Supplementary Material

Perception and Semantic Aware Regularization for Sequential Confidence Calibration

Table of Contents

We provide further details about experimental setting, as well as additional results on scene text recognition (STR) and speech recognition (SR) tasks. This supplementary material is structured as follows:

A.1 Detailed Description of Benchmark Datasets

We have used the following datasets in our experiments:

1. **Synth90K** [8]: This dataset contains 9 million synthetic text instance images composed of 90k common English words, which can simulate the distribution of scene text images and replace the real data set for deep learning algorithm training. Each image is annotated with a ground-truth word.
2. **SynthTex** [4]: This dataset contains $800,000$ images with 6 million synthetic text instances. Each image is annotated with a ground-truth word.
3. **IIIT5K** [13]: This dataset contains 5000 text instance images: 2000 for training and 3000 for testing. We use the provided test set as one of the test sets for our experiments and the provided train set for our validation.
4. **SVT** [23]: This dataset contains 350 images: 100 for training and 250 for testing. We use the provided test set as one of the test sets for our experiments and the provided train set for our validation.
5. **IC03** [12]: This dataset contains 509 images: 258 for training and 251 for testing. After discarding images that contain non-alphanumeric characters or less than three characters, it contains 867 instances of cropped text. We use the provided test set as one of the test sets for our experiments and the provided train set for our validation.
6. **IC13** [10]: This dataset contains 561 images: 420 for training and 141 for testing. After removing words with non-alphanumeric characters, the IC13 dataset contains 1051 cropped text instance images. We use the provided test set as one of the test sets for our experiments.
7. **IC15** [9]: This dataset contains 1500 images: 1000 for training and 500 for testing. We use the provided test set as one of the test sets for our experiments and the provided train set for our validation.
8. **SVTP** [18]: This dataset contains 238 images with 639 cropped text instances, which are mostly instances of distorted perspectives. We use this dataset as a test set.
9. **CUTE80** [19]: This dataset contains 80 high-resolution images with 288 cropped text instances, which are designed specifically for curvy text recognition. We use this dataset as a test set.

10. **RCTW** [20]: This dataset provides 12263 annotated Chinese text images from natural scenes. And 44420 text lines are exported from the training set for our experiment.

11. **ReCTS** [26]: This dataset provides 25000 annotated street-view Chinese text images, mainly derived from natural signboards. In the training set, 107657 cropped text samples are utilized for experiments.

12. **LSVT** [21]: This dataset is a large-scale Chinese and English scene text dataset, including 50000 full-labeled samples and 400000 partial-labeled samples. The full-labeled samples contain polygon boxes and text labels, while the partial-labeled samples contain only text instances. We only utilize the full-labeled training set and crop 243063 text line images for experiments.

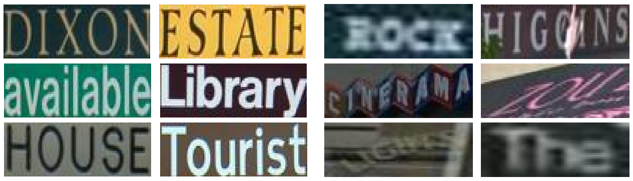
13. **ArT** [3]: This dataset contains text samples of various text layouts captured in a natural scene, such as rotated text and curved text. We obtain 49951 cropped text images from the training set and use them in our experiments.

14. **CTW** [25]: This dataset contains annotated 30000 street view images with rich diversity. We crop 191364 text lines from both the training and testing sets for experiments.

15. **AISHELL-1** [2]: This dataset is split into a training set, a development set, and a test set. The training set contains 120098 utterances from 340 speakers; the development set contains 14326 utterances from the 40 speakers; the Test set contains 7176 utterances from 20 speakers. And we use the training set for training, the test set for testing, and the development set for validation.

A.2 Construction of Hard and Easy Datasets

In this section, we provide a brief description of the construction of hard and easy datasets used in the main manuscript. As shown in Fig. 1, the visual feature quality of samples in IC03 dataset is relatively higher than the samples in SVTP dataset. The samples in IC03 are clear and easy to be recognized, while the samples in SVTP are blurred, noisy, incomplete, and thus relatively harder to be recognized. Therefore, we take the IC03 as the easy dataset and the SVTP as the hard dataset. And we construct a hardness-adjustable mixed dataset, where samples are derived from the easy and hard datasets, and thus the hardness is adjusted by changing the proportion of samples from easy and hard datasets.



(a) IC03 (b) SVTP

Figure 1. The illustration of IC03 and SVTP samples.

B. Comparison Techniques

In this section, we provide a brief description of the methods and hyper-parameter settings used in training for each comparison.

Brier Score [1] (BS): BS is defined as the squared loss between the one-hot target vector and the predicted probability vector.

Label Smoothing [22] (LS): LS softens the hard one-hot label with a smoothing parameter α as $y_k^{LS} = y_k(1-\alpha) + \alpha/K$, where y_k denotes the one-hot label for the k-th class and K is the class size. And we trained using $\alpha \in \{0.01, 0.05, 0.1\}$.

Focal Loss [14] (FL): FL is defined as $\text{FL} = -(1-p(y|x))^\gamma \log p(y|x)$, where γ is a hyper-parameter. And we trained using $\gamma \in \{1, 2, 3\}$.

Entropy Regularization [17] (ER): ER performs confidence calibration by directly penalizing the entropy of the predicted distribution, which is defined as $\text{ER} = \mathcal{L}_{\text{CE}} - \beta H(p(y|x))$, where β is a hyper-parameter. And We trained using $\beta \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$ as performed in [17].

Margin-based Label Smoothing [11] (MBLS): MBLS is defined as: $\mathcal{L}_{\text{CE}} + \lambda \sum_k \max(0, \max_j(l_j) - l_k - m)$, where λ and m are hyper-parameters. We set m to 10 following [11], and use $\lambda \in \{0.01, 0.05, 0.1, 0.2\}$.

Multi-class Difference in Confidence and Accuracy [5] (MDCA): As for MDCA, we trained on the following loss: FL + $\beta \cdot$ MDCA, where MDCA is defined in [5], and β is a hyper-parameter. We trained varying $\beta \in \{1, 5, 10, 15, 20, 25\}$ as performed in [5] and the γ used in FL is in $\{1, 2, 3\}$.

Graduated Label Smoothing [24] (GLS): The smoothing penalty α of GLS varies with the token's confidence. We set the α to 0.015 for tokens with confidence above 0.7, 0.0 for tokens with confidence below 0.3, and 0.005 for the remaining tokens.

Context-Aware Selective Label Smoothing [7] (CASLS): As for CASLS, we trained with $\alpha \in \{0.01, 0.05, 0.1\}$ following [7].

Perception and Semantic Aware Regularization (PSSR): As for PSSR, we set ε_e and ε_h to 0.01 and 1, respectively, and we trained with $\alpha \in \{0.1, 0.5, 1.0\}$.

According to the ECE attained on the validation set, we present the results of the models that perform the best with each of the aforementioned techniques.

C. Evaluation Metrics

In this section, we provide a brief description of the evaluation metrics involved in the main manuscript: ECE, ACE, and MCE. And the size of bins of ECE, ACE, and MCE is set to 15 following [7].

Expected Calibration Error [15] (ECE): ECE approximates the expected absolute difference between the predicted confidence and the accuracy of model. Given a finite number N of samples, the ECE cannot be directly computed using this definition. Instead, we partition the confidence range $[0, 1]$ into M equispaced bins, where i^{th} bin is the confidence interval in $(\frac{i-1}{M}, \frac{i}{M}]$. Let B_i represent the set of samples whose confidence falls into the i^{th} bin, and $|B_i|$ denote the number of samples in i^{th} bin. Further, the accuracy of B_i is defined as $A_i = \frac{1}{|B_i|}\sum_{j\in B_i}\mathbb{I}(\hat{Y}_j = Y_j)$, where $\mathbb{I}$ denotes indicator function. Similarly, the average confidence C_i of B_i is calculated as $C_i = \frac{1}{|B_i|}\sum_{j\in B_i}\mathbb{P}(\hat{Y}_j \mid X_j)$. And the ECE can be calculated as the weighted average of the absolute difference between the accuracy and confidence of each bin:

$$ECE = \sum_{i=1}^{M} \frac{B_i}{N}\Big|A_i - C_i\Big|. \quad (1)$$

Adaptive ECE [16] (ACE): The vanilla ECE is uniform bin width. In the case of a well-trained model, most of the samples are located within the high confidence ranges and thus dominate the value of the ECE. Therefore, we utilize Adaptive ECE (ACE), where bin sizes are determined so that samples are distributed equally throughout the bins:

$$ACE = \sum_{i=1}^{M} \frac{B_i}{N}\Big|A_i - C_i\Big|. \quad (2)$$

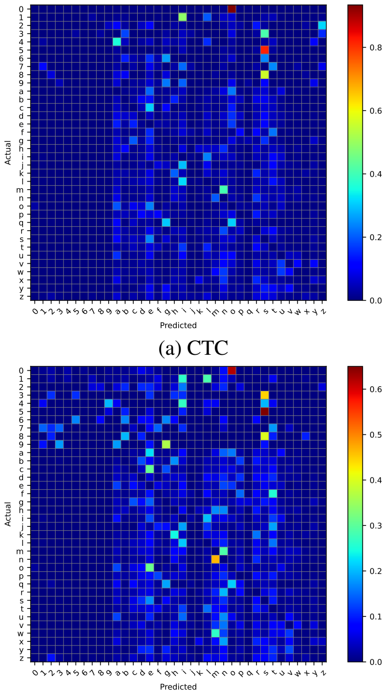
where $\forall_{i,j}, |B_i| = |B_j|$.

Maximum Calibration Error [5] (MCE): MCE is defined as the maximum absolute difference between the average accuracy and average confidence of each bin:

$$MCE = \max_{i\in 1,\ldots,M} \Big|A_i - C_i\Big|. \quad (3)$$

D. Confusion Matrices and Quantitative Metrics

The complete confusion matrices of the mispredictions are shown in Fig. 2.



(a) CTC

(b) Attention

Figure 2. The illustration of the complete confusion matrices of the mispredictions.

Frequency and Average probability: And the calculation of frequency (F_{vis}) and average probability (P_{vis}) mentioned in section 3.1 of the main manuscript are as follow:

$$F_{vis}(i-j) = \frac{N_j}{N_e} \quad (4)$$

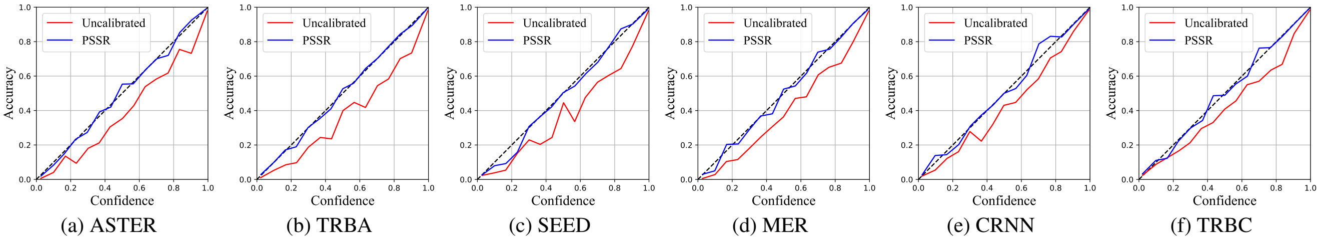
$$P_{vis}(i-j) = \frac{1}{N_j}\sum_{n=1}^{N_j} p(y_t = j|X) \quad (5)$$

where $F_{vis}(i-j)$ and $P_{vis}(i-j)$ refers to the frequency and average probability of a pair (i-j). N_j is the number of target token class i incorrectly predicted to class j, and N_e is the number of all incorrect predictions of class i. $p(y_t = j|X)$ refers to the probability that the prediction is class j while the target of y_t is i.

Perplexity: Perplexity (PPL) is the exponentiation of the average cross entropy of a corpus. The language models provide a probability distribution over full sentences or texts, and this makes PPL a natural assessment metric for these models. It is defined as:

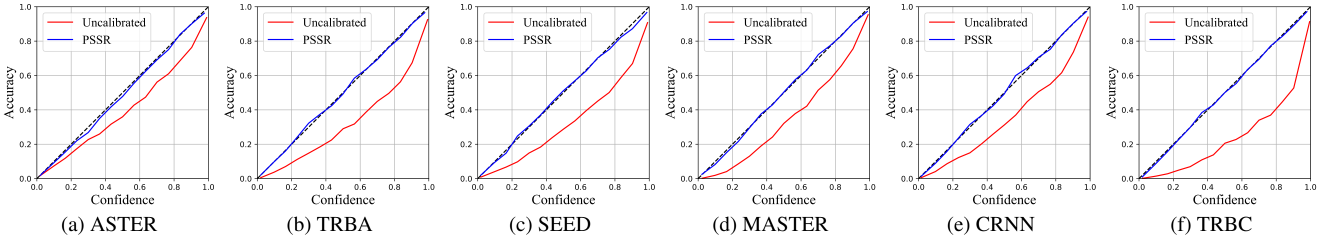
$$PPL = \exp(-\frac{1}{T}\sum_{t=1}^{T} p(x_t|x_{<t})) \quad (6)$$

where $x_{<t} = [x_0, \cdots, x_{t-1}]$.



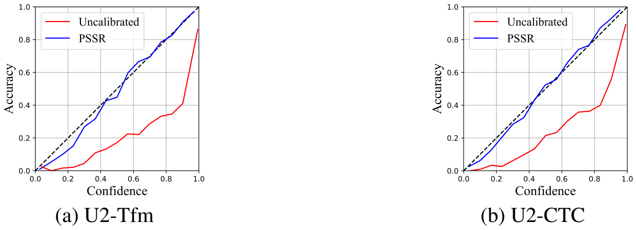
(a) ASTER (b) TRBA (c) SEED (d) MER (e) CRNN (f) TRBC

Figure 3. The reliability diagrams of the models calibrated with PSSR on English STR benchmark.



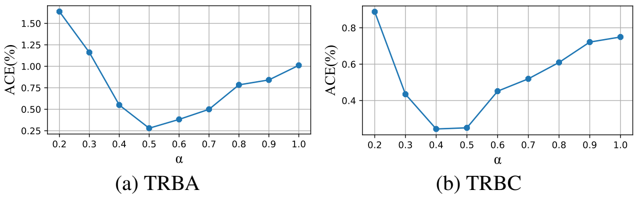
(a) ASTER (b) TRBA (c) SEED (d) MASTER (e) CRNN (f) TRBC

Figure 4. The reliability diagrams of the models calibrated with PSSR on Chinese STR benchmark.



(a) U2-Tfm (b) U2-CTC

Figure 5. The reliability diagrams of the models calibrated with PSSR on the AISHELL-1 dataset of speech recognition task.



(a) TRBA (b) TRBC

Figure 6. The effect of α on PSSR on TRBA and TRBC.

E.1 Ablation Study on the α

In the proposed PSSR, the main hyper-parameter is the calibration factor α, which controls the global calibration strength of all the samples. Here we conduct experiments by varying α from 0.2 to 1 to validate the effect of α for calibration performance. Fig. 6 shows how calibration performance is affected when we increase α in different model. We observe a general trend that the calibration error first decreases as α increases, and it achieves the best performance when α is close to 0.5 for TRBA and 0.4 for TRBC. As we continue to increase α, the calibration error starts to increase. This is because the calibration intensity is too high, which already leads to underconfident.

E.2 Additional Results on SR

The complete results of the different calibration methods on the speech recognition task on the AISHELL-1 dataset are presented in Tab. 1 and Tab. 2. We can observe that, regardless of the decoding schemes, the proposed PSSR outperforms other competitive calibration methods on ECE, ACE, and MCE metrics. This consolidates the thesis of this paper and further demonstrates the generalization of the proposed method.

Table 1. The calibration results of U2-Tfm on AISHELL-1. The best method is highlighted in bold.

Method	Acc	ECE	ACE	MCE
NLL	58.81	22.75	22.75	50.85
BS	58.84	23.25	23.25	49.79
LS	58.67	4.32	4.49	12.12
FL	58.12	18.86	18.86	36.95
ER	58.61	13.40	13.40	29.34
MBLS	58.65	4.17	4.20	10.36
MDCA	58.15	19.02	19.02	37.30
GLS	**59.18**	3.36	3.33	10.45
CASLS	58.92	2.91	3.00	8.85
PSSR	57.36	**2.21**	**2.06**	**7.01**

E.3 Reliability Diagrams on STR and SR benchmarks

In this section, we further investigate the calibration performance of the proposed method with reliability diagrams,

Table 2. The calibration results of U2-CTC on AISHELL-1. The best method is highlighted in bold.

Method	Acc	ECE	ACE	MCE
CTC	**58.14**	20.20	20.20	41.28
PSSR	57.44	**2.47**	**2.35**	**3.82**

whose results for STR and SR tasks of the different datasets are shown in Fig. 3, Fig. 4, and Fig. 5. The reliability diagram is the accuracy function of confidence. As for the perfectly calibrated model, the confidence of each bin ideally matches the corresponding accuracy, and thus its reliability diagram approximates a diagonal line (see the dashed line in Fig. 3, Fig. 4, and Fig. 5). In contrast, the curve of underconfident models lie mostly above the diagonal, while the curve of overconfident models lie mostly below the diagonal.

E.4 Additional Results of Combining with Temperature Scaling

In this section, we investigate the calibration performance of combining temperature scaling with various training methods including ours (PSSR). In Tab. 3 and Tab. 4, we show the results of TRBA and TRBC on the English and Chinese STR benchmark. And we can observe that our proposed method achieves the best performance across different models and datasets. Furthermore, the optimal temperature value of our method is very close to 1, which indicates that the models calibrated by our method are already nearly perfectly calibrated.

E.5 Additional Results under Dataset Shift

Data distribution shift is prevalent in sequence recognition tasks. For example, the scene text recognition models are usually trained on synthetic data, while applied to real scene data. Therefore, maintaining performance even under dataset shift is necessary for sequence recognition calibration methods.

To evaluate the calibration performance under dataset shift, we construct four corruption version of English STR benchmark following [6]. The [6] presents four types algorithmically generated corruptions: noise, blur, weather, and digital categories. And we select a representative corruption from each of the above four types corruptions as follow:

- **Speckle noise**. Speckle noise is a type of additive noise, and the noise added to a pixel tends to be larger if the original pixel intensity is larger.
- **Gaussian blur**. Gaussian blur is a low-pass filter in which a blurred pixel is the result of a weighted average of its neighbors, with more distant pixels being given less weight in this average.

Table 3. ECE(%) for different methods with pre- and post-temperature scaling. Optimal T is indicated in brackets. The best method is highlighted in bold.

Method	Dataset	TRBA		TRBC	
		PreT	PosT	PreT	PosT
Uncalibrated	English	3.88	1.16 (1.3)	2.73	0.90 (1.2)
BS		3.44	0.98 (1.3)	-	-
LS		1.59	0.93 (1.1)	-	-
FL		1.36	1.36 (1.0)	-	-
ER		1.31	1.31 (1.0)	-	-
MBLS		1.34	1.34 (1.0)	-	-
MDCA		1.5	1.0 (1.1)	-	-
GLS		0.92	0.92 (1.0)	-	-
CASLS		1.02	1.02 (1.0)	-	-
PSSR		**0.36**	**0.36 (1.0)**	**0.47**	**0.47 (1.0)**
Uncalibrated	Chinese	10.78	6.83 (1.5)	15.02	3.72 (1.5)
BS		10.18	3.40 (1.3)	-	-
LS		1.25	1.25 (1.0)	-	-
FL		9.74	3.57 (1.3)	-	-
ER		3.42	3.42 (1.0)	-	-
MBLS		1.29	1.29 (1.0)	-	-
MDCA		9.97	4.15 (1.2)	-	-
GLS		1.31	1.31 (1.0)	-	-
CASLS		1.4	1.4 (1.0)	-	-
PSSR		**0.72**	**0.72 (1.0)**	**0.79**	**0.79 (1.0)**

Table 4. ACE(%) for different methods with pre- and post-temperature scaling. Optimal T is indicated in brackets. The best method is highlighted in bold.

Method	Dataset	TRBA		TRBC	
		PreT	PosT	PreT	PosT
Uncalibrated	English	3.88	0.81(1.3)	2.71	0.73(1.2)
BS		3.42	0.72(1.3)	-	-
LS		1.52	0.74(1.1)	-	-
FL		0.99	0.99(1.0)	-	-
ER		1.10	1.10(1.0)	-	-
MBLS		1.16	1.16(1.0)	-	-
MDCA		1.44	0.85(1.1)	-	-
GLS		0.90	0.90(1.0)	-	-
CASLS		0.98	0.98(1.0)	-	-
PSSR		**0.28**	**0.28(1.0)**	**0.25**	**0.25(1.0)**
Uncalibrated	Chinese	10.78	6.89(1.5)	15.02	3.73(1.5) (1.5)
BS		10.18	3.24(1.3)	-	-
LS		1.23	1.23(1.0)	-	-
FL		9.74	3.57(1.3)	-	-
ER		3.35	3.35(1.0)	-	-
MBLS		1.18	1.18(1.0)	-	-
MDCA		9.97	4.11(1.2)	-	-
GLS		1.27	1.27(1.0)	-	-
CASLS		1.40	1.40(1.0)	-	-
PSSR		**0.63**	**0.63(1.0)**	**0.73**	**0.73(1.0)**

- **Spatter**. Spatter can occlude the lens in the form of rain or mud.
- **Saturate**. Saturate is common in edited images, in which case the image becomes more or less colorful.

Table 5. The calibration results comparison of Uncalibrated, BS, LS, FL, ER, MBLS, MDCA, GLS, CASLS, and PSSR on the English STR benchmark for four corruption. The accuracy and three calibration metrics: Acc(%), ECE(%), ACE(%) and MCE(%), are listed. The best method is highlighted in bold.

Corruption	Method	TRBA				MASTER				TRBC			
		Acc	ECE	ACE	MCE	Acc	ECE	ACE	MCE	Acc	ECE	ACE	MCE
Speckle Noise	Uncal.	65.71	3.80	3.85	15.83	63.45	3.37	3.37	13.43	65.63	1.51	1.46	**7.59**
	LS	65.82	1.57	1.57	7.22	**66.10**	2.64	2.60	9.29	-	-	-	-
	FL	66.95	0.84	0.68	**4.47**	65.37	2.49	2.72	11.10	-	-	-	-
	ER	67.08	1.10	0.76	6.19	65.60	1.80	1.83	8.24	-	-	-	-
	MBLS	66.76	1.27	1.25	6.14	65.99	2.12	2.12	7.14	-	-	-	-
	MDCA	67.35	1.09	0.78	5.22	65.37	2.04	2.20	12.57	-	-	-	-
	GLS	67.31	3.03	2.95	6.11	63.23	3.26	3.26	14.26	-	-	-	-
	CASLS	**67.48**	1.36	1.22	6.64	66.07	1.72	1.49	11.20	-	-	-	-
	PSSR	67.01	**0.64**	**0.66**	5.84	65.89	**1.24**	**1.31**	**5.34**	**66.45**	**1.19**	**0.54**	9.26
Gaussian Blur	Uncal.	42.10	19.10	19.10	57.63	41.41	17.78	17.78	51.60	**40.52**	14.49	14.50	44.80
	LS	43.18	12.73	12.74	40.55	42.48	10.19	10.19	26.99	-	-	-	-
	FL	42.78	9.71	9.84	38.07	42.03	6.85	6.99	24.82	-	-	-	-
	ER	43.14	3.32	3.17	**9.49**	**43.27**	3.54	3.69	11.06	-	-	-	-
	MBLS	43.47	14.11	14.14	43.52	42.64	10.84	10.84	30.77	-	-	-	-
	MDCA	43.17	13.91	13.91	44.24	41.95	8.06	8.16	28.27	-	-	-	-
	GLS	42.53	12.95	12.99	31.17	42.56	8.94	8.94	22.95	-	-	-	-
	CASLS	**43.62**	14.79	14.79	47.36	42.22	12.75	12.75	32.47	-	-	-	-
	PSSR	42.29	**2.45**	**2.55**	10.92	41.76	**1.82**	**1.95**	**8.44**	40.50	**1.45**	**1.25**	**7.80**
Spatter	Uncal.	59.91	4.12	4.12	11.79	59.80	6.54	6.54	16.09	58.12	2.23	1.89	6.41
	LS	59.96	1.83	1.72	8.05	61.58	1.66	1.64	7.40	-	-	-	-
	FL	60.63	1.89	1.30	7.00	61.20	1.19	1.09	4.27	-	-	-	-
	ER	59.82	1.09	1.05	**2.74**	61.55	1.55	1.15	6.17	-	-	-	-
	MBLS	60.15	1.56	1.46	5.04	61.73	1.41	1.53	9.14	-	-	-	-
	MDCA	60.38	1.74	1.55	7.73	62.20	1.48	1.36	8.13	-	-	-	-
	GLS	60.02	2.60	2.06	7.82	**62.85**	2.13	1.93	4.60	-	-	-	-
	CASLS	60.55	1.28	1.00	5.32	62.01	1.07	1.20	4.51	-	-	-	-
	PSSR	**61.68**	**1.06**	**0.86**	4.89	62.57	**0.95**	**0.88**	**3.01**	**58.82**	**1.99**	**1.87**	**5.13**
Saturate	Uncal.	81.04	3.95	3.95	16.96	79.99	4.34	4.34	22.19	80.41	2.56	2.48	17.56
	LS	81.03	2.00	1.79	9.25	80.82	1.58	1.59	5.58	-	-	-	-
	FL	81.52	1.09	0.87	4.59	79.93	1.25	0.65	7.23	-	-	-	-
	ER	81.56	1.20	0.85	8.19	80.84	1.00	0.67	**4.68**	-	-	-	-
	MBLS	81.24	1.20	1.10	7.28	80.78	1.15	0.96	6.52	-	-	-	-
	MDCA	**81.97**	1.16	0.91	**4.51**	80.69	1.12	0.88	6.97	-	-	-	-
	GLS	81.95	2.91	2.45	11.74	80.79	2.91	2.89	8.60	-	-	-	-
	CASLS	81.80	1.09	0.59	8.47	80.00	1.07	0.91	6.12	-	-	-	-
	PSSR	81.56	**0.74**	**0.54**	7.07	**82.11**	**0.62**	**0.53**	6.37	**80.92**	**0.64**	**0.38**	**6.78**

In Tab. 5, we show the complete experiment results of data shift. It can be found that for ECE and ACE metrics, our method shows superior performance compared to other methods.

References

[1] Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950. 2

[2] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. AISHELL-1: an open-source mandarin speech corpus and a speech recognition baseline. In *20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment*, pages 1–5, Seoul, Korea, 2017. IEEE Computer Society. 2

[3] Chee Kheng Chng, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, Chee Seng Chan, Lianwen Jin, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, Chuan-Ming Fang, Shuaitao Zhang, and Junyu Han. ICDAR2019 robust reading challenge on arbitrary-shaped text - rrc-art. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1571–1576. IEEE, 2019. 2

[4] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2315–2324, Las Vegas, 2016. IEEE Computer Society. 1

[5] Ramya Hebbalaguppe, Jatin Prakash, Neelabh Madan, and Chetan Arora. A stitch in time saves nine: A train-time regularizing loss for improved neural network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16081–16090, 2022. 2, 3

[6] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 5

[7] Shuangping Huang, Yu Luo, Zhenzhou Zhuang, Jin-Gang Yu, Mengchao He, and Yongpan Wang. Context-aware selective label smoothing for calibrating sequence recognition model. In *MM '21: ACM Multimedia Conference*, pages 4591–4599. ACM, 2021. 2, 3

[8] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition, 2014. 1

[9] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman K. Ghosh, Andrew D. Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. ICDAR 2015 competition on robust reading. In *13th International Conference on Document Analysis and Recognition*, pages 1156–1160, Nancy, France, 2015. IEEE Computer Society. 1

[10] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernández Mota, Jon Almazán, and Lluís-Pere de las Heras. ICDAR 2013 robust reading competition. In *12th International Conference on Document Analysis and Recognition*, pages 1484–1493, Washington, DC, USA, 2013. IEEE Computer Society. 1

[11] Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. The devil is in the margin: Margin-based label smoothing for network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 80–88, 2022. 2

[12] Simon M. Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, Robert Young, Kazuki Ashida, Hiroki Nagai, Masayuki Okamoto, Hiroaki Yamamoto, Hidetoshi Miyao, JunMin Zhu, WuWen Ou, Christian Wolf, Jean-Michel Jolion, Leon Todoran, Marcel Worring, and Xiaofan Lin. ICDAR 2003 robust reading competitions: entries, results, and future directions. *International Journal on Document Analysis & Recognition*, 7(2-3):105–122, 2005. 1

[13] Anand Mishra, Karteek Alahari, and C. V. Jawahar. Top-down and bottom-up cues for scene text recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2687–2694, Providence, RI, USA, 2012. IEEE Computer Society. 1

[14] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip HS Torr, and Puneet K Dokania. Calibrating deep neural networks using focal loss. *arXiv preprint arXiv:2002.09437*, 2020. 2

[15] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2901–2907, Austin, Texas, USA, 2015. AAAI Press. 3

[16] Khanh Nguyen and Brendan O'Connor. Posterior calibration and exploratory analysis for natural language processing models. *arXiv preprint arXiv:1508.05154*, 2015. 3

[17] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017. 2

[18] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *IEEE International Conference on Computer Vision*, pages 569–576, Sydney, Australia, 2013. IEEE Computer Society. 1

[19] Anhar Risnumawan, Palaiahnakote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, 41(18):8027–8048, 2014. 1

[20] Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge J. Belongie, Shijian Lu, and Xiang Bai. ICDAR2017 competition on reading chinese text in the wild (RCTW-17). In *14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 9-15, 2017*, pages 1429–1434. IEEE, 2017. 2

[21] Yipeng Sun, Dimosthenis Karatzas, Chee Seng Chan, Lianwen Jin, Zihan Ni, Chee Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, and Jingtuo Liu. ICDAR 2019 competition on large-scale street view text with partial labeling - RRC-LSVT. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1557–1562. IEEE, 2019. 2

[22] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, Las Vegas, NV, USA, 2016. IEEE Computer Society. 2

[23] Kai Wang, Boris Babenko, and Serge J. Belongie. End-to-end scene text recognition. In Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc Van Gool, editors, *IEEE International Conference on Computer Vision*, pages 1457–1464, Barcelona, Spain, 2011. IEEE Computer Society. 1

[24] Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. On the inference calibration of neural machine translation. *arXiv preprint arXiv:2005.00963*, 2020. 2

[25] Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, and Shi-Min Hu. A large chinese text dataset in the wild. *J. Comput. Sci. Technol.*, 34(3):509–521, 2019. 2

[26] Rui Zhang, Mingkun Yang, Xiang Bai, Baoguang Shi, Dimosthenis Karatzas, Shijian Lu, C. V. Jawahar, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, and Minghui Liao. ICDAR 2019 robust reading challenge on reading chinese text on signboard. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1577–1581. IEEE, 2019. 2