6	63.3	69.1	75.1
W$_{\text{text-read}}$ w/ SCOB	✓	88.6	92.1	59.3	65.5	73.1	78.2
Point-based Localization							
SPTS [32]	✓	-	93.3	-	79.5	74.2	82.4
W$_{\text{OCR-read}}$ w/ SCOB	✓	-	**96.6**	-	**90.9**	**78.9**	**86.0**

Table 6. The end-to-end recognition F-measure results on ICDAR2013 [14], ICDAR2015 [13] and TotalText [6]. *Abbr.* *G*: generation model, *N, S, F*: none, strong and full lexicon, respectively.

sequence generation framework, which distinguishes them from previous methods that require specific architectural designs for each task.

2.4. Comparison on OCR

Table 5 provides the F-measure scores for the end-to-end OCR benchmark on scene text. Our model achieves state-of-the-art performance in IC13 and shows comparable results in IC15 and Total-Text compared to the alternative state-of-the-art methods.

2.4.1 W$_{\text{OCR-read}}$ vs. SPTS

Our W$_{\text{OCR-read}}$ architecture is similar to SPTS [32], but outperforms it significantly. Notably, W$_{\text{OCR-read}}$ and SPTS differ in parameters (202M vs. 36M), dataset, encoder backbone, and resolution. SPTS pre-trains on Curved Synthetic Dataset 150K [22], MLT-2017 [30], ICDAR2013 [14], ICDAR2015 [13], and TotalText [6], followed by fine-tuning on each target dataset. In contrast, W$_{\text{OCR-read}}$ generalizes not only to OCR but also to diverse downstream tasks, using various datasets of VDU and STU. While SPTS employs ResNet-50 and Transformer 6 layers as its encoder backbone, we use the Swin-transformer. Moreover, SPTS adopts 1600 resolution, while W$_{\text{OCR-read}}$ uses 768 resolution in pre-training and 2560 resolution in fine-tuning. Notably, the resolution has a significant impact on OCR performance. We also accelerate end-to-end recognition convergence using the auxiliary loss [3]. Unlike SPTS, we do not use word

Method	#Param	G	Category					mAP
			Text	Title	List	Table	Figure	
PubLayNet [40]	-		91.6	84.0	88.6	96.0	94.9	91.0
LayoutLMv3 [12]	133M		94.5	90.6	95.5	97.9	97.0	**95.1**
UDoc [8]	272M		93.9	88.5	93.7	97.3	96.4	93.9
DiT [20]	304M		94.4	89.3	96.0	97.8	97.2	94.9
$W_{\text{OCR-read}}$	202M	✓	93.1	88.4	92.9	97.3	97.1	93.8
$W_{\text{OCR-read}}$ w/ SCOB	202M	✓	93.2	88.8	93.7	97.0	96.6	93.9
$W_{\text{text-read}}$	202M	✓	93.2	88.8	92.4	97.1	96.7	93.6
$W_{\text{text-read}}$ w/ SCOB	202M	✓	93.5	89.1	93.0	96.9	96.8	93.9

Table 7. The public benchmark on PubLayNet [40] validation set (mAP @ IOU [0.50:0.95]) for document layout analysis. *Abbr. G*: generation model.

Models	TEDS	Model Params.	OCR Params.	Time (s/img)
LayoutLMv3 + EasyOCR	56.2	133M	25M	0.7
LayoutLMv3 + PaddleOCR	60.5	133M	12M	0.3
LayoutLMv3 + MSAzure	84.4	133M	n/a	1.8
$W_{\text{OCR-read}}$	88.2	202M	None	1.1

Table 8. The performance and inference time on CORD testset. LayoutLMv3 can be combined with various OCR models.

instance padding, allowing our model to learn many more words with the same decoder sequence length by saving redundant decoding sequences. While word instance padding may help the model converge at the early training stages, it does not significantly improve final performance.

To compare two methods that have different outputs for coordinate information, the central point of the box of our model is taken as a single point, and we follow the SPTS evaluation protocol. Figures 6-8 also present the qualitative comparisons: $W_{\text{OCR-read}}$, $W_{\text{OCR-read}}$ with SCOB, and SPTS. As can be seen, W is more robust to small or dense words. As shown in Figure 8, W also predicts well for curved text.

2.5. Comparison with LayoutLMv3

As shown in Table 8, we present the performance of LayoutLMv3 based on different OCR models. When employing lightweight OCR models like EasyOCR[*] and PaddleOCR[†], LayoutLMv3 demonstrates improved speed compared to our model. However, it is important to note that the scores obtained with these lightweight OCR models are significantly inferior when compared to the utilization of a commercial OCR model like MSAzure[‡].

3. Visualization of t-SNE

Figures 9 and 10 provide more diverse results visualized by t-SNE [36] with various perplexities. To color-code points according to ground truth classes in sequence generation architecture, we use a teacher-forcing scheme. Note that Figures 9 and 10 use pre-trained and fine-tuned models, respectively.

As shown in the figures, our visualizations show similar trends regardless of perplexity values, but visualization for pre-trained models have a different tendency with fine-tuned models. The ICDAR2015 dataset contains small and blurry texts that can be inaccurately identified in low-resolution images. Considering models are pre-trained with small resolution (768 × 768) images, all models are inherently incapable of identifying very small and blurry texts. These texts might be visualized as multi-colored clusters composed of different classes in Figure 9. In $W_{\text{OCR-read}}$ with SCOB and $W_{\text{OCR-read}}$ with online text rendering, a large multi-colored cluster remains, but scattered small multi-colored clusters disappear. This could be because our online text renderer makes the model robust against small and blurry texts by rendering texts with various image augmentations. As can be seen in Figure 10, our $W_{\text{OCR-read}}$ with SCOB extracts the latent representations more discriminatively than other models in the embedding space.

References

[1] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9365–9374, 2019. 1

[2] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019. 2, 3, 4

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2, 4

[4] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021. 1

[5] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-

[*] https://github.com/JaidedAI/EasyOCR

[†] https://github.com/PaddlePaddle/PaddleOCR

[‡] https://learn.microsoft.com/en-us/azure/cognitive-services/computer-vision/overview-ocr

image model. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 4

[6] Chee Kheng Ch'ng and Chee Seng Chan. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 935–942. IEEE, 2017. 2, 3, 4

[7] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 3

[8] Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Nikolaos Barmpalios, Rajiv Jain, Ani Nenkova, and Tong Sun. Unified pretraining framework for document understanding. *arXiv preprint arXiv:2204.10939*, 2022. 5

[9] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 3

[10] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE, 2015. 3, 4

[11] Mingxin Huang, Yuliang Liu, Zhenghao Peng, Chongyu Liu, Dahua Lin, Shenggao Zhu, Nicholas Yuan, Kai Ding, and Lianwen Jin. Swintextspotter: Scene text spotting via better synergy between text detection and text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4593–4603, 2022. 4

[12] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. *arXiv preprint arXiv:2204.08387*, 2022. 3, 4, 5

[13] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1156–1160. IEEE, 2015. 1, 2, 3, 4

[14] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*, pages 1484–1493. IEEE, 2013. 2, 3, 4

[15] Jihyung Kil, Soravit Changpinyo, Xi Chen, Hexiang Hu, Sebastian Goodman, Wei-Lun Chao, and Radu Soricut. Prestu: Pre-training for scene-text understanding. *arXiv preprint arXiv:2209.05534*, 2022. 4

[16] 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 (ECCV)*, 2022. 4

[17] Seonghyeon Kim, Seung Shin, Yoonsik Kim, Han-Cheol Cho, Taeho Kil, Jaeheung Surh, Seunghyun Park, Bado Lee, and Youngmin Baek. Deer: Detection-agnostic end-to-end recognizer for scene text spotting. *arXiv preprint arXiv:2203.05122*, 2022. 4

[18] Yair Kittenplon, Inbal Lavi, Sharon Fogel, Yarin Bar, R Manmatha, and Pietro Perona. Towards weakly-supervised text spotting using a multi-task transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4604–4613, 2022. 4

[19] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017. 3

[20] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer. *arXiv preprint arXiv:2203.02378*, 2022. 5

[21] 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. 4

[22] 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. 4

[23] Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. Towards end-to-end unified scene text detection and layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[24] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017. 1

[25] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 67–83, 2018. 4

[26] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 3, 4

[27] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 2, 3, 4

[28] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 3

[29] Ahmed Nassar, Nikolaos Livathinos, Maksym Lysak, and Peter Staar. Tableformer: Table structure understanding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4614–4623, 2022. 4

[30] 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)*, volume 1, pages 1454–1459. IEEE, 2017. 4

[31] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019. 2, 3, 4

[32] Dezhi Peng, Xinyu Wang, Yuliang Liu, Jiaxin Zhang, Mingxin Huang, Songxuan Lai, Jing Li, Shenggao Zhu, Dahua Lin, Chunhua Shen, et al. Spts: Single-point text spotting. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4272–4281, 2022. 4

[33] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 2, 3, 4

[34] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8802–8812, 2021. 3

[35] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19254–19264, 2023. 3, 4

[36] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 1, 5

[37] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 3, 4

[38] Xiang Zhang, Yongwen Su, Subarna Tripathi, and Zhuowen Tu. Text spotting transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9519–9528, 2022. 4

[39] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. In *European Conference on Computer Vision*, pages 564–580. Springer, 2020. 3, 4

[40] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE, 2019. 3, 4, 5

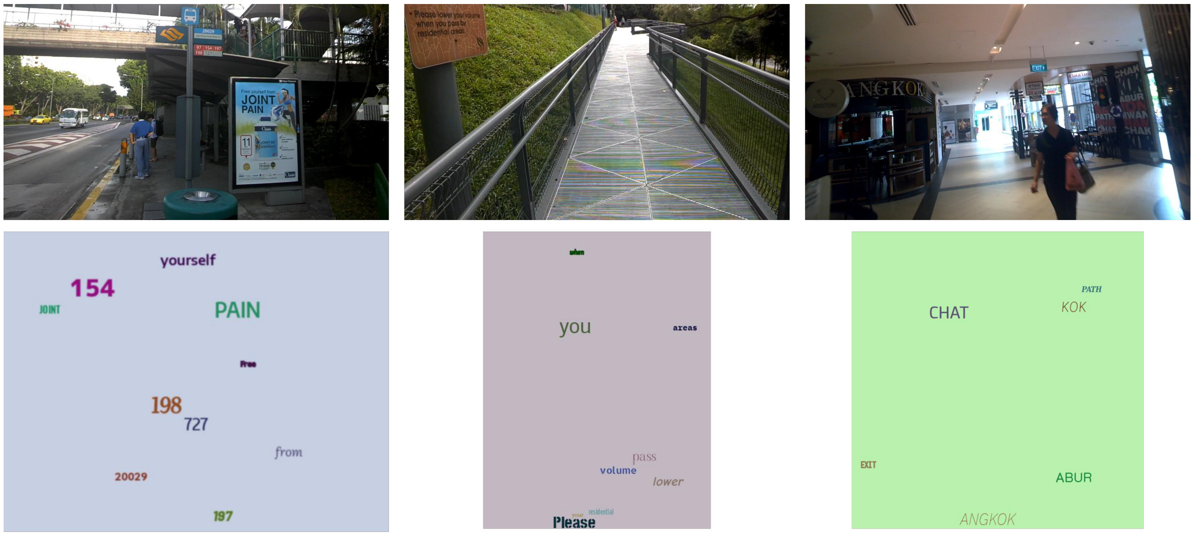
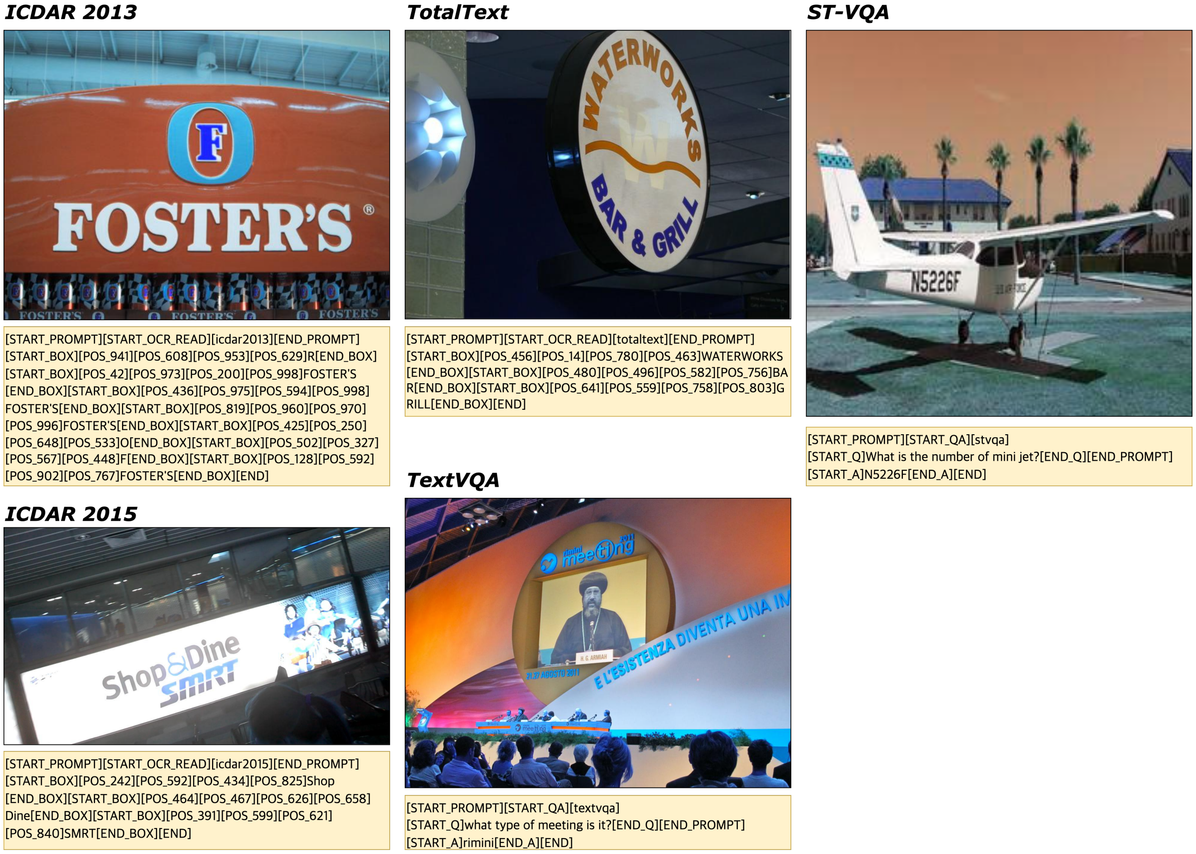


Figure 3. The visualization of images generated in a one-to-one correspondence to the ICDAR2015 test set using the proposed online renderer.



ICDAR 2013

```
[START_PROMPT][START_OCR_READ][icdar2013][END_PROMPT]
[START_BOX][POS_941][POS_608][POS_953][POS_629]R[END_BOX]
[START_BOX][POS_42][POS_973][POS_200][POS_998]FOSTER'S
[END_BOX][START_BOX][POS_436][POS_975][POS_594][POS_998]
FOSTER'S[END_BOX][START_BOX][POS_819][POS_960][POS_970]
[POS_996]FOSTER'S[END_BOX][START_BOX][POS_425][POS_250]
[POS_648][POS_533]O[END_BOX][START_BOX][POS_502][POS_327]
[POS_567][POS_448]F[END_BOX][START_BOX][POS_128][POS_592]
[POS_902][POS_767]FOSTER'S[END_BOX][END]
```

TotalText

```
[START_PROMPT][START_OCR_READ][totaltext][END_PROMPT]
[START_BOX][POS_456][POS_14][POS_780][POS_463]WATERWORKS
[END_BOX][START_BOX][POS_480][POS_496][POS_582][POS_756]BA
R[END_BOX][START_BOX][POS_641][POS_559][POS_758][POS_803]G
RILL[END_BOX][END]
```

ST-VQA

```
[START_PROMPT][START_QA][stvqa]
[START_Q]What is the number of mini jet?[END_Q][END_PROMPT]
[START_A]N5226F[END_A][END]
```

ICDAR 2015

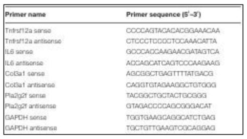
```
[START_PROMPT][START_OCR_READ][icdar2015][END_PROMPT]
[START_BOX][POS_242][POS_592][POS_434][POS_825]Shop
[END_BOX][START_BOX][POS_464][POS_467][POS_626][POS_658]
Dine[END_BOX][START_BOX][POS_391][POS_599][POS_621]
[POS_840]SMRT[END_BOX][END]
```

TextVQA

```
[START_PROMPT][START_QA][textvqa]
[START_Q]what type of meeting is it?[END_Q][END_PROMPT]
[START_A]rimini[END_A][END]
```

Figure 4. Examples of STU tasks. We provide a sample image for each dataset and its ground truth sequence.

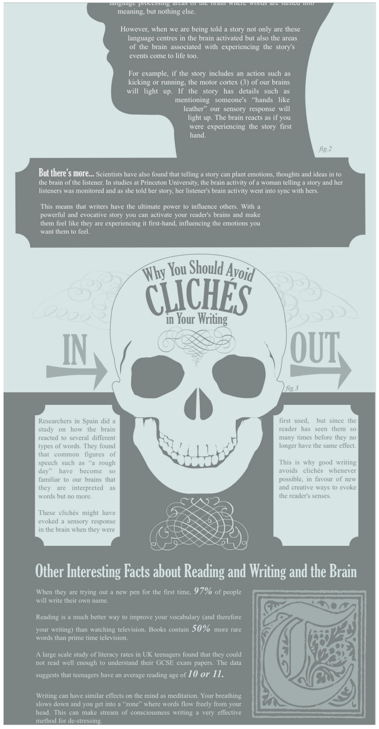
PubTabNet



Primer name	Primer sequence (5′–3′)
Tnfrsf12a sense	CCCCAGTACACACGGAAACAA
Tnfrsf12a antisense	CTCCCTCCCCTCCAAACATTA
IL6 sense	GCCCACCAAGAACGATAGTCA
IL6 antisense	ACCAGCATCAGTCCCAAGAAG
Col3a1 sense	AGCGGCTGAGTTTTATGACG
Col3a1 antisense	CAGGTGTAGAAGGCTGTGGG
Pla2g2f sense	TACGGCTGCTACTGCGGG
Pla2g2f antisense	GTAGACCCCAGCGGGACAT
GAPDH sense	TGGTGAAGCAGGCATCTGAG
GAPDH antisense	TGCTGTTGAAGTCGCAGGAG

[START_PROMPT][START_TABLE_PARSE][pubtabnet][END_PROMPT]
[START_TP]<html><body><table><thead><tr><td>Primer name
</td><td>Primer sequence (5′-3′)</td></tr></thead><tbody><tr>
<td>Tnfrsf12a sense</td><td>CCCCAGTACACACGGAAACAA</td>
</tr><tr><td>Tnfrsf12aantisense</td><td>CTCCCTCCCCTCCAAACAT
TA</td></tr><tr><td>IL6 sense</td><td>GCCCACCAAGAACGATAGT
CA</td></tr><tr><td>IL6 antisense</td><td>ACCAGCATCAGTCCCA
AGAAG</td></tr><tr><td>Col3a1 sense</td><td>AGCGGCTGAGTTT
TATGACG</td></tr><tr><td>Col3a1 antisense</td><td>CAGGTGTAG
AAGGCTGTGGG</td></tr><tr><td>Pla2g2f sense</td><td>TACGGC
TGCTACTGCGGG</td></tr><tr><td>Pla2g2f antisense</td><td>GTA
GACCCCAGCGGGACAT</td></tr><tr><td>GAPDH sense</td><td>TG
GTGAAGCAGGCATCTGAG</td></tr><tr><td>GAPDH antisense</td>
<td>TGCTGTT GAAGTCGCAGGAG</td></tr></tbody>/table></body
></html>[END_TP][END]

InfoVQA



meaning, but nothing else.

However, when we are being told a story not only are these language centres in the brain activated but also the areas of the brain associated with experiencing the story's events come to life too.

For example, if the story includes an action such as kicking or running, the motor cortex (3) of our brains will light up. If the story has details such as mentioning someone's "hands like leather" our sensory response will light up. The brain reacts as if you were experiencing the story first hand.

fig.2

But there's more... Scientists have also found that telling a story can plant emotions, thoughts and ideas in to the brain of the listener. In studies at Princeton University, the brain activity of a woman telling a story and her listeners was monitored and as she told her story, her listener's brain activity went into sync with hers.

This means that writers have the ultimate power to influence others. With a powerful and evocative story you can activate your reader's brains and make them feel like they are experiencing it first-hand, influencing the emotions you want them to feel.

Why You Should Avoid CLICHÉS in Your Writing

IN

OUT

fig.3

Researchers in Spain did a study on how the brain reacted to several different types of words. They found that common figures of speech such as "a rough day" have become so familiar to our brains that they are interpreted as words but no more.

These clichés might have evoked a sensory response in the brain when they were first used, but since the reader has seen them so many times before they no longer have the same effect.

This is why good writing avoids clichés whenever possible, in favour of new and creative ways to evoke the reader's senses.

Other Interesting Facts about Reading and Writing and the Brain

When they are trying out a new pen for the first time, *97%* of people will write their own name.

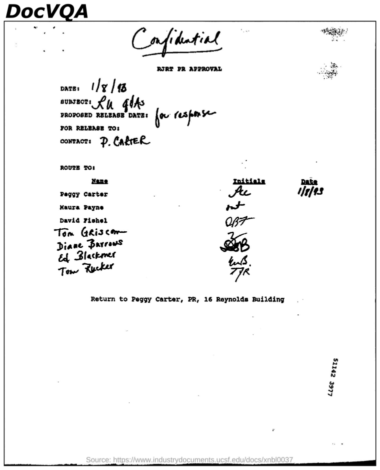
Reading is a much better way to improve your vocabulary (and therefore your writing) than watching television. Books contain *50%* more rare words than prime time television.

A large scale study of literacy rates in UK teenagers found that they could not read well enough to understand their GCSE exam papers. The data suggests that teenagers have an average reading age of *10 or 11.*

Writing can have similar effects on the mind as meditation. Your breathing slows down and you get into a "zone" where words flow freely from your head. This can make stream of consciousness writing a very effective method for de-stressing.

[START_PROMPT][START_QA][docvqa_task3]
[START_Q]what is mentioned on the forehead of the skull[END_Q]
[END_PROMPT][START_A]why you should avoid cliches in your writin
g[END_A][END]

DocVQA



Confidential

RJRT PR APPROVAL

DATE: 1/8/93
SUBJECT: RJR Q&As
PROPOSED RELEASE DATE: for response
FOR RELEASE TO:
CONTACT: P. CARTER

ROUTE TO:

Name	Initials	Date
Peggy Carter	PC	1/8/93
Maura Payne	out	
David Fishel	DBF	
Tom Griscom		
Diane Barrows	DB	
Ed Blackmer	EwB.	
Tom Rucker	TJR	

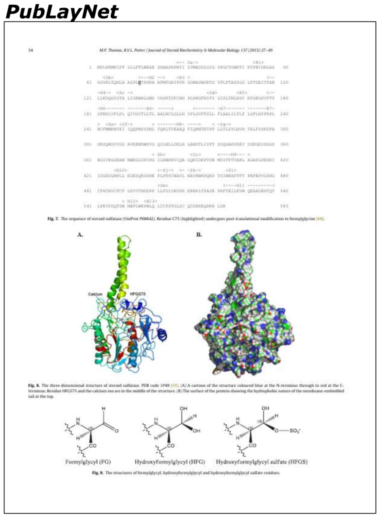
Return to Peggy Carter, PR, 16 Reynolds Building

51142 3977

Source: https://www.industrydocuments.ucsf.edu/docs/xnbl0037

[START_PROMPT][START_QA][docvqa_task1]
[START_Q]what is the date mentioned in this letter?[END_Q]
[END_PROMPT][START_A]1/8/93[END_A][END]

PubLayNet



34 M.P. Thomas, B.V.L. Potter / Journal of Steroid Biochemistry & Molecular Biology 137 (2013) 27–49

Fig. 7. The sequence of steroid sulfatase (UniProt P08842). Residue C75 (highlighted) undergoes post-translational modification to formylglycine [60].

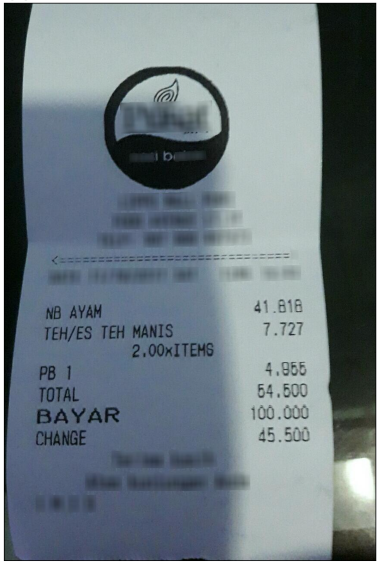
Fig. 8. The three-dimensional structure of steroid sulfatase. PDB code 1P49 [59]. (A) A cartoon of the structure coloured blue at the N-terminus through to red at the C-terminus. Residue HFGS75 and the calcium ion are in the middle of the structure. (B) The surface of the protein showing the hydrophobic nature of the membrane-embedded tail at the top.

Formylglycyl (FG) Hydroxyformylglycyl (HFG) Hydroxyformylglycyl sulfate (HFGS)

Fig. 9. The structures of formylglycyl, hydroxyformylglycyl and hydroxyformylglycyl sulfate residues.

[START_PROMPT][START_OD][publaynet][END_PROMPT]
[START_OBJECT][POS_118][POS_396][POS_864][POS_407]
[publaynet__text][END_OBJECT][START_OBJECT][POS_66][POS_735]
[POS_916][POS_768][publaynet__text][END_OBJECT][START_OBJECT]
[POS_238][POS_912][POS_744][POS_923][publaynet__text]
[END_OBJECT][START_OBJECT][POS_215][POS_73][POS_766]
[POS_382][publaynet__figure][END_OBJECT][START_OBJECT]
[POS_197][POS_423][POS_785][POS_726][publaynet__figure]
[END_OBJECT][START_OBJECT][POS_164][POS_784][POS_755]
[POS_903][publaynet__figure][END_OBJECT][END]

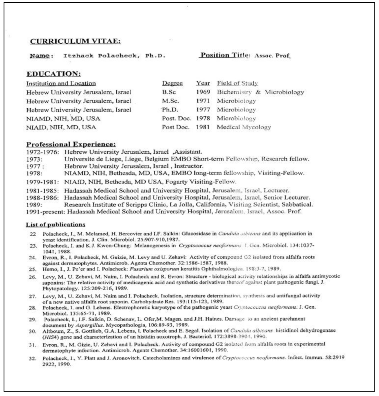
CORD



NB AYAM	41.818
TEH/ES TEH MANIS	7.727
2.00xITEMS	
PB 1	4,955
TOTAL	54,500
BAYAR	100.000
CHANGE	45.500

[START_PROMPT][START_KIE][cord][END_PROMPT]
[START_menu.nm]NB AYAM[END_menu.nm][START_menu.price]
41.818[END_menu.price][START_menu.nm]TEH/ES TEH MANIS
[END_menu.nm][START_menu.price]7.727[END_menu.price]
[START_total.menuqty_cnt]2.00xITEMS[END_total.menuqty_cnt]
[START_sub_total.discount_price]PB 1 4,955[END_sub_total.discount_p
rice][START_total.total_price]TOTAL 54,500[END_total.total_price]
[START_total.cashprice]BAYAR 100.000[END_total.cashprice]
[START_total.changeprice]CHANGE 45.500[END_total.changeprice]
[END]

RVL-CDIP



CURRICULUM VITAE:

Name: Itzhack Polacheck, Ph.D. Position Title: Assoc. Prof.

EDUCATION:

Institution and Location	Degree	Year	Field of Study
Hebrew University Jerusalem, Israel	B.Sc	1969	Bichemistry & Microbiology
Hebrew University Jerusalem, Israel	M.Sc.	1971	Microbiology
Hebrew University Jerusalem, Israel	Ph.D.	1977	Microbiology
NIAMD, NIH, MD, USA	Post. Doc.	1978	Microbiology
NIAID, NIH, MD, USA	Post Doc.	1981	Medical Mycology

Professional Experience:

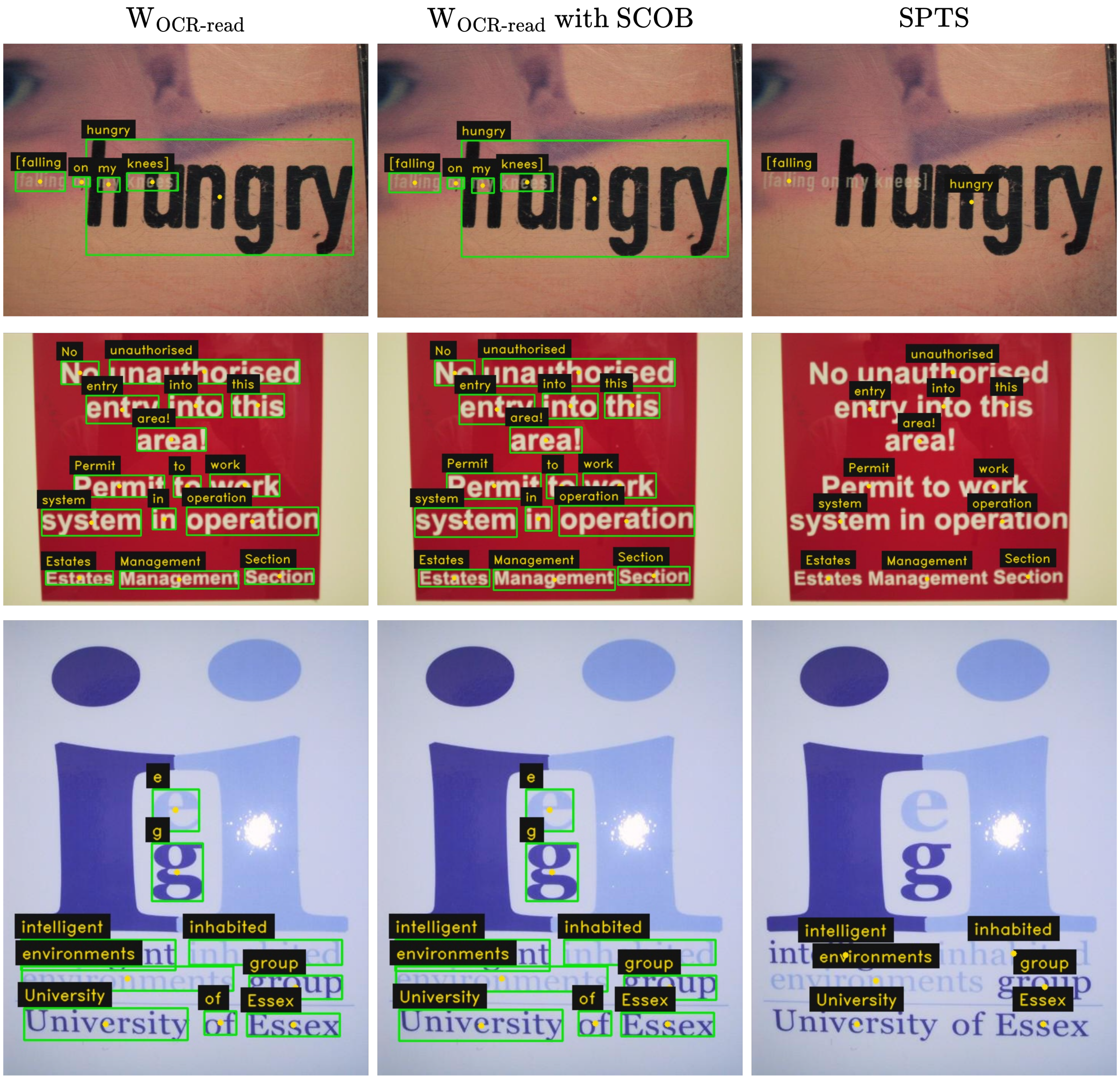
1972-1976: Hebrew University Jerusalem, Israel ,Assistant.
1973: Universite de Liege, Liege, Belgium EMBO Short-term Fellowship, Research fellow.
1977 : Hebrew University Jerusalem, Israel , Instructor.
1978: NIAMD, NIH, Bethesda, MD, USA, EMBO long-term fellowship, Visiting-Fellow.
1979-1981: NIAID, NIH, Bethesda, MD USA, Fogarty Visiting-Fellow.
1981-1985: Hadassah Medical School and University Hospital, Jerusalem, Israel, Lecturer.
1988-1986: Hadassah Medical School and University Hospital, Jerusalem, Israel, Senior Lecturer.
1989: Research Institute of Scripps Clinic, La Jolla, California, Visiting Scientist, Sabbatical.
1991-present: Hadassah Medical School and University Hospital, Jerusalem, Israel, Assoc. Prof.

List of publications

22 Polacheck, I., M. Melamed, H. Bercovier and I.F. Salkin: Glucosidase in *Candida albicans* and its application in yeast identification. J. Clin. Microbiol. 25:907-910,1987.
23. Polacheck, I. and K.J. Kwon-Chung: Melanogenesis in *Cryptococcus neoformans*. J. Gen. Microbiol. 134:1037-1041, 1988.
24. Evron, R., I. Polacheck, M. Guizie, M. Levy and U. Zehavi: Activity of compound G2 isolated from alfalfa roots against dermatophytes. Antimicrob. Agents Chemother. 32:1586-1587, 1988.
25. Hemo, I., J. Pe'er and I. Polacheck: *Fusarium oxisporum* keratitis Ophthalmologica. 198:3-7, 1989.
26. Levy, M., U. Zehavi, M. Naim, I. Polacheck and R. Evron: Structure - biological activity relationships in alfalfa antimycotic saponins: The relative activity of medicagenic acid and synthetic derivatives thereof against plant pathogenic fungi. J. Phytopatology. 125:209-216, 1989.
27. Levy, M., U. Zehavi, M. Naim and I. Polacheck. Isolation, structure determination, synthesis and antifungal activity of a new native alfalfa root saponin. Carbohydrate Res. 193:115-123, 1989.
28. Polacheck, I. and G. Lebens. Electrophoretic karyotype of the pathogenic yeast *Cryptococcus neoformans*. J. Gen. Microbiol. 135:65-71, 1989.
29. Polacheck, I., I.F. Salkin, D. Schenav, L. Ofer,M. Magen. and J.H. Haines. Damage to an ancient parchment document by *Aspergillus*. Mycopathologia, 106:89-93, 1989.
30. Altboum, Z., S. Gottlieb, G.A. Lebens, I. Polacheck and E. Segal. Isolation of *Candida albicans* histidinol dehydrogenase (*HIS4*) gene and characterization of an histidin auxotroph. J. Bacteriol. 172:3898-3904, 1990.
31. Evron, R., M. Gizie, U. Zehavi and I. Polacheck. Activity of compound G2 isolated from alfalfa roots in experimental dermatophyte infection. Antimicrob. Agents Chemother. 34:16001601, 1990.
32. Polacheck, I., Y. Platt and J. Aronovitch. Catecholamines and virulence of *Cryptococcus neoformans*. Infect. Immun. 58:2919 2922, 1990.

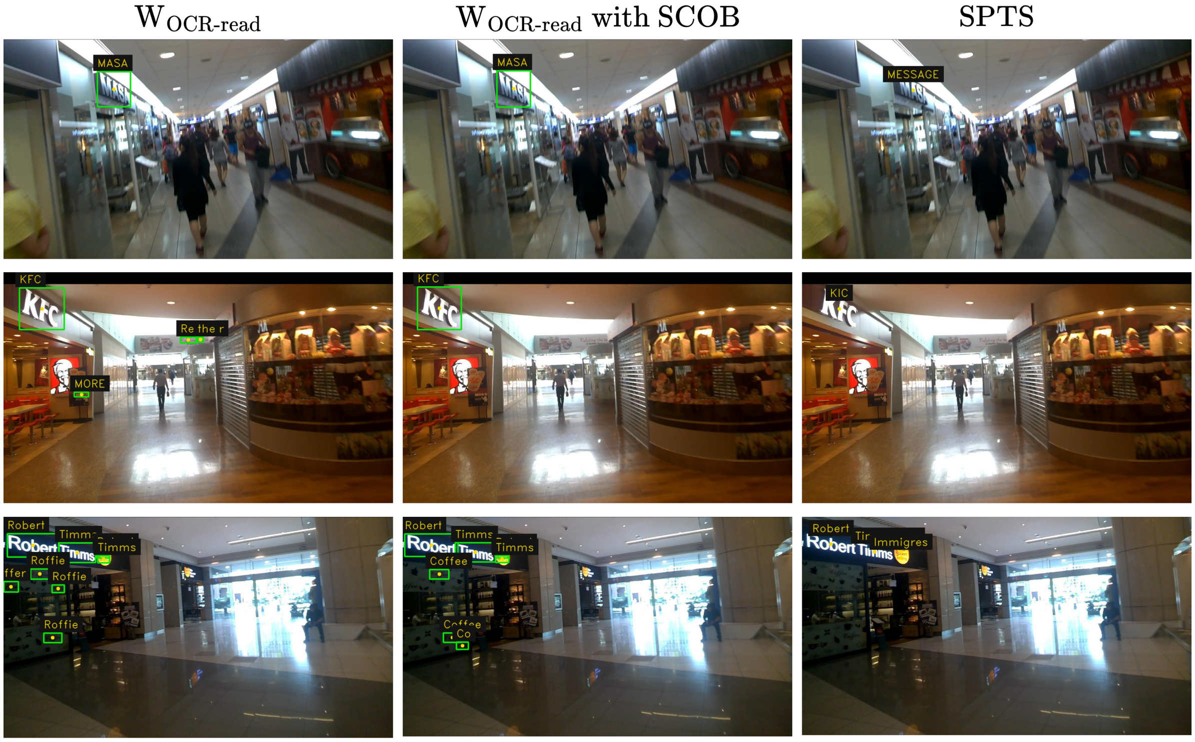
[START_PROMPT][START_CLASSIFICATION][rvl_cdip][END_PROMPT]
[START_CLASS][rvl_cdip__resume][END_CLASS][END]

Figure 5. Examples of VDU tasks. We provide a sample image for each dataset and its ground truth sequence. Note that a sample image of InfoVQA dataset is too long vertically, thus we crop the answer part of the question and report that.



$\mathrm{W}_{\text{OCR-read}}$ $\mathrm{W}_{\text{OCR-read}}$ with SCOB SPTS

Figure 6. The visualization of OCR prediction on ICDAR2013 test set. Our W models predicts in the form of a bounding box and SPTS predicts in the form of single-points. Note that the central points of the bounding boxes predicted by our W models are displayed for comparison.

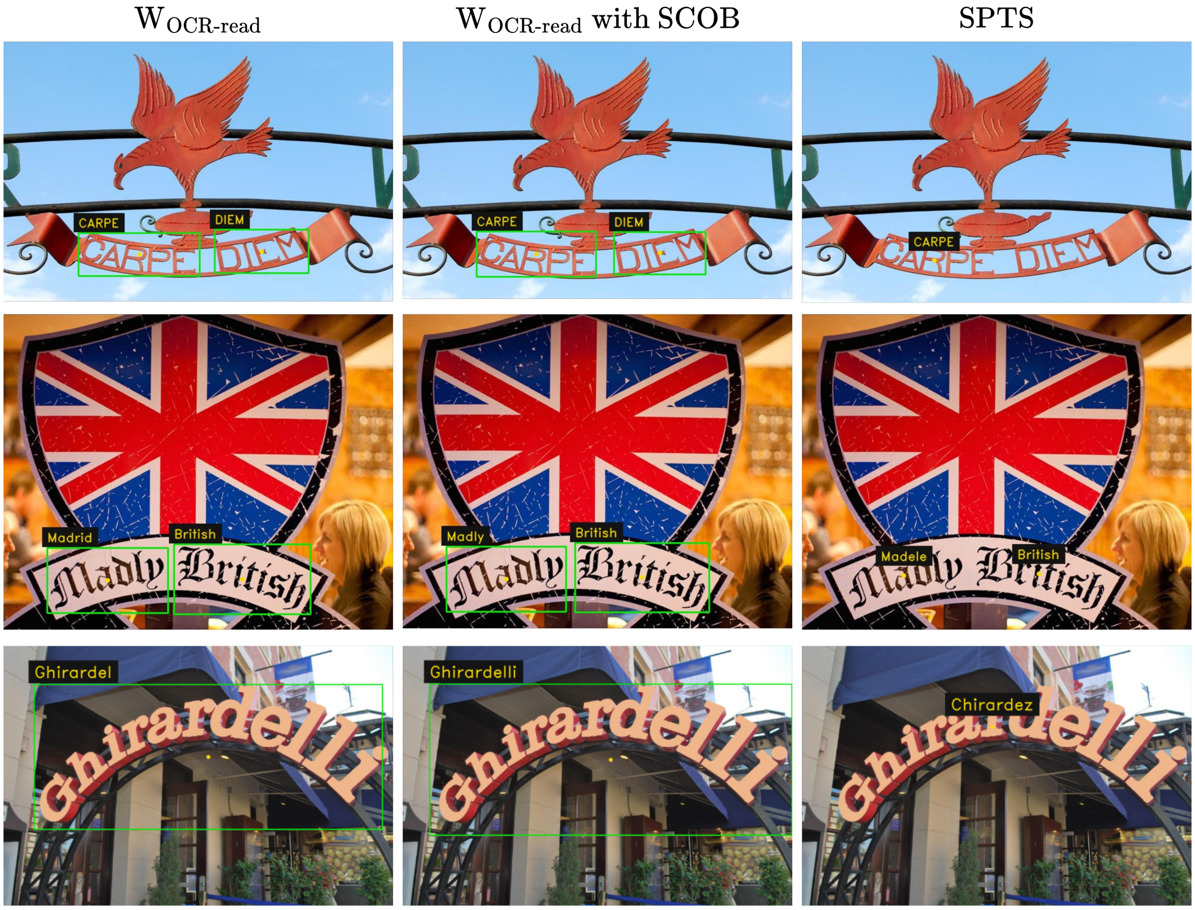


$W_{\text{OCR-read}}$

$W_{\text{OCR-read}}$ with SCOB

SPTS

Figure 7. The visualization of OCR prediction on ICDAR2015 test set. Our W models predicts in the form of a bounding box and SPTS predicts in the form of single-points. Note that the central points of the bounding boxes predicted by our W models are displayed for comparison.



$W_{\text{OCR-read}}$

$W_{\text{OCR-read}}$ with SCOB

SPTS

Figure 8. The visualization of OCR prediction on TotalText test set. Our W models predicts in the form of a bounding box and SPTS predicts in the form of single-point. Note that the central points of the bounding boxes predicted by our W models are displayed for comparison.

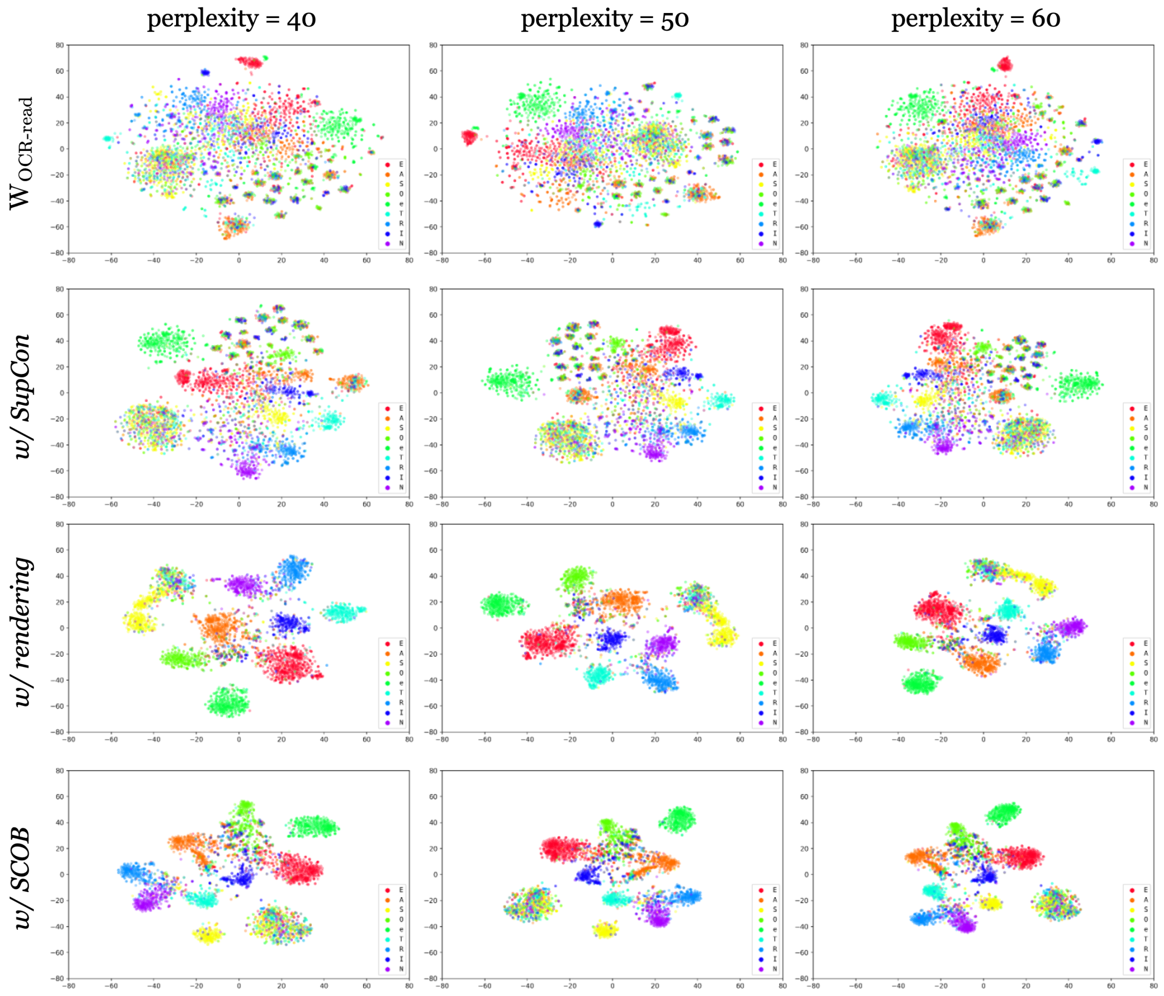


Figure 9. Examples of t-SNE visualization. We visualize representations extracted from the final layer of the decoder. Note that the models are pre-trained weights. *Data*: ICDAR2015 test set.

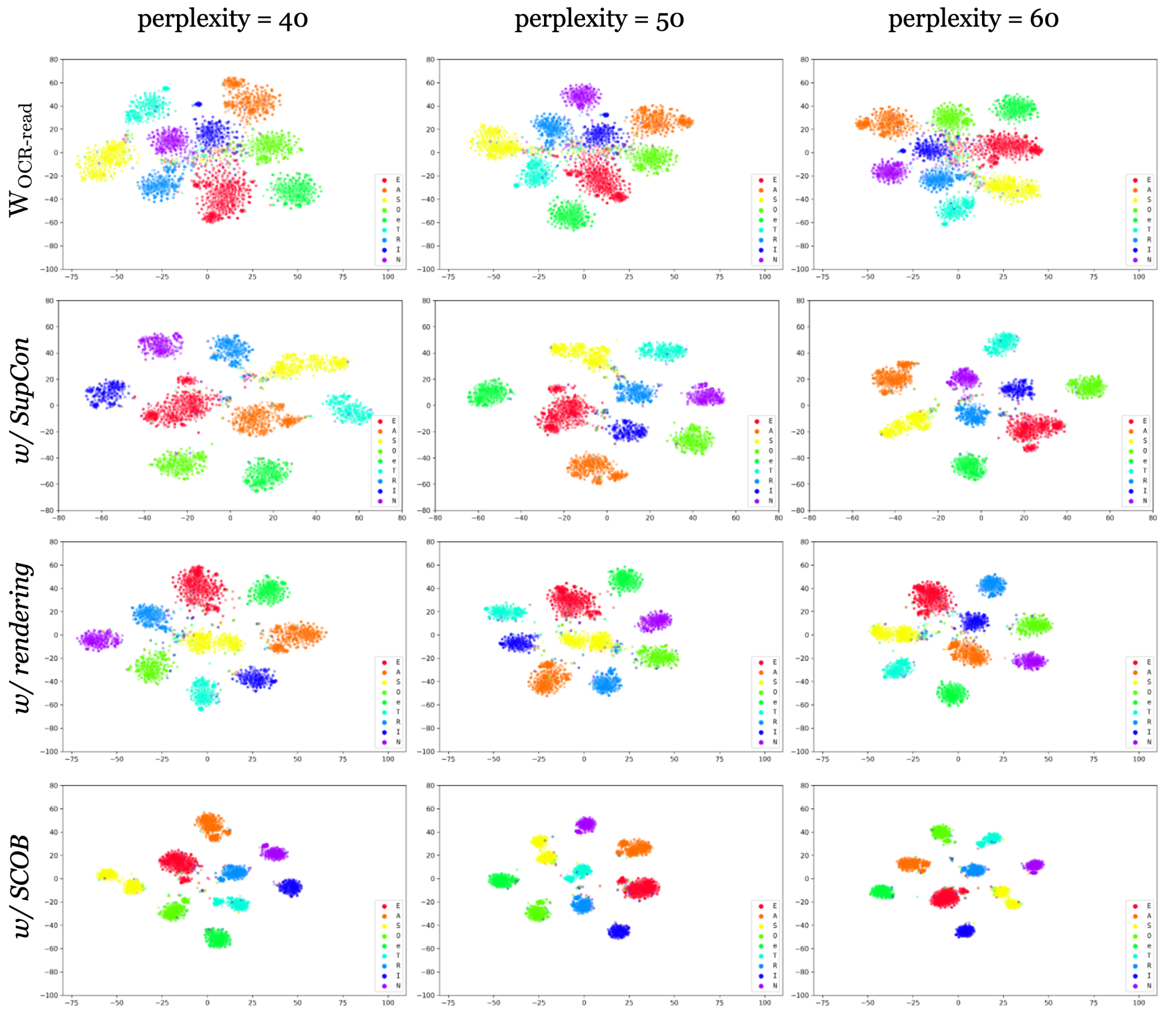


Figure 10. Examples of t-SNE visualization. We visualize representations extracted from the final layer of the decoder. The models are fine-tuned for scene text OCR. *Data*: ICDAR2015 test set.