Perception and Semantic Aware Regularization for Sequential Confidence Calibration

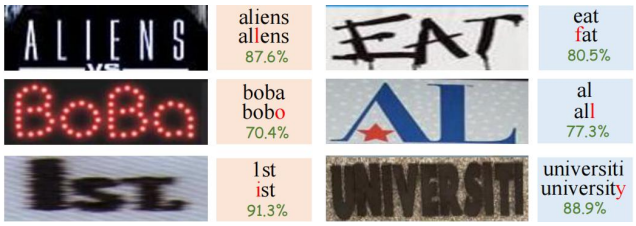
Zhenghua Peng[1], Yu Luo[1], Tianshui Chen[2], Keke Xu[1], Shuangping Huang[1,3,*]
[1]South China University of Technology, [2]Guangdong University of Technology, [3]Pazhou Laboratory
eepzh@mail.scut.edu.cn, luoyurl@126.com, tianshuichen@gmail.com,
eexkk@mail.scut.edu.cn, eehsp@scut.edu.cn

Abstract

Deep sequence recognition (DSR) models receive increasing attention due to their superior application to various applications. Most DSR models use merely the target sequences as supervision without considering other related sequences, leading to over-confidence in their predictions. The DSR models trained with label smoothing regularize labels by equally and independently smoothing each token, reallocating a small value to other tokens for mitigating overconfidence. However, they do not consider tokens/sequences correlations that may provide more effective information to regularize training and thus lead to sub-optimal performance. In this work, we find tokens/sequences with high perception and semantic correlations with the target ones contain more correlated and effective information and thus facilitate more effective regularization. To this end, we propose a Perception and Semantic aware Sequence Regularization framework, which explore perceptively and semantically correlated tokens/sequences as regularization. Specifically, we introduce a semantic context-free recognition and a language model to acquire similar sequences with high perceptive similarities and semantic correlation, respectively. Moreover, over-confidence degree varies across samples according to their difficulties. Thus, we further design an adaptive calibration intensity module to compute a difficulty score for each samples to obtain finer-grained regularization. Extensive experiments on canonical sequence recognition tasks, including scene text and speech recognition, demonstrate that our method sets novel state-of-the-art results. Code is available at https://github.com/husterpzh/PSSR.

1. Introduction

Deep neural networks (DNNs) have shown remarkable performance in sequence recognition tasks, such as scene text recognition (STR) [11, 40, 47] and speech recognition (SR) [3, 25]. Despite impressive accuracy, recent studies have indicated that DNNs [7, 8, 19], including deep sequence recognition (DSR) models, are usually poorly calibrated and tend to be overconfident [15, 27, 28]. In the sense that the confidence values associated with the predicted labels are higher than the true likelihood of the correctness of these labels, even for the wrong predictions, the overconfident DSR models may assign high confidences. This property may lead to potentially disastrous consequences for many safety-critical applications, such as autonomous driving [12] and medical diagnosis [26, 35].



(a) Perception overconfidence (b) Semantic overconfidence

Figure 1. Text strings placed along the right side of images are target, prediction, and sequence confidence respectively from top to bottom. Fig. 1 (a): the model assigns higher confidence to the character that extremely resembles to the ground-truth character in visual perception (*e.g.*, texture and topological shape); Fig. 1 (b): the word that are semantically correlated to the ground-truth label will be predicted with a high confidence.

Current DSR models use merely the target sequence as supervision and consider little information about any other sequences. Thus, they may tend to blindly give an overconfident score for their predictions, leading to the overconfidence dilemma. Presently, some works [13, 58] introduce label smoothing, which smooth each token by reallocating a small value to all non-target token class from the target class, to prevent the DSR models from assigning the full probability mass to target sequences. However, these algorithms do not consider token/sequences correla-

*Corresponding author.

tions, and are difficult to provide effective and sufficient information. In this work, we find that tokens/sequences with high perception or semantic correlations, which refer to tokens/sequences with high visual/auditory similarities and with high co-occurrence similarities respectively, may be mistakenly given a highly-confident score. Taking STR for example, the Figure 1 shows that token "l" shares highly visual similarity with "i", and thus the models may easily predict it to "i" with high confidence. On the other hand, word "universiti" is semantically similar to word "university" and thus it is also predicted to "university". These tokens/sequences are easily ambiguous with the target ones and thus may provide more effective information to regularize training.

In this paper, we propose a calibration method for DSR models: **P**erception and **S**emantic aware **S**equence **R**egularization (PSSR). The PSSR enables the DSR models with stronger vital perception discrimination ability and richer semantic contextual correlation knowledge by incorporating additional perception similarity and semantic correlation information into training. Specifically, we construct a similar sequence set that comprises sequences either perception similar to the instantiated sequence input or semantic correlated with the target text sequence. During the training stage, these similar sequences are used as weighted additional supervision signals to offer more perception similarity of different token classes and semantic correlation in the same context. Furthermore, we discover that the degree of overconfidence of the model on its predictions varies across samples and is related to the hardness of recognizing samples. Hence, we further introduce a modulating factor function to adjust the calibration among different samples adaptively. To evaluate the effectiveness of the proposed method, we conduct experiments on two canonical sequence recognition tasks, including scene text recognition and speech recognition. Experimental results demonstrate that our method is superior to the state-of-the-art calibration methods across different benchmarks.

The major contributions of this paper are fourfold. First, we discovered the overconfidence of DSR models comprises perception overconfidence and semantic overconfidence. Second, following our observations, we propose a calibration method for DSR models that enables the DSR models with more vital perception discrimination ability and richer semantic contextual correlation knowledge, so as to obtain more calibrated predictions. Third, we introduce a modulating factor function to achieve adaptive calibration. Fourth, we provide comprehensive experiments over multiple sequence recognition tasks with various network architectures, datasets, and in/out-domain settings. We also verify its effectiveness on the downstream application active learning. The results suggest that our method yields substantial improvements in DSR models calibration.

2. Related Works

Sequence Recognition. Sequence recognition generally involves dealing with instantiated sequential data, which usually carries rich information on perception and semantic modalities. Previous methods, such as segmentation [10,49] and CTC-based methods [2, 14, 50], predict the sequence mainly depending on the perception feature of tokens, hardly taking semantic information into consideration. For example, the CTC-based model splits the input sequence into several vertical pixel frames and outputs per-frame predictions, which are purely based on the perception features of the corresponding frame at each time step. Recent works increasingly pay attention to conjointly exploiting both perception and semantic information [11, 60, 64], since the two types of information complement each other in the recognition process. Some works implicitly incorporate the semantic correlation to the models using RNNs with attention [3, 30, 51, 59] or Transformers [54, 62]. Additionally, [25, 33, 47] explicitly integrates a language model to learn semantics for supervision. Although remarkable progress has been achieved in the public benchmarks, we discover that it meanwhile incurs a problem, that is, these state-of-the-art methods are biased towards the commonly-seen perception pattern or the semantic context in the training set and produce overconfident predictions [55].

Confidence Calibration. Calibration of scalar classification has been extensively studied for a long time [29, 41, 46, 61]. A simple yet efficient manner is post-hoc calibration, which directly rescales the prediction confidence of already trained models to the calibrated confidence during the inference stage [15, 22, 42, 46]. While showing favorable effectiveness for in-domain samples calibration, they fail to be applied under the condition of dataset shift, since a held-out dataset is required to learn the re-calibration function [27]. As another prevalent line of research, several studies calibrate networks by modifying the training process during the training stage [4]. Label smoothing [53], originally proposed as a regularization technique, has shown a favorable effect on model calibration [44]. [44] and [37] fix overconfidence from the perspective of maximizing the entropy of the prediction distribution. More recently, Liu *et al*. argue that Label smoothing pushes all the logit distances to zero and lead to a non-informative solution, and propose a margin-based Label smoothing to realize better calibration [32]. [17] developed an auxiliary loss function that calibrates the entire confidence distribution in a multi-class setting.

The aforementioned methods mostly focus on the improvement and analysis of scalar classification task. However, almost little literature is proposed to study the calibration for DSR models calibration [20, 52]. Slossberg *et al*. simply extend the temperature scaling for scene text recognition calibration, which rescale the logits on

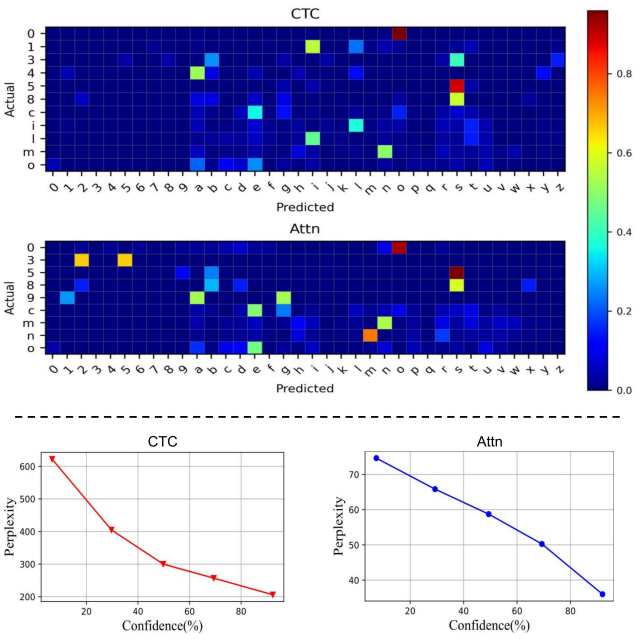


Figure 2. The upper part illustrates the confusion matrix of the mispredictions, which represents the distribution that the actual tokens in a sequence are recognized as other classes. And the bottom part plots the correlation between perplexity and the confidence of sequence.

each time-step individually with a specific temperature value [52]. However, it essentially calibrates individual tokens to achieve the calibration of a sequence, unaware of perception and semantic correlation in calibration of sequences. Huang *et al*. achieve the adaptive calibration on each token with taking the contextual dependency underlying the sequence [20]. Despite of showing a certain extent effectiveness, the method is insufficient for DSR models calibration, since only the inter-token context is considered, the potential cause of the overconfident prediction brought by the overfiting on the perception features are ignored.

3. What Causes Overconfidence of DSR models?

In this section, we delve into reasons for the observed overconfidence of DSR models and identify that the perception similarity and semantic correlation of sequence are responsible for the phenomenon. All the statistics are derived from the prediction results of a CTC-based model (NRNC [1]), and an attention-based model (MASTER [33]) on the ensembled testing set [21, 23, 24, 34, 36, 45, 48, 57].

3.1. Perception Similarity

To study how the perception information influences the miscalibration of DSR models, we build confusion matrices to count the frequency that the ground-truth class is recognized as other classes. The upper part of Fig. 2 displays part of the confusion matrices (see appendix for the complete confusion matrices), from which we can observe that the ground-truth class is more likely to be confused with the classes with higher perception similarity, that is, these classes suffer from more severe overconfidence. For example, the ground-truth token "0" is almost exclusively confused with "*o*" in either attention or CTC models.

Table 1 further presents the quantitative metrics, including the frequency (F_{vis}) and the average probability (P_{vis}) of the ground-truth token being confused to most perception similar token (see appendix for detailed calculation of the two metrics). As shown, a similar token owns a relatively high frequency and probability to be mispredicted. Note that, due to the data bias of the training set, where the proportion of letters is much larger than numbers, perception overconfidence mainly occurs in the letter-related classes. For example, the F_{vis} in Table 1 show that, the letter "*o*" is seldom predicted as number "0".

3.2. Semantic Correlation

We additionally compute the perplexity of the misprediction of models to measure how well the predicted sequences are formed (see appendix for details of perplexity). In general, the lower perplexity score represents that the prediction has a stronger semantic correlation. As shown in the bottom part of Fig 2, the semantically correlated mispredictions with lower perplexity scores demonstrate a more severe overconfidence problem. Another interesting observation is that, although all the models tend overconfident in wrong predictions, the perplexity varies from the models based on different decoders. Compared with the CTC-based models that rely more on visual information, the context-aware attention-based models generally have lower perplexity scores. The phenomenon indicates that introducing the semantic information during training makes the model tend to predict legitimate sequences in the training set.

4. Proposed Methodology

4.1. Preliminaries

Let $\{(X_i, Y_i)\}_{i=1}^{N} \in \mathcal{D}(\mathcal{X}, \mathcal{Y})$ denotes a dataset, where $X_i \in \mathcal{X}$ is a sequential input sequence (*e.g*. text image, speech audio, etc), and $Y_i = \{y_{i,1}, y_{i,2}, ..., y_{i,n_i}\} \in \mathcal{Y}$ is the corresponding target sequence consisting of multiple tokens. Let $\mathbb{P}(\tilde{Y}|X_i)$ denotes the posterior probability that a sequence recognition network predicts for a candidate sequence $\tilde{Y}$ on the given input X_i. And the predicted sequence is obtained as $\hat{Y}_i = \text{argmax}_{\tilde{Y} \in \mathcal{Y}} \mathbb{P}(\tilde{Y}|X_i)$ with its confidence as $\mathbb{P}(\hat{Y}_i|X_i) = \max_{\tilde{Y} \in \mathcal{Y}} \mathbb{P}(\tilde{Y}|X_i)$. Generally, the DSR model are said to be perfectly calibrated when, for

Table 1. The frequency (F_{vis}) and the average probability (P_{vis}) of ground-truth token being confused to most visually similar token

	Pair	0-o	1-i	3-s	4-a	5-s	8-s	c-e	i-l	l-i	m-n	o-0
CTC	F_{vis}	95.92	55.10	40.74	52.17	88.89	57.14	35.80	36.88	45.11	50.20	4.72
	P_{vis}	84.01	75.02	60.92	78.79	76.87	76.57	68.10	70.86	69.11	66.23	65.13
	Pair	0-o	3-2	3-5	5-s	8-s	9-a	9-g	c-e	m-n	n-m	o-0
Attn	F_{vis}	70.00	50.00	50.00	72.73	44.44	40.00	40.00	37.33	40.75	56.64	2.58
	P_{vis}	70.85	68.13	96.40	81.63	81.42	41.71	90.81	75.06	73.79	74.66	85.25

each sample $(X_i, Y_i) \in \mathcal{D}(\mathcal{X}, \mathcal{Y})$:

$$\mathbb{P}(\hat{Y}_i = Y_i \mid \mathbb{P}(\hat{Y}_i|X_i)) = \mathbb{P}(\hat{Y}_i|X_i). \tag{1}$$

4.2. Sequence-level Calibration

The vanilla training process of the DSR model adopts one-hot encoding that places all the probability mass in one target sequence and thus encourages the probability of the target sequence being biased toward one-hot distribution. This myopic training algorithm may be useful for recognition accuracy, but it ignores the perception similarity between different token classes and various semantic contextual correlations. This lack of knowledge makes the model predict recklessly without considering various conditions. To alleviate this problem, we attempt to incorporate additional information into the training stage, which comprises the perception similarity information between different token classes and more semantic contextual correlations.

Specifically, we construct a similarity sequence set that comprises sequences either perception similar to the sequence instance inside the input sequence or semantic correlated with the corresponding target sequence. And we introduce a regularization term to the vanilla loss to smooth the empirical loss over these similar sequences. Formally, the entire loss is defined as:

$$\mathcal{L}_i^{total} = \mathcal{L}_G(Y_i, \hat{Y}_i) + \alpha\ f(p_i) \sum_{Y_i' \in \mathcal{S}(X_i, Y_i)} \mathcal{L}_G(Y_i', \hat{Y}_i) \tag{2}$$

where $\mathcal{L}_G$ refers to the empirical loss function (*e.g.*, cross-entropy and CTC loss) generally used in DSR models of different decoding mechanisms, α is a hyperparameter used for adjusting the global calibration intensity, $f(p_i)$ is an adaptive calibration intensity function which is used for local adjustment of calibration intensity among different samples (see Sec. 4.4 for more details), and $\mathcal{S}(X_i, Y_i)$ is the similarity sequence set consisting of perception similarity and semantic correlation sequences of sample (X_i, Y_i).

Most previous calibration methods for DSR models are implemented at the token-level, which require a token-to-token alignment relationship between input and output sequence and is therefore limited to the partial decoders (*e.g.*, attention). In contrast, our proposed loss is computed among different sequences, which can avoid the complicated alignment strategies operated on token-level and thus can be applied to different decoding schemes.

4.3. Similar Sequence Mining

In this section, we describe how to obtain the similar sequence set, which consists of perception similarity and semantic correlation sequences

Perception Similarity Sequences. The prediction distribution of DSR models is affected by both perception and semantic contextual features. Thus, the critical challenge of effectively modeling the perception similar between sequences is to eliminate the effect of semantic context. Hence, we resort to the semantic context-free model (*e.g.*, CTC-based model). Specifically, we first fed the input sequences X_i into a well-trained CRNN [50] model to obtain the probability matrix consisting of token prediction distribution at each time step. Then, we can calculate the posterior probability $\mathbb{P}(\tilde{Y}|X_i)$ of any candidate sequence $\tilde{Y}$ over the entire sequence space. Benefiting from the context-free attribute, the higher the probability of a candidate sequence, the higher its perception similarity to the input sequence. Thus, we conduct a search algorithm based on the probability matrix to rank the posterior probability among the sequence space and finally collect the top *N* probable sequences as the perception similarity sequences.

Semantic Correlation Sequences. Recently, [56] discovered models tend to assign high probabilities to sequences that share a highly similar context to the target sequence and appear more frequently in training, even if these sequences obviously deviate from the perception feature of the input sequence. Here, we define them as semantic correlation sequences of the target sequence. And we search for these sequences with the help of a pre-trained language model BCN [11], which is a variant of transformer decoder with a diagonal attention mask to prevent the model from attending to the current time-step token of the target sequence. As a result, the token distribution at each time step is conditioned on its bidirectional context, that is $P(y_t|y_{1:t-1}, y_{t+1:n})$. In this setting, we can efficiently model the correlation between tokens and their contexts. Specifically, a higher probability for a certain token class means that the semantics of the combination of this token

class with its context is stronger, i.e., this combination appears more frequently in training. Then, we multiply the probability of each token together as the probability of semantic context correlation between the candidate sequence and the target sequence. Similarly, we perform a search algorithm to rank the probability of sequences in the entire sequence space and collect the top N probable sequences as the semantic correlation sequences of the target sequence.

4.4. Hardness ranking adaptive calibration

The models differ in the degree of overconfidence of their predictions on different samples, with more or less. Applying the identical calibration intensity to each sample may result in underconfident in some samples while may still be overconfident in others, which makes it challenging to achieve co-calibration. To analyze the claim, we take STR as an example and construct a dataset with adjustable hardness property (see appendix for details). We compare the calibration performance of TRBA [1] and TRBC [1] on the dataset with different hardness ratios. As shown in Fig. 3 (b), the ECE values all increase with increasing hardness ratio, indicating that the models become more overconfident. One reason for this may be that the confidence of target sequences continuously increase when the model is trained with hard label, irrespective of the fact that the actual posterior probabilities of target sequences of difficult samples should be low intuitively. And the training process only leads to the predicted confidence scores become further greater than the actual probabilities.

Following our observation, and inspired by the *Focal loss* [31] that views the posterior probability of the target class as a measure of the sample hardness (i.e. the lower the probability, the harder the sample), we propose a modulating factor function $f(p_i)$ that is integrated into the regularization term to achieve adaptive calibration based on sample hardness. It is defined as:

$$f(p_i) = \varepsilon_e + (\varepsilon_h - \varepsilon_e)(1 - p_i)^2 \tag{3}$$

where ε_e and ε_h are the hyper-parameters that control the calibration intensity for the easiest and hardest samples ($\varepsilon_h \geq \varepsilon_e$), respectively, and p_i is the posterior probability of the target sequence (i.e. $\mathbb{P}(Y_i|X_i)$). When the sample is hard to recognize and the p_i is small, the result of $f(p_i)$ is close to the ε_h, so that more probability is smoothed from the target sequence towards similar sequences, and vice versa. In this work, we set ε_e and ε_h are 0.01 and 1.0, respectively.

5. Experiment

5.1. Experimental Setup

We evaluate our method on the two classic sequence recognition tasks: scene text recognition (STR) and speech recognition (SR). Detailed settings are described below.

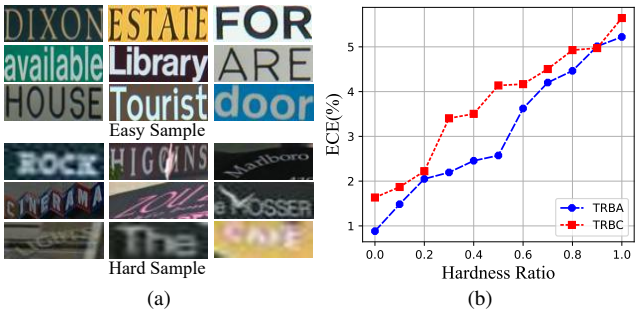


Figure 3. The illustration of: **(a)** easy and hard recognition samples; **(b)** the ECE results of TRBA and TRBC on different degrees of hardness of dataset.

Datasets: For STR, we conduct the experiments on the English and Chinese benchmarks: 1) The English benchmark contains two synthetic datasets for training, i.e., Synth90K [21] and SynthText [16], and the ensemble of seven realistic datasets for testing, including IIIT5K [36], SVT [57], IC03 [34], IC13 [24], IC15 [23], SVTP [45], and CUTE80 [48]. 2) The Chinese benchmark [6] ensembles five public datasets, consisting of 509,164 and 63,645 images for training and testing, respectively. For SR, we use the AISHELL-1 [5], which is a large-scale mandarin speech dataset containing 141,600 sentences with 120,098 for training, 14326 for validation, and 7,176 for testing.

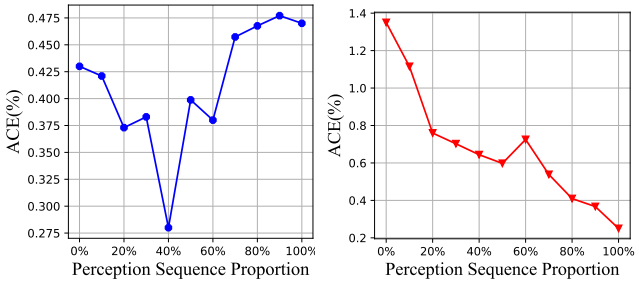
Models: For STR, we adopt six models, including ASTER [51], TRBA [1], SEED [47], MASTER [33], CRNN [50], and TRBC [1], which cover the advanced and classical attention-based and CTC-based models. For SR, we use U2-Tfm [63] and U2-CTC [63], which use a shared Comformer [43] encoder with self-attention and CTC as two branch decoders.

Evaluation Metrics: We adopt the widely used expected calibration error (ECE) [38], adaptive ECE (ACE) [39], maximum calibration error (MCE) [17], and reliability diagram [9] as calibration metrics. Following [20], these metrics are calculated by taking the entire sequence as a unit to calculate the sequence-level confidence and accuracy.

Comparison Methods: We compare our method with SOTA scalar and sequential calibration methods. Specifically, scalar calibration methods, including Brier Score (BS) [4], Label Smoothing (LS) [53], Focal Loss (FL) [37], Entropy Regularization (ER) [44], Margin-based Label Smoothing (MBLS) [32], and MDCA [17], are extended to sequence recognition by applying them to each token. In addition, sequential calibration methods, including Graduated Label Smoothing (GLS) [58], Context-Aware Selective Label Smoothing (CASLS) [20], are adopted for comparison. However, the two methods are limited to attention-based models due to the utilization of one-hot encoding.

Table 2. How hardness ranking adaptive calibration affects the sequence recognition calibration. The best method is highlighted in bold.

Method	TRBA			TRBC		
	ECE	ACE	MCE	ECE	ACE	MCE
PSSR w/o $f(p_i)$	0.74	0.93	8.97	1.19	0.85	10.65
PSSR	**0.36**	**0.28**	**3.99**	**0.47**	**0.25**	**6.22**



(a) TRBA (b) TRBC

Figure 4. The results of different perception similar sequence proportions in similarity sequence set on TRBA and TRBC models.

5.2. Ablation Study

We conduct ablation studies to discuss and analyze the actual contribution of each component. All the experiments are conducted on the English STR benchmark. TRBA and TRBC are adopted to validate the effectiveness of the proposed method on the sequence recognition models of attention and CTC decoders, respectively.

5.2.1 Effect of Adaptive Calibration

As discussed in Sec. 4.4, the hardness of recognizing a sample plays an important role in the calibration performance. We remove the component of hardness ranking $f(p_i)$ from the PSSR in Eq. 2, and show the comparison results of the PSSR with and without hardness ranking component in Table 2. From the results, the resulting method suffers from a severe performance drop on all the metrics across TRBA and TRBC models. The calibration performance is more evident in the TRBA model, where the ECE, ACE, and MCE are increased by 0.38%, 0.65%, and 4.98%, respectively.

5.2.2 Effect of Similar Sequence Set

The similar sequence set comprises both perception and semantic similar sequences. Here, we explore how the different combined proportions of these two kinds of sequences affect the calibration performance. The results are shown in the Fig. 4. As for TRBA, the calibration performance is better when the number of visually similar sequences is approximately equal to that of semantically similar sequences. However, the TRBC is better with the increase of visually similar sequences. This confirms our claim above that the CTC-based models, such as TRBC, mainly occurring perception overconfidence, while overconfidence in attention-based models derives from both perception overconfidence and semantic overconfidence.

5.3. Comparison with State-of-the-arts

In this section, we compare the proposed method against the state-of-the-art method on the two tasks: scene text recognition (STR) and speech recognition (SR).

5.3.1 Results on STR

We present the quantitative calibration results of attention-based models on the English STR benchmark in Table 3. The results show that our proposed PSSR outperforms all the compared state-of-the-art methods across all the models in terms of ECE, ACE, and MCE metrics. Among other comparison methods, the two calibration methods for sequential data, GLS and CASLS, generally perform better than the methods for scalar data and achieve the second-best performances. Compared with the sub-optimal GLS method, particularly in the TRBA model, the proposed method still reduces 0.56%, 0.62%, and 3.18% in ECE, ACE, and MCE, respectively. Moreover, Table 4 reports the calibration results of CTC-based models on the English STR benchmark. Compared with the uncalibrated models trained with CTC loss, the models trained with PSSR perform much better in terms of all the metrics, including accuracy, ECE, ACE, and MCE. Combined with the above, the satisfying performance demonstrates that the proposed method can be well adapted to the model with different decoding schemes.

We further verify the effectiveness of our method on the Chinese STR benchmark, and the calibration results of attention and CTC models are presented in Table 5 and 6, respectively. Notably, our PSSR outperforms other approaches and sets a new state-of-the-art with better accuracy and confidence calibration on almost all the models.

5.3.2 Results on SR

Table 7 reports the calibration results of uncalibrated models and PSSR on the AISHELL-1 dataset. As shown, the proposed PSSR performs better than uncalibrated models in ECE, ACE, and MCE metrics. Compared to uncalibrated models, the proposed PSSR reduces 20.54%, 20.69%, 43.84% in terms of ECE, ACE, and MCE on the attention-based model and reduces 17.73%, 17.85%, 37.46% in terms of ECE, ACE, and MCE on the CTC-based model. More results are presented in the appendix.

Table 3. The calibration results comparison of NLL, BS, LS, FL, ER, MBLS, MDCA, GLS, CASLS and PSSR on the English STR benchmark of attention-based models. The accuracy and three calibration metrics: Acc(%), ECE(%), ACE(%) and MCE(%), are listed. The best method is highlighted in bold.

Method	ASTER				TRBA				SEED				MASTER			
	Acc	ECE	ACE	MCE	Acc	ECE	ACE	MCE	Acc	ECE	ACE	MCE	Acc	ECE	ACE	MCE
NLL	**85.27**	3.82	3.82	17.10	85.51	3.88	3.88	21.49	85.34	4.04	4.04	23.09	84.52	3.86	3.86	16.01
BS [4]	85.17	3.46	3.41	16.53	86.06	3.44	3.42	23.72	85.20	4.14	4.14	21.23	85.83	3.26	3.26	16.17
LS [53]	84.35	0.99	0.81	10.27	84.12	1.59	1.52	10.38	84.62	1.23	1.20	9.61	85.16	1.37	1.32	8.11
FL [37]	84.94	1.79	1.40	9.55	85.34	1.36	0.99	11.04	85.89	2.23	2.24	16.01	84.86	1.22	0.97	7.37
ER [44]	76.33	7.25	7.21	23.85	85.64	1.31	1.10	9.18	85.50	1.07	0.95	13.73	85.09	1.40	1.02	10.86
MBLS [32]	84.42	1.12	1.03	7.63	84.51	1.34	1.16	9.47	84.55	1.39	1.38	10.22	85.01	1.03	1.05	**5.72**
MDCA [17]	85.09	2.18	2.14	10.70	85.98	1.50	1.44	7.85	**86.08**	2.54	2.47	20.58	84.92	1.25	0.82	6.70
GLS [58]	84.12	0.93	0.71	6.36	83.83	0.92	0.90	7.17	85.13	1.26	1.11	11.24	85.05	2.66	2.64	11.76
CASLS [20]	84.65	0.86	0.77	5.55	85.41	1.02	0.98	7.94	85.71	1.59	1.36	13.15	84.89	1.16	0.93	12.20
PSSR	85.06	**0.69**	**0.48**	**5.26**	**86.45**	**0.36**	**0.28**	**3.99**	85.54	**0.94**	**0.77**	**7.48**	**86.03**	**0.78**	**0.40**	8.36

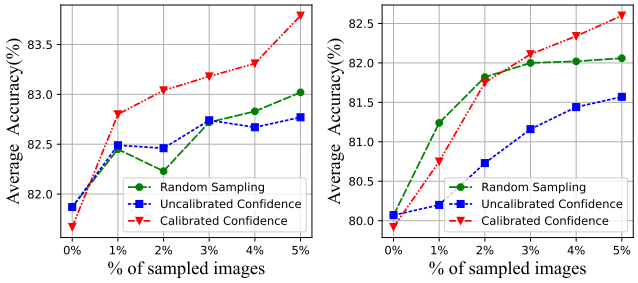
Table 4. The calibration results of CTC-based models on the English STR benchmark. The best method is highlighted in bold.

Method	CRNN				TRBC			
	Acc	ECE	ACE	MCE	Acc	ECE	ACE	MCE
CTC	78.91	2.80	2.80	14.45	84.94	2.73	2.71	16.62
PSSR	**79.53**	**0.97**	**0.49**	**8.52**	**85.48**	**0.47**	**0.25**	**6.22**

5.4. Calibration Performance under Dataset Shift

The DNNs are discovered to be overconfident and highly uncalibrated under the condition of data shift. Inspired by [18], the data distribution drift test datasets are derived from the English benchmark test dataset after four diverse corruption types, including speckle noise, Gaussian blur, spatter, and saturate. Figure 5 shows the clean and the four corrupted examples. Table 8 reports the calibration results of uncalibrated models and PSSR on the English STR benchmark of corrupted datasets, which demonstrates that the model trained with the proposed PSSR can still achieve a good calibration performance even under data shift. And compared with the state-of-the-art calibration methods, our method performs best in terms of all the metrics across all the drift datasets. And the details on other methods and their corrupted calibration results are presented in the appendix.

Figure 5. Clean and four corruption examples.



(a) TRBA (b) TRBC

Figure 6. The results of active learning task on TRBA and TRBC.

5.5. Downstream Application

We argue that calibration benefits the downstream active learning task when adopting a confidence-based query strategy. In general, active learning trains an initial model based on a small amount of labeled data, and then a query strategy is applied to the output of models to select the most informative samples with the least confidence for an oracle to annotate. The model is then retrained with the additional labeled data. The above process will be repeated until model accuracy is satisfied or the labeling resource is exhausted.

The active learning experiment is conducted on the English STR benchmark, where an attention-based TRBA model and a CTC-based TRBC model are adopted. Specifically, only 10% of training samples are used initially to train the base model. Then, 1% samples of the unlabeled data pool (consisting of the remaining 90% of training samples) are queried, combining the original labeled samples to retrain the model. We compare three query strategies: random sampling, least uncalibrated confidence, and confidence calibrated with our PSSR. And the querying process is iterated five times.

Fig. 6 shows the average accuracy on the test set against

Table 5. The calibration results comparison of NLL, BS, LS, FL, ER, MBLS, MDCA, GLS, CASLS and PSSR on the Chinese STR benchmark of attention-based models. The accuracy and three calibration metrics: Acc(%), ECE(%), ACE(%) and MCE(%), are listed. The best method is highlighted in bold.

Method	ASTER				TRBA				SEED				MASTER			
	Acc	ECE	ACE	MCE	Acc	ECE	ACE	MCE	Acc	ECE	ACE	MCE	Acc	ECE	ACE	MCE
NLL	56.12	6.69	6.14	16.00	56.23	10.78	10.78	27.14	42.09	11.27	11.27	26.55	61.28	9.01	9.01	21.19
BS [4]	56.18	6.14	5.58	16.03	56.82	10.18	10.18	25.34	**44.15**	10.41	10.41	24.55	65.06	8.77	8.77	20.88
LS [53]	56.31	1.95	1.53	4.71	56.16	1.25	1.23	4.18	42.54	1.31	1.34	4.21	65.28	1.33	1.33	3.22
FL [37]	55.98	5.73	5.19	12.95	56.78	9.74	9.74	24.58	43.02	8.69	8.72	21.45	64.04	2.96	2.96	7.43
ER [44]	55.66	3.62	3.42	6.45	55.40	3.42	3.35	7.59	42.70	3.70	3.71	11.74	63.53	2.39	2.39	5.66
MBLS [32]	**56.32**	1.96	1.52	5.15	56.26	1.29	1.18	**2.29**	42.22	1.25	1.19	**2.96**	65.66	1.13	1.13	3.31
MDCA [17]	56.06	5.63	5.08	13.01	**56.85**	9.97	9.97	26.74	43.43	9.68	9.69	22.78	64.12	3.02	3.02	8.93
GLS [58]	56.16	1.38	1.15	**2.83**	56.38	1.31	1.27	3.35	41.54	1.16	1.18	3.62	64.89	1.54	1.46	4.82
CASLS [20]	56.10	1.40	1.05	2.96	56.18	1.40	1.40	3.41	41.45	1.27	1.15	3.34	64.78	1.50	1.42	4.46
PSSR	55.91	**1.02**	**0.58**	3.14	56.55	**0.72**	**0.63**	**2.29**	41.64	**1.01**	**0.86**	2.99	**65.86**	**1.03**	**0.93**	**2.11**

Table 6. The calibration results of CTC-based models on the Chinese STR benchmark. The best method is highlighted in bold.

Method	CRNN				TRBC			
	Acc	ECE	ACE	MCE	Acc	ECE	ACE	MCE
CTC	**41.10**	8.62	8.62	21.85	**58.07**	15.02	15.02	39.80
PSSR	40.44	**0.64**	**0.48**	**3.38**	57.25	**0.79**	**0.73**	**1.78**

Table 7. The calibration results of U2-Tfm and U2-CTC on AISHELL-1. The best method is highlighted in bold.

Models	Method	Acc	ECE	ACE	MCE
U2-Tfm	NLL	**58.81**	22.75	22.75	50.85
	PSSR	57.36	**2.21**	**2.06**	**7.01**
U2-CTC	CTC	**58.14**	20.20	20.20	41.28
	PSSR	57.44	**2.47**	**2.35**	**3.82**

Table 8. Corrupted calibration results on the English STR benchmark. Uncal is short for Uncalibrated. The best method is highlighted in bold.

Corruption	Method	TRBA				TRBC			
		Acc	ECE	ACE	MCE	Acc	ECE	ACE	MCE
Speckle Noise	Uncal	65.71	3.80	3.85	15.83	65.63	1.51	1.46	**7.59**
	PSSR	**67.01**	**0.64**	**0.66**	**5.84**	**66.45**	**1.19**	**0.54**	9.26
Gaussian Blur	Uncal	42.10	19.10	19.10	57.63	40.52	14.49	14.50	44.80
	PSSR	**42.29**	**2.45**	**2.55**	**10.92**	**40.50**	**1.45**	**1.25**	**7.80**
Spatter	Uncal	59.91	4.12	4.12	11.79	58.12	2.23	1.89	6.41
	PSSR	**61.68**	**1.06**	**0.86**	**4.89**	**58.82**	**1.99**	**1.87**	**5.13**
Saturate	Uncal	81.04	3.95	3.95	16.96	80.41	2.56	2.48	17.56
	PSSR	**81.56**	**0.74**	**0.54**	**7.07**	**80.92**	**0.64**	**0.38**	**6.78**

the percentage of images sampled from the unlabeled data pool for different models. It can be seen that the accuracy using the confidence-based strategy performs better than other query strategies. And it further outperforms the uncalibrated confidence-based strategy with accuracy improvement by 1.02% and 1.03% after the final iteration on TRBA and CRNN, respectively.

6. Conclusion

Despite the superior performance of deep sequence recognition models, they have been proven to suffer from the over-confidence dilemma. In this paper, we investigate the overconfidence problem of the DSR model and discover that tokens/sequences with higher perception and semantic correlations to the target ones contain more sufficient and correlated information to supervise the regularization of labels and facilitate more effective regularization. Motivated by the observation, we propose a Perception and Semantic aware Sequence Regularization framework, which explores perceptively and semantically correlated tokens/sequences as regularization. Comprehensive experiments are conducted on classic DSR tasks: scene text and speech recognition, and our method achieves state-of-the-art confidence calibration performance. In the future, we will explore more effective strategies to conjointly utilize perception and semantic information for better DSR model calibration.

Acknowledgment

This research was supported in part by NSFC (Grant No. 62176093, 61673182, 62206060), Key Realm R&D Program of Guangzhou (No. 202206030001), Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515012282) and Guangdong Provincial Key Laboratory of Human Digital Twin (No. 2022B1212010004).

References

[1] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwal-

suk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *International Conference on Computer Vision*, pages 4714–4722, 2019. 3, 5

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR*, San Diego, CA, USA, 2015. 2

[3] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016*, pages 4945–4949. IEEE, 2016. 1, 2

[4] Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950. 2, 5, 7, 8

[5] 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 *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment*, pages 1–5, 2017. 5

[6] Jingye Chen, Haiyang Yu, Jianqi Ma, Mengnan Guan, Xixi Xu, Xiaocong Wang, Shaobo Qu, Bin Li, and Xiangyang Xue. Benchmarking chinese text recognition: Datasets, baselines, and an empirical study. *CoRR*, abs/2112.15093, 2021. 5

[7] Tianshui Chen, Liang Lin, Riquan Chen, Xiaolu Hui, and Hefeng Wu. Knowledge-guided multi-label few-shot learning for general image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1371–1384, 2022. 1

[8] Tianshui Chen, Tao Pu, Hefeng Wu, Yuan Xie, Lingbo Liu, and Liang Lin. Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9887–9903, 2022. 1

[9] Morris Degroot and Stephen Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983. 5

[10] Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai. Pixellink: Detecting scene text via instance segmentation. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 6773–6780. AAAI Press, 2018. 2

[11] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 7098–7107. Computer Vision Foundation / IEEE, 2021. 1, 2, 4

[12] Di Feng, Lars Rosenbaum, Claudius Gläser, Fabian Timm, and Klaus Dietmayer. Can we trust you? on calibration of a probabilistic object detector for autonomous driving. *arXiv*, 1909.12358, 2019. 1

[13] Yingbo Gao, Weiyue Wang, Christian Herold, Zijian Yang, and Hermann Ney. Towards a better understanding of label smoothing in neural machine translation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 212–223, Seattle, USA, 2020. Association for Computational Linguistics. 1

[14] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML*, volume 148, pages 369–376, Pittsburgh, Pennsylvania, USA, 2006. ACM. 2

[15] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 1, 2

[16] 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. 5

[17] 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, 5, 7, 8

[18] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 7

[19] Hongxiang Huang, Daihui Yang, Gang Dai, Zhen Han, Yuyi Wang, Kin-Man Lam, Fan Yang, Shuangping Huang, Yongge Liu, and Mengchao He. Agtgan: Unpaired image translation for photographic ancient character generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5456–5467, 2022. 1

[20] 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, 5, 7, 8

[21] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition, 2014. 3, 5

[22] Byeongmoon Ji, Hyemin Jung, Jihyeun Yoon, Kyungyul Kim, and Younghak Shin. Bin-wise temperature scaling (BTS): improvement in confidence calibration performance through simple scaling techniques. In *IEEE/CVF International Conference on Computer Vision Workshops*, pages 4190–4196, Seoul, Korea, 2019. IEEE Computer Society. 2

[23] 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. 3, 5

[24] 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. 3, 5

[25] Shigeki Karita, Nelson Enrique Yalta Soplin, Shinji Watanabe, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. In Gernot Kubin and Zdravko Kacic, editors, *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1408–1412. ISCA, 2019. 1, 2

[26] Young-Min Kim and Tae-Hoon Lee. Korean clinical entity recognition from diagnosis text using BERT. *BMC Medical Informatics and Decision Making*, 20-S(7):242, 2020. 1

[27] Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. Calibrated language model fine-tuning for in-and out-of-distribution data. *arXiv preprint arXiv:2010.11506*, 2020. 1, 2

[28] Aviral Kumar and Sunita Sarawagi. Calibration of encoder decoder models for neural machine translation. *arXiv preprint arXiv:1903.00802*, 2019. 1

[29] Christian Leibig, Vaneeda Allken, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. 2016. 2

[30] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. Show, attend and read: A simple and strong baseline for irregular text recognition. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI*, pages 8610–8617. AAAI Press, 2019. 2

[31] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, pages 2999–3007, Venice, Italy, 2017. IEEE Computer Society. 5

[32] 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, 5, 7, 8

[33] Ning Lu, Wenwen Yu, Xianbiao Qi, Yihao Chen, Ping Gong, Rong Xiao, and Xiang Bai. MASTER: multi-aspect non-local network for scene text recognition. *Pattern Recognit.*, 117:107980, 2021. 2, 3, 5

[34] 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. 3, 5

[35] Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging*, 39(12):3868–3878, 2020. 1

[36] 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. 3, 5

[37] 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, 5, 7, 8

[38] 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. 5

[39] Khanh Nguyen and Brendan O'Connor. Posterior calibration and exploratory analysis for natural language processing models. *arXiv preprint arXiv:1508.05154*, 2015. 5

[40] Nguyen Nguyen, Thu Nguyen, Vinh Tran, Minh-Triet Tran, Thanh Duc Ngo, Thien Huu Nguyen, and Minh Hoai. Dictionary-guided scene text recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 7383–7392. Computer Vision Foundation / IEEE, 2021. 1

[41] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, volume 119 of *ACM International Conference Proceeding Series*, pages 625–632. ACM, 2005. 2

[42] Kanil Patel, William Beluch, Bin Yang, Michael Pfeiffer, and Dan Zhang. Multi-class uncertainty calibration via mutual information maximization-based binning. *arXiv preprint arXiv:2006.13092*, 2020. 2

[43] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 367–376, 2021. 5

[44] 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, 5, 7, 8

[45] 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. 3, 5

[46] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999. 2

[47] Zhi Qiao, Yu Zhou, Dongbao Yang, Yucan Zhou, and Weiping Wang. SEED: semantics enhanced encoder-decoder framework for scene text recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 13525–13534, Seattle, USA, 2020. Computer Vision Foundation / IEEE. 1, 2, 5

[48] 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. 3, 5

[49] David Rybach, Christian Gollan, Ralf Schlüter, and Hermann Ney. Audio segmentation for speech recognition using segment features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan*, pages 4197–4200. IEEE, 2009. 2

[50] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(11):2298–2304, 2017. 2, 4, 5

[51] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. ASTER: an attentional scene text recognizer with flexible rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2035–2048, 2019. 2, 5

[52] Ron Slossberg, Oron Anschel, Amir Markovitz, Ron Litman, Aviad Aberdam, Shahar Tsiper, Shai Mazor, Jon Wu, and R Manmatha. On calibration of scene-text recognition models. *arXiv preprint arXiv:2012.12643*, 2020. 2, 3

[53] 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, 5, 7, 8

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, volume 30, pages 5998–6008, Long Beach, CA, USA, 2017. Curran Associates, Inc. 2

[55] Zhaoyi Wan, Jielei Zhang, Liang Zhang, Jiebo Luo, and Cong Yao. On vocabulary reliance in scene text recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 11422–11431. Computer Vision Foundation / IEEE, 2020. 2

[56] Zhaoyi Wan, Jielei Zhang, Liang Zhang, Jiebo Luo, and Cong Yao. On vocabulary reliance in scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11425–11434, 2020. 4

[57] 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. 3, 5

[58] Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. On the inference calibration of neural machine translation. *arXiv preprint arXiv:2005.00963*, 2020. 1, 5, 7, 8

[59] Xiao Yang, Dafang He, Zihan Zhou, Daniel Kifer, and C. Lee Giles. Learning to read irregular text with attention mechanisms. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*, pages 3280–3286. ijcai.org, 2017. 2

[60] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Towards accurate scene text recognition with semantic reasoning networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12110–12119. Computer Vision Foundation / IEEE, 2020. 2

[61] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the 18th International Conference on Machine Learning (ICML 2001)*, pages 609–616, Williamstown, MA, USA, 2001. Morgan Kaufmann. 2

[62] Binbin Zhang, Di Wu, Chao Yang, Xiaoyu Chen, Zhendong Peng, Xiangming Wang, Zhuoyuan Yao, Xiong Wang, Fan Yu, Lei Xie, and Xin Lei. Wenet: Production first and production ready end-to-end speech recognition toolkit, 2021. 2

[63] Binbin Zhang, Di Wu, Zhuoyuan Yao, Xiong Wang, Fan Yu, Chao Yang, Liyong Guo, Yaguang Hu, Lei Xie, and Xin Lei. Unified streaming and non-streaming two-pass end-to-end model for speech recognition. *CoRR*, abs/2012.05481, 2020. 5

[64] Xinyun Zhang, Binwu Zhu, Xufeng Yao, Qi Sun, Ruiyu Li, and Bei Yu. Context-based contrastive learning for scene text recognition. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 3353–3361. AAAI Press, 2022. 2