Supplementary Material of Pure Transformer with Integrated Experts for Scene Text Recognition

Yew Lee Tan[1], Adams Wai-Kin Kong[1], and Jung-Jae Kim[2]

[1] Nanyang Technological University, Singapore
[2] Institute for Infocomm Research, A*STAR, Singapore

A Number of PTIE Layers

Multiple PTIE models were trained with varying number of layers for the encoder and decoder noting that both have the same number of layers. The results of the latency and weighted average accuracy are shown in Table A1. The PTIE models show a trade-off between accuracy to number of parameters and latency. All the PTIE models are competitive/outperform other recent works that are open source in terms of accuracy.

Table A1. Inference time and weighted average accuracy of recent works. The total count of 7672 uses IC15 (2077) on top of the 5 other datasets namely IIIT, IC13, SVT, SVT-P, and CT. 7406 uses IC15 (1811) while 7248 uses IC15 (1811) and a filtered version of IC13. The variation in total count is due to other works using varied set of benchmarks

Method	**Year**	**Avg. accuracy**			**Parameters**	**Time**
		7672	7406	7248	**(mil.)**	**(ms)**
Wang et al. [4]	AAAI '20	86.9	-	-	18.4	22
Lu et al. [2]	PR '21	89.3	-	-	54.6	53
Fang et al. [1]	CVPR '21	-	92.8	-	36.7	27
Yan et al. [5]	CVPR '21	-	-	91.5	29.1	29
PTIE–4 layers		91.6	92.8	92.9	31.2	36
PTIE–5 layers		91.9	93.2	93.2	38.6	45
PTIE–6 layers		92.4	94.1	93.5	45.9	52

B Impact of Padding the Images

The scene text images are resized to a fixed height and width before being passed on as input to the model. Most of the recent works resize the images without preserving the original aspect ratios which is also the method we adopted. Shi et al. [3] stated that padding the images while maintaining the original aspect ratios resulted in worse performance in most cases. This is also the case for our transformer-only model as shown in Table A2 where padding the images has a weighted average accuracy of 89.5% while resizing the images without preserving the original aspect ratios has an accuracy of 90.9%.

Table A2. Results of model trained with and without padding

Method	**Regular Text**			**Irregular Text**			
	IIIT	IC13	SVT	IC15	SVT-P	CT	Avg.
	3000	1015	647	2077	645	288	7672
With padding	95.3	96.0	91.7	79.0	85.7	86.8	89.5
Without padding	95.6	96.4	93.4	81.9	88.1	89.2	90.9

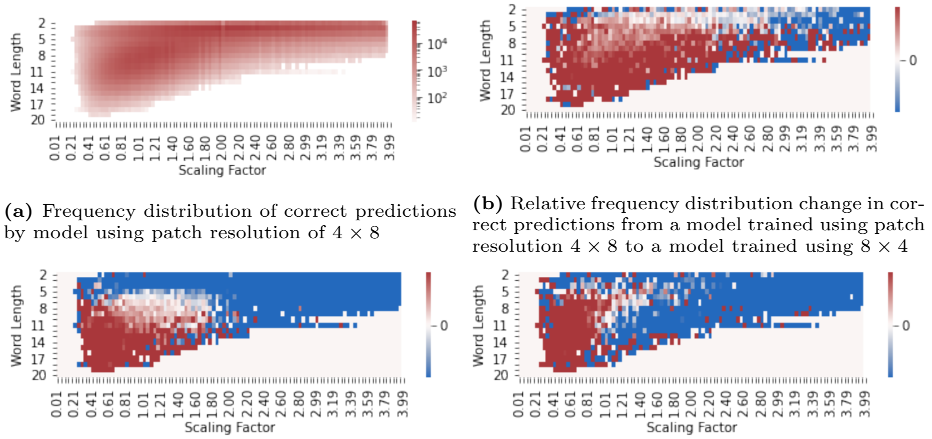
C Impact of Patch Resolutions and Sizes

As per Sec. 3.1 in the main paper, six models were trained with three pairs of inverting resolutions namely: (1) patch resolution (height × width) of 4×8 vs 8×4, (2) patch resolution of 2×16 vs 16×2, and (3) patch resolution of 4×16 vs 16×4. The three relative distribution changes are visualized in Fig. A1. The distributions come from the models' results on the training dataset as large amount of samples are required to provide a reliable visualization. Word length is ranged from 2 to 20. The scaling factor ranges from 0 to 4 and bins with frequency counts lesser than 100 are removed. The remaining counts account for 95% of the total counts. These arrangements seek to reduce the noise caused by bins with low frequency and provide better visualizations. Figs. A1b to A1d suggest that patches with resolution of height lower than width will result in more correct predictions on samples with higher scaling factor and vice-versa. Fig. A6 shows samples from the train dataset with respect the various word lengths, l, and scaling factors, s.

Results of models trained with various patch sizes and resolutions are tabulated in Table A3. All models were trained with the same hyperparameters as specified in the main paper. Generally, the weighted average accuracy of the models decreases with the increase in patch size. The highest accuracies come from patch resolutions of 4×8 and 8×4 with a patch size of 32, therefore they were chosen as the resolutions for PTIE.

Table A3. Results of models trained with different patch sizes and resolutions

Patch Size	**Patch Resolution**	**Regular Text**			**Irregular Text**			
		IIIT	IC13	SVT	IC15	SVT-P	CT	Avg.
		3000	1015	647	2077	645	288	7672
16	4×4	95.2	95.4	91.2	81.3	87.8	89.2	90.3
32	2×16	95.1	96.2	92.0	80.5	86.0	88.9	90.0
32	16×2	95.1	95.7	92.3	79.5	87.0	86.8	89.7
32	4×8	95.4	96.6	93.4	80.7	88.1	87.8	90.5
32	8×4	95.6	96.4	93.4	81.9	88.1	89.2	90.9
64	8×8	94.0	95.7	91.7	80.0	87.0	87.2	89.4
64	4×16	94.3	95.6	91.3	79.5	87.1	84.4	89.2
64	16×4	94.5	95.6	91.2	78.9	85.7	86.1	89.1



(a) Frequency distribution of correct predictions by model using patch resolution of 4 × 8

(b) Relative frequency distribution change in correct predictions from a model trained using patch resolution 4 × 8 to a model trained using 8 × 4

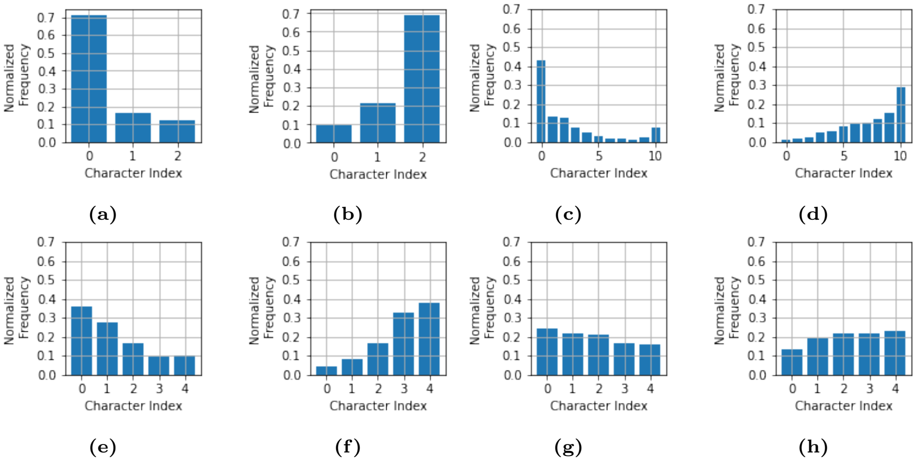
(c) Relative frequency distribution change in correct predictions from a model trained using patch resolution 2 × 16 to a model trained using 16 × 2

(d) Relative frequency distribution change in correct predictions from a model trained using patch resolution 4 × 16 to a model trained using 16 × 4

Fig. A1

D Errors in First Character Prediction

Two models were trained as per Sec. 3.1 of the main paper where one model uses the original ground-truth while the other uses an reversed ground-truth. The normalized frequency distributions of wrong character(s) prediction on words with various lengths are plotted in Fig. A2. The models have the highest error rate when decoding the first character.



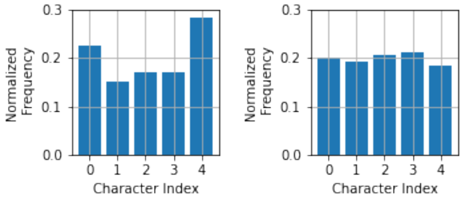
(a) (b) (c) (d) (e) (f) (g) (h)

Fig. A2. Normalized frequency distributions of 1 wrong character prediction conditioned on ground truth characters for (a, b) word length 3; (c, d) word length 11; (e, f) 2 wrongly predicted characters on words with length 5; (g, h) 4 wrongly predicted characters on words with length 5. (a, c, e, g) are trained with original ground-truth while (b, d, f, h) are trained with inverted ground-truth

Apart from PTIE, non-autoregressive decoding method using the transformer decoder is also adopted. Zhu et al. [6] proposed the use of learnable positional encoding as queries vector to replace the sequence input for the decoder in object detection tasks. We hypothesize that the weak first character prediction may be due to less information being available when decoding earlier characters, as compared with later characters, in the autoregressive decoding process. Basing off the method by Zhu et al., all the characters in the text sequence for STR would be predicted in parallel and therefore, would have equal amount of information thereby solving the problem with first character prediction. The normalized frequency distributions of wrong character predictions as shown in Fig. A3 shows that the non-autoregressive method indeed mitigates the problem of weak first character. However, the overall word prediction accuracy is lower than that of the autoregressive method as shown in Table A4. This may be due to the lack of previous character grounding during training as query vectors are used as a replacement to the ground-truth text input.

Table A4. Results of autoregressive and non-autoregressive models

Method	**Regular Text**			**Irregular Text**			
	IIIT	IC13	SVT	IC15	SVT-P	CT	Avg.
	3000	1015	647	2077	645	288	7672
Non-autoregressive	92.7	92.5	89.5	74.3	84.8	85.8	86.5
Autoregressive	95.6	96.4	93.4	81.9	88.1	89.2	90.9

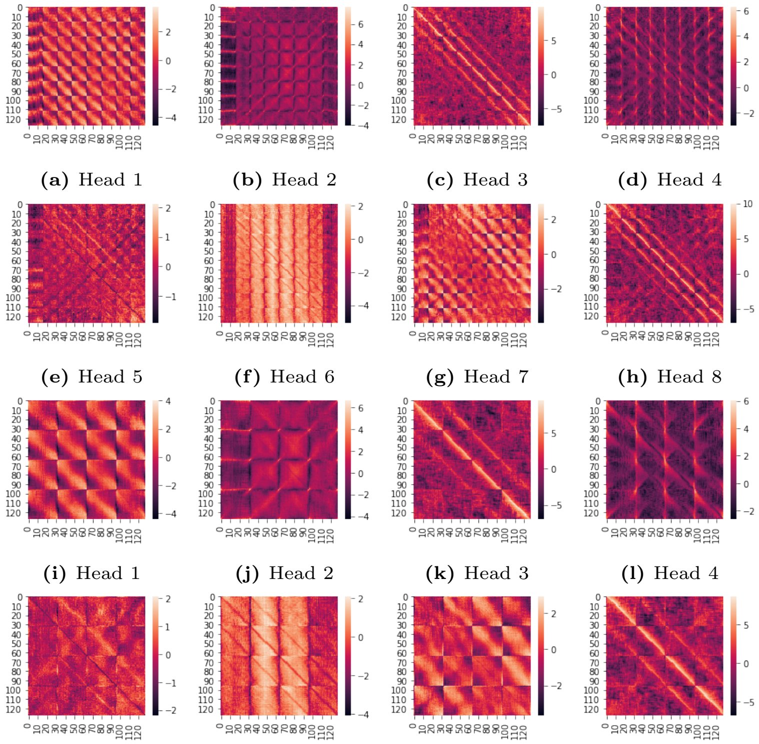


(a) **(b)**

Fig. A3. Normalized frequency distributions of wrong predictions by the non-autoregressive method for word length 5. (a) Predictions with one wrong character. (b) Predictions with two wrong characters

E Positional Attention Maps

The flatten patches of different patch resolutions have different spatial layouts. Since most of the parameters in PTIE are shared, the handling of spatial layouts (for patch resolution 4 × 8 and 8 × 4) are done by the unnormalized positional attention maps as shown in Fig. A4 and Fig. A5. As PTIE contains 16 attentions heads, Fig. A4 shows the first 8 heads of each resolution and Fig. A5 shows the last 8 heads of each resolution. It seems that the patterns are denser in unnormalized attention maps of resolution 4 × 8 as they have more vertically adjacent patches before row-major order flattening.



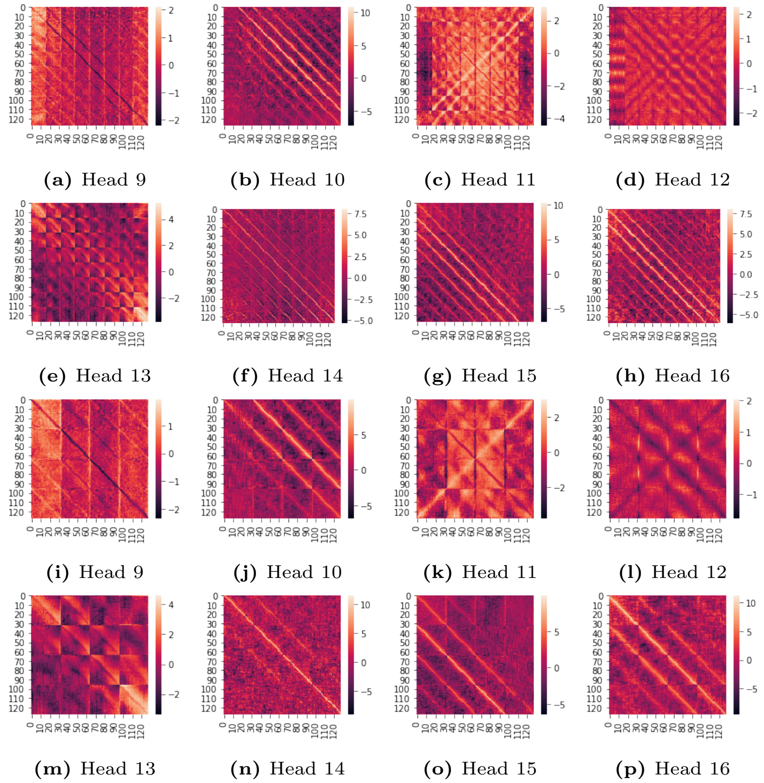
(a) Head 1 **(b)** Head 2 **(c)** Head 3 **(d)** Head 4

(e) Head 5 **(f)** Head 6 **(g)** Head 7 **(h)** Head 8

(i) Head 1 **(j)** Head 2 **(k)** Head 3 **(l)** Head 4

(m) Head 5 **(n)** Head 6 **(o)** Head 7 **(p)** Head 8

Fig. A4. Unnormalized positional attention maps from first 8 heads of trained PTIE encoder for (a-h) patch resolutions of 4 × 8, and (i-p) 8 × 4



(a) Head 9 **(b)** Head 10 **(c)** Head 11 **(d)** Head 12

(e) Head 13 **(f)** Head 14 **(g)** Head 15 **(h)** Head 16

(i) Head 9 **(j)** Head 10 **(k)** Head 11 **(l)** Head 12

(m) Head 13 **(n)** Head 14 **(o)** Head 15 **(p)** Head 16

Fig. A5. Unnormalized positional attention maps from last 8 heads of trained PTIE encoder for (a-h) patch resolutions of 4 × 8, and (i-p) 8 × 4

F Sample Images

Sample images with different word lengths and scaling factors are shown in Fig. A6 where l and s represent word length and scaling factor respectively. Images with word length 3-5 and scaling factor of 1.2-2.4 are least affected by the patch resolution used. Images with (1) word length of 2-3; scaling factor < 1, and (2) word length 2-11; scaling factor > 2.6, favours patch resolution of 4×8. Images with word length > 5 and scaling factor < 2.6 favours patch resolution of 8×4. Examples of success and failure cases are shown in Fig. A7 and Fig. A8 respectively.

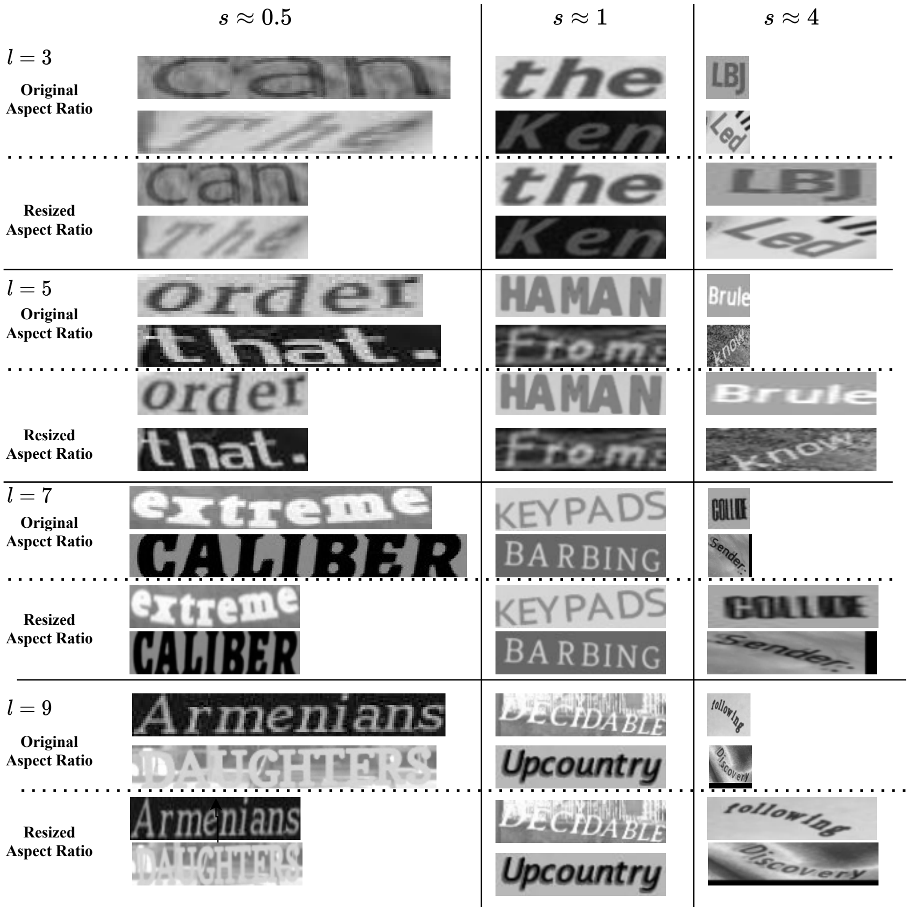
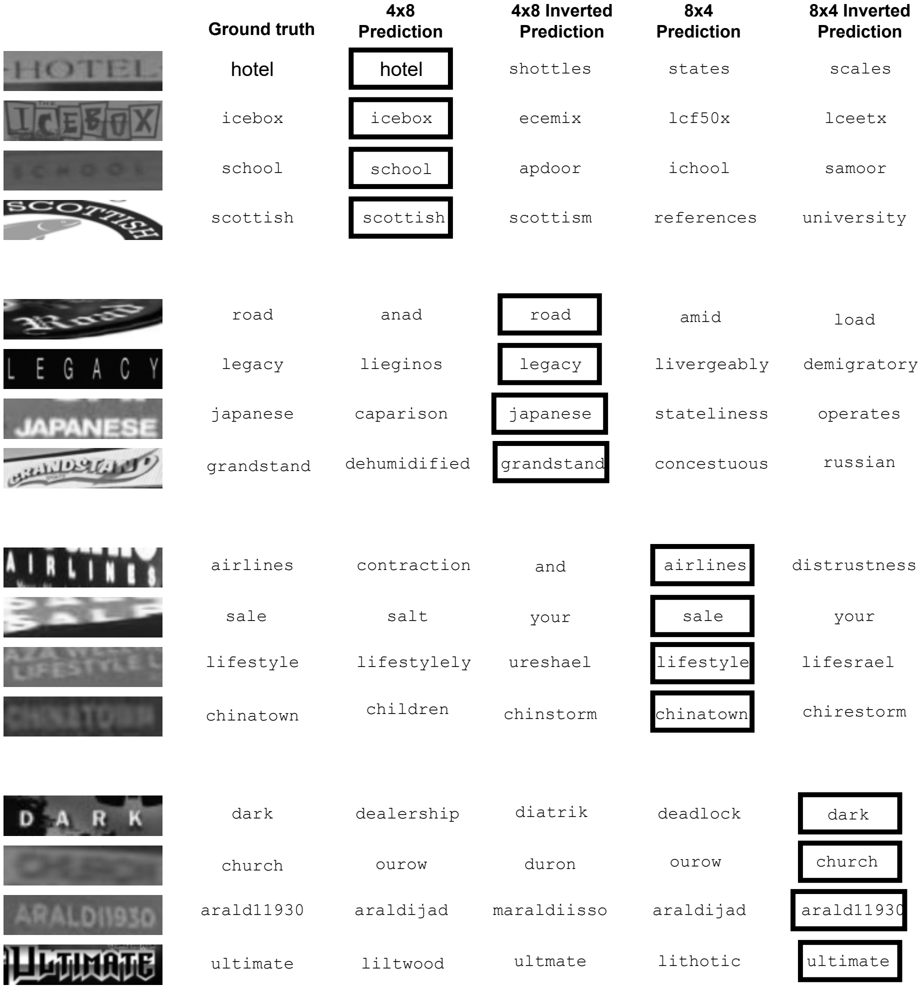
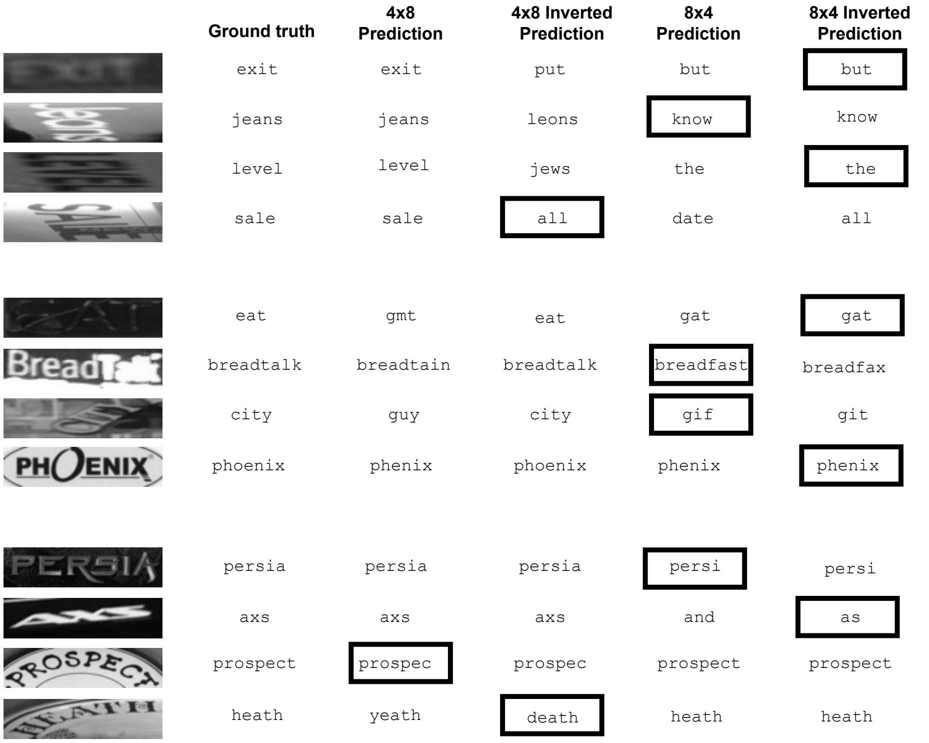


Fig. A6. Sample images with different word lengths and scaling factors



Ground truth	4x8 Prediction	4x8 Inverted Prediction	8x4 Prediction	8x4 Inverted Prediction
hotel	**hotel**	shottles	states	scales
icebox	**icebox**	ecemix	lcf50x	lceetx
school	**school**	apdoor	ichool	samoor
scottish	**scottish**	scottism	references	university
road	anad	**road**	amid	load
legacy	lieginos	**legacy**	livergeably	demigratory
japanese	caparison	**japanese**	stateliness	operates
grandstand	dehumidified	**grandstand**	concestuous	russian
airlines	contraction	and	**airlines**	distrustness
sale	salt	your	**sale**	your
lifestyle	lifestylely	ureshael	**lifestyle**	lifesrael
chinatown	children	chinstorm	**chinatown**	chirestorm
dark	dealership	diatrik	deadlock	**dark**
church	ourow	duron	ourow	**church**
arald11930	araldijad	maraldiisso	araldijad	**arald11930**
ultimate	liltwood	ultmate	lithotic	**ultimate**

Fig. A7. Examples of success cases with PTIE. The boxed text represents final output from PTIE



	Ground truth	4x8 Prediction	4x8 Inverted Prediction	8x4 Prediction	8x4 Inverted Prediction
	exit	exit	put	but	$\boxed{\text{but}}$
	jeans	jeans	leons	$\boxed{\text{know}}$	know
	level	level	jews	the	$\boxed{\text{the}}$
	sale	sale	$\boxed{\text{all}}$	date	all
	eat	gmt	eat	gat	$\boxed{\text{gat}}$
	breadtalk	breadtain	breadtalk	$\boxed{\text{breadfast}}$	breadfax
	city	guy	city	$\boxed{\text{gif}}$	git
	phoenix	phenix	phoenix	phenix	$\boxed{\text{phenix}}$
	persia	persia	persia	$\boxed{\text{persi}}$	persi
	axs	axs	axs	and	$\boxed{\text{as}}$
	prospect	$\boxed{\text{prospec}}$	prospec	prospect	prospect
	heath	yeath	$\boxed{\text{death}}$	heath	heath

Fig. A8. Examples of failure cases with PTIE. The boxed text represents final output from PTIE

References

1. Fang, S., Xie, H., Wang, Y., Mao, Z., Zhang, Y.: Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In: CVPR. pp. 7098–7107 (2021)
2. Lu, N., Yu, W., Qi, X., Chen, Y., Gong, P., Xiao, R., Bai, X.: Master: Multi-aspect non-local network for scene text recognition. PR **117**, 107980 (2021)
3. Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: Aster: An attentional scene text recognizer with flexible rectification. PAMI **41**(9), 2035–2048 (2018)
4. Wang, T., Zhu, Y., Jin, L., Luo, C., Chen, X., Wu, Y., Wang, Q., Cai, M.: Decoupled attention network for text recognition. In: AAAI. vol. 34, pp. 12216–12224 (April 2020). https://doi.org/10.1609/aaai.v34i07.6903
5. Yan, R., Peng, L., Xiao, S., Yao, G.: Primitive representation learning for scene text recognition. In: CVPR. pp. 284–293 (2021)
6. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)