GLaLT: Global-Local Attention-Augmented Light Transformer for Scene Text Recognition

Hui Zhang, *Member, IEEE*, Guiyang Luo, Jian Kang, Shan Huang, Xiao Wang, *Senior Member, IEEE*, and Fei-Yue Wang, *Fellow, IEEE*

***Abstract*— Recent years have witnessed the growing popularity of connectionist temporal classification (CTC) and attention mechanism in scene text recognition (STR). CTC-based methods consume less time with few computational burdens, while they are not as effective as attention-based methods. To retain computational efficiency and effectiveness, we propose the global–local attention-augmented light Transformer (GLaLT), which adopts a Transformer-based encoder–decoder structure to orchestrate CTC and attention mechanism. The encoder integrates the self-attention module with the convolution module to augment the attention, where the self-attention module pays more attention to capturing long-term global dependencies and the convolution module focuses on local context modeling. The decoder consists of two parallel modules: one is the Transformer-decoder-based attention module and the other is the CTC module. The first one is removed in the testing phase and can guide the second one to extract robust features in the training phase. Extensive experiments on standard benchmarks demonstrate that GLaLT achieves state-of-the-art performance for both regular and irregular STR. In terms of tradeoffs, the proposed GLaLT is at or near the frontiers for maximizing speed, accuracy, and computational efficiency at the same time.**

***Index Terms*— Attention mechanism, connectionist temporal classification (CTC), scene text recognition (STR).**

NOMENCLATURE

I	Input image.
Y	Encoder-length vector/ The input of attention-augmented encoder.
Y_i	Input of the encoder block i.
O_i	Output of the encoder block i.
G	Ground truth label sequence.
S	Output sequence produced by the encoder.
C	Ground-truth character set.
M	Sequence-to-sequence mapping function.
T	Length of the sequence.
Q	Queries in the MSAP.
K	Keys in the MSAP.
V	Values in the MSAP.
d_{in}	Embedding size of the encoder.
$T \times d_k^h$	Dimension of queries in the MSAP.
$T \times d_v^h$	Dimension of values in the MSAP.
$T \times d_v$	Output dimension of MSAP.
N_e	Number of encoder blocks.
N_d	Number of decoder blocks.
h	Number of attention heads.
L_{ctc}	CTC loss.
L_{attn}	Attention loss.
α	Tunable parameter for balancing L_{ctc} and L_{attn}.

I. INTRODUCTION

SCENE text recognition (STR) is the task of recognizing character sequences in natural scenes and plays an essential role in various real-world applications such as information retrieval [1], [2], image understanding [3], [4], vehicle license plate recognition [5], [6], information security [7], [8], and automatic driving [9], [10], [11]. STR has been widely studied in both academia and industry, which has achieved significant progress [12], [13], [14]. However, it is still challenging to perform efficient and effective text recognition, specifically for uneven lighting conditions, complex background, low contrast, high occlusion, and heavy perspective distortion.

The encoder–decoder structure is gaining momentum in STR, and in the light of the decoding mechanism, there exist two main approaches, i.e., connectionist temporal classification (CTC) and attention-based methods. Shi et al. [15] introduce the CTC mechanism into the image-based sequence recognition first [see Fig. 1(a)]. Hu et al. [16] use a graph convolutional network to model the sequence correlations and

Manuscript received 1 November 2022; revised 9 January 2023; accepted 17 January 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62203040 and Grant 62102041, and in part by the National Key Research and Development Program of China under Grant 2021YFB1600402. *(Corresponding author: Xiao Wang.)*

Hui Zhang is with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China (e-mail: huizhang1@bjtu.edu.cn).

Guiyang Luo is with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: luoguiyang@bupt.edu.cn).

Jian Kang is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China, and also with China Telecom, Beijing 100033, China (e-mail: kangj30@chinatelecom.cn).

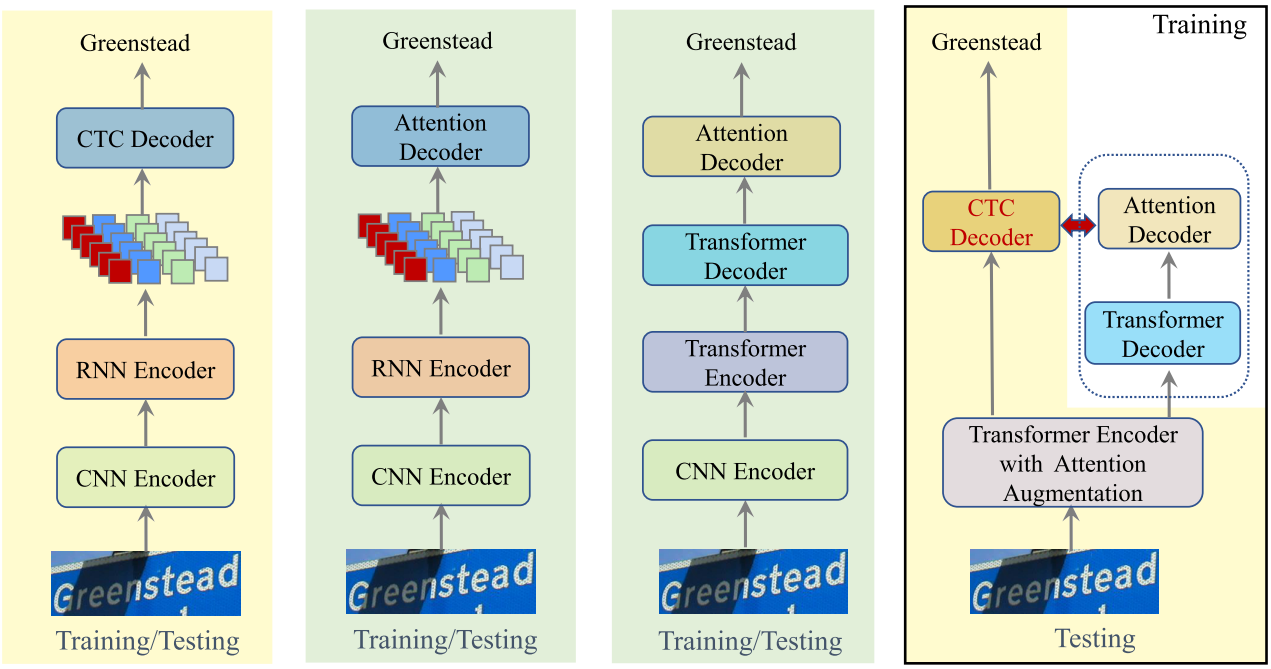
Shan Huang is with the Department of Information Security, Tencent, Beijing 100193, China (e-mail: lattehuang@tencent.com).

Xiao Wang is with the School of Artificial Intelligence, Anhui University, Hefei 230093, China (e-mail: x.wang@ia.ac.cn).

Fei-Yue Wang is with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: feiyue@ieee.org).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3239696.

Digital Object Identifier 10.1109/TNNLS.2023.3239696



(a) (b) (c) (d)

Fig. 1. Typical framework and our proposed GLaLT for STR. (a) Encoder–decoder structure with CTC mechanism. (b) Encoder–decoder structure with the attention mechanism. (c) Transformer-based methods, which are also based on the attention mechanism. (d) Our proposed GLaLT (in the inference stage, the part inside the dotted line can be removed).

achieve a more effective CTC-based model by learning from guidance. The encoder–decoder structure with the attention mechanism was first proposed for natural language processing (NLP) tasks such as machine translation [17], [18], image captioning [19], and dialog system [20]. Subsequently, it was devoted to achieving alignment between text sequence and image local region [21] in STR. Since then, a growing number of attention-based methods [see Fig. 1(b)] have emerged [22], [23], [24]. Luo et al. [22] propose a multiobject rectified attention network to deal with various shapes and distorted patterns of irregular text. Wan et al. [23] use two separate character classification and position prediction branches to predict the class and geometry information. Recently, the Transformer-based structure is introduced into text recognition for capturing long-distance context information [25], [26], shown in Fig. 1(c).

Generally, CTC-based methods assume that the character order in the label and that in the image is monotonous, and thereby design a series of many-to-one rules to align the dense framewise output with the label. On the other hand, attention-based methods use a parameterized attention decoder to align and transcribe each character on the image [27]. They predict text according to the features in the previous time step by recurrent neural network (RNN). This causes heavy computational resources and impedes decoding in parallel, which greatly slows down the inference time. However, attention-based methods have been demonstrated to achieve higher recognition accuracy. On the contrary, CTC-based methods consume less time and few computational burdens. They are not as effective as attention-based methods, since CTC loss misguides the training of feature representations. In this article, we intend to take advantage of both the CTC-based and attention-based methods and put much more attention on balancing accuracy, speed, and efficiency.

To overcome the drawbacks of CTC and make full use of the parallel-computing self-attention mechanism, we present the global–local attention-augmented light transformer, named GLaLT. The encoder of the GLaLT uses stacked encoder blocks, each of which consists of multihead self-attention, convolution module, and two feedforward modules. The multihead self-attention focuses on long-term global relationship modeling while the convolution module concentrates on local context modeling. The decoder consists of two parallel modules: one is the Transformer-decoder-based attention module and the other is the CTC module. The CTC and attention modules share the encoder of the GLaLT. When making predictions, the CTC module assumes that the probability of occurrence of each character in the sentence is a conditional-independent event. CTC usually misrecognizes characters that are difficult to distinguish. Therefore, additional language models are sometimes required to assist in prediction. While the attention module is data-driven and predicts the next character based on the input and previous characters, which implies a language model. However, when there is noise in the data, the attention module performs poorly. Meanwhile, if the input sequence is long, the attention module is difficult to learn in the early training stage. Since the CTC's forward–backward algorithm can enforce the monotonic alignment of input and output, stable alignment can be obtained even when the input data are too noisy. In the early stage of training, even if the attention module has not learned how to align correctly, the CTC module can help the encoder to learn in a monotonically aligned manner, thereby accelerating the convergence of the entire network. In the inference stage, we simply use the encoder and the CTC module, which remove the heavy Transformer-decoder-based attention module, achieving an attention-augmented light Transformer method.

Our contributions can be summarized as follows.

1) A novel GLaLT method is designed for STR, which puts much emphasis on balancing the tradeoff between recognition accuracy and inference efficiency. The inference process simply consists of an attention-augmented encoder and a CTC decoder so that the text recognizer is

with more computation parallelization and fewer computing resources.
2) The proposed global–local range attention (GLRA) is adopted to integrate the self-attention module with the convolution module, where the convolution module focuses on local context modeling while the self-attention module pays more attention to capturing long-term global dependencies.
3) The decoder is a hybrid CTC and attention module to improve accuracy and accelerate learning. The attention module can provide strong and effective guidance in training the CTC module within the multiloss constraint framework.

The remainder of this article is organized as follows. In Section II, we present a brief introduction to related works. The technical details of GLaLT are presented in Section III. Section IV details experimental procedures and results over public datasets. The conclusion is drawn in Section V.

II. Related Work

In recent years, a series of well-known text recognition methods have been constructed in natural scenes. According to the characteristics of these methods, they can be roughly divided into three categories: character-based recognition, word-based recognition, and sequence-based recognition.

The character-based recognition first detects each character by leveraging connected components [28] or sliding window [29], and then conducts classification using hand-engineered features or learned features. After that, the individual characters are incorporated into the whole word by means of dynamic programming or other heuristic algorithms [30], [31]. Inspired by the success of the deep convolutional neural network in visual understanding, Liu et al. [32] incorporate unsupervised feature learning with deep neural network to achieve an accurate character recognizer. For postprocessing, the character responses with character spacings, the beam search algorithm, or the weighted finite state transducer-based representation are applied to recognize target words in a defined lexicon.

Word-based recognition is another alternative approach for text recognition. Goel et al. [33] recognize the text in the image by matching the scene and synthetic image features with the proposed weighted dynamic time warping approach. Rodriguez-Serrano et al. [34] use the structured support vector machine (SVM) framework [35], [36], [37] to embed word labels and word images into a common space and cast the text recognition problem as retrieving the closest word label in this space. Almazán et al. [38] propose to address the spotting and recognition tasks by creating a common embedding space for word images and text strings. This is extended in [39] where Gordo leverages character bounding box annotations on a small set of training images to learn local mid-level features and aggregates that to produce a global word image signature. Jaderberg et al. [40] formulate text recognition as a classification problem in a large lexicon of $90k$ possible words. Benefiting from highly realistic and sufficient synthetic data [41], [42], [43], they propose a CNN text recognizer and regress all the characters simultaneously.

As for recent methods, they tend to address text recognition in a sequence-to-sequence manner. Shi et al. [15] propose an end-to-end trainable neural network, which integrates the advantages of both CNN and RNN. CNN is built for extracting feature sequences from each input and RNN is used to predict a label distribution for each frame of feature sequences. The CTC proposed by Graves et al. [44] is introduced to transcript preframe distribution made by RNN into a label sequence. Luo et al. [22] propose a multiobject rectification network to handle the complex deformations that are robust to irregular texts and an attention-based sequence recognition network which is built for recognizing the rectified image. Instead of rectifying the whole image, Liu et al. [45] propose to detect and rectify individual character regions through a simple local Transformer network. Besides, Cheng et al. [46] combine the arbitrary orientation network which extracts scene text features in four directions into an attention-based decoder to generate character sequence. With extra use of character-level annotations, Liao et al. [47] recognize the text of arbitrary shapes with a semantic segmentation network.

With the unprecedented success of the Transformer [48], [49], [50] framework in computer vision [51], [52], [53], there are also some researches on the Transformer-based scene text recognizer. Sheng et al. [25] propose a no-recurrence encoder–decoder recognizer, where the encoder uses stacked self-attention to extract image features and the decoder adopts stacked self-attention to perform text recognition. Yang et al. [26] directly connect 2-D CNN features to an attention-based sequence decoder with no recurrent module, which is trained in parallel and is guided by the holistic representation. Lu et al. [54] propose a novel multiaspect nonlocal block and incorporate it into the conventional CNN backbone, which motivates the feature extractor to model the global context information. Atienza [14] propose a simple single-stage text recognition method built upon a compute and parameter efficient vision Transformer. Nevertheless, the encoder and decoder module in the Transformer consists of multiple blocks, each of which contains a series of fully connected layers, largely increasing the number of parameters and computational flops.

In this article, we also regard text recognition as a sequence-to-sequence problem and propose a GLaLT for recognition. In particular, we put much attention to balancing the tradeoff between recognition accuracy and inference efficiency, where the inference process simply contains an attention-augmented encoder and a CTC decoder. Compared with the CTC-based recognizers, the proposed GLaLT uses auxiliary guidance from the attention module, which is more effective and robust to recognize irregular scene text images. Besides, compared with the attention-based recognizers, GLaLT incorporates both the global and local context dependencies and has a parallel attention mechanism, which is more efficient than predicting sequence frame by frame. To the best of our knowledge, this is the first work of introducing a multiloss constraint scheme into the Transformer in STR, which reduces the computing resource consumption and improves the accuracy of the CTC-based methods.

III. Methodology

GLaLT, as shown in Fig. 2, contains an attention-augmented encoder and a hybrid CTC and attention decoder. The decoder consists of two parallel modules: one is the Transformer-decoder-based attention module and the other is the CTC module. During training, the Transformer-decoder-based attention module provides effective guidance in learning better alignment and feature representations. During inference, we can simply use the encoder and the CTC module to predict the

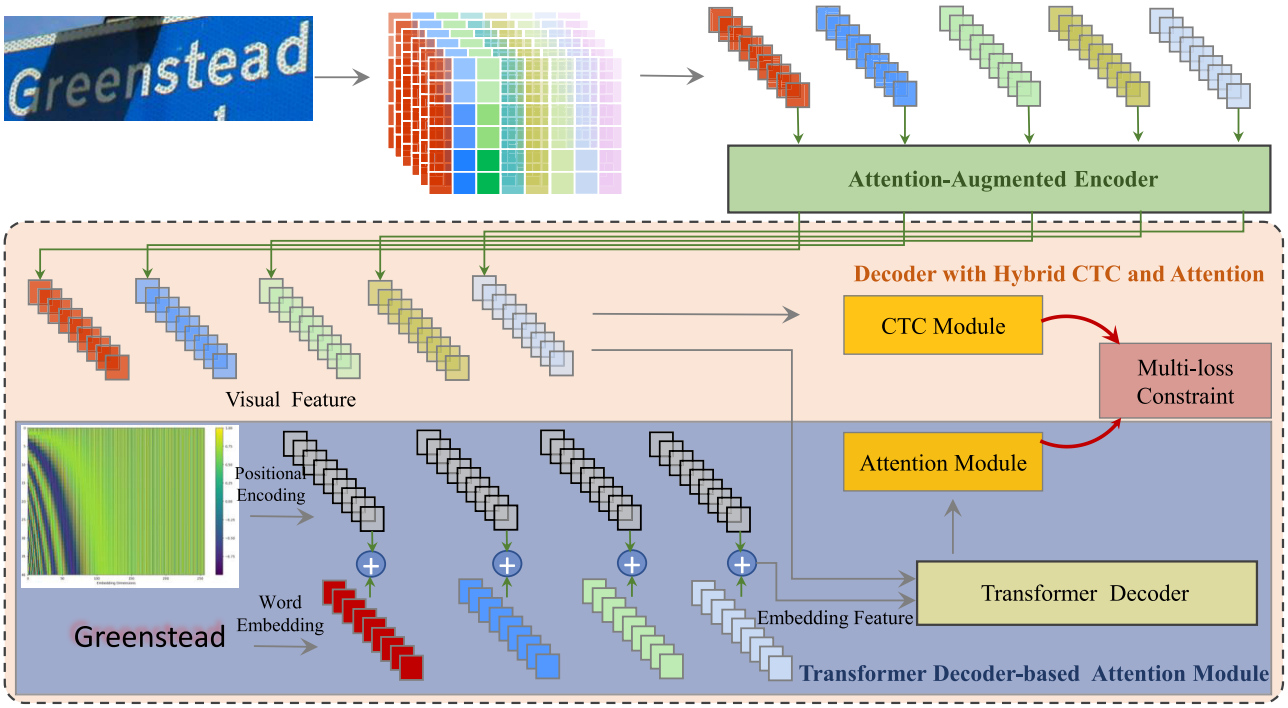


Fig. 2. Overall architecture of GLaLT. The text image is first fed into a two-layer subsampling module, and then a linear operation is applied to reshape it into an encoder-length vector. After that, the attention-augmented encoder is attached to transform the encoder-length vector into visual feature representation. Finally, the visual feature is put into two modules, one is the CTC module, aiming at efficient inference, and the other is the Transformer-decoder-based attention module, consisting of a Transformer decoder and an attention module.

character sequences, which can largely reduce the number of parameters and flops compared with the Transformer-based recognizer.

A. Attention-Augmented Encoder

The self-attention mechanism [55], [56], [57] has emerged as an integral part of compelling sequence modeling and transduction models in various tasks, allowing capturing long-range interactions with no regard to their distance. However, they possess a weak capacity for extracting fine-grained local information. CNN exploits a shared position-based kernel to gradually scan the 2-D input data. The local connectivity makes it capable of extracting local feature patterns. The limitation of that is lack in long-term global relationship modeling. Inspired by Gulati et al. [58], we propose the GLRA, which combines the self-attention module with the convolution module in text recognition model. The self-attention module specializes in long-term global relationship modeling, while the convolution module focuses on local context modeling.

1) Global Range Attention: Unlike conventional feature extraction stage, which puts too much emphasis on deeper CNNs, such as VGG [59], [60], RCNN [61], [62], and ResNet [63], [64], we obtain the visual features by leveraging two convolutional layers. This two-layer subsampling module not only saves a lot of memory consumption but also draws more discriminative features. More information is detailed in experiements. Then a linear operation is applied to reshape it into an encoder-length vector Y, with the shape of (T, d_{in}), where T is the length of the sequence and d_{in} is the embedding size. After that, they are inputted into the encoder, consisting of a stack of N_e encoder blocks, each of which has two feedforward modules and a GLRA module, as shown in Fig. 3. The GLRA module consists of a self-attention module and a

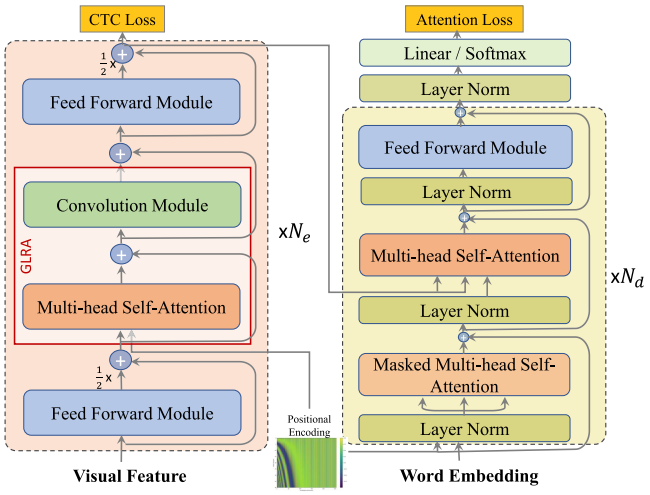


Fig. 3. Encoder and decoder.

convolution module. The self-attention module is a multihead self-attention with relative positional encoding (MSAP), which is proposed in Transformer-XL [65]. Instead of the elementwise addition of the word embedding and the positional encoding in the Transformer [66], [67], we can incorporate the positional encoding into the attention score in each layer, to generalize well on different lengths of input and be robust to the varying length. The output of the MSAP for a single head h is computed as

$$\text{head}_h = \text{softmax}\left(\frac{QK^T + QR^T}{\sqrt{d_k^h}}\right) * V \tag{1}$$

where Q denotes the queries and $Q = YW_q, W_q \in \mathbb{R}^{d_{\text{in}} \times d_k^h}$, K denotes the keys and $K = YW_k, W_k \in \mathbb{R}^{d_{\text{in}} \times d_k^h}$, and V denotes the values and $V = YW_v, W_v \in \mathbb{R}^{d_{\text{in}} \times d_v^h}$. R represents the relative positional encoding and $R = PW_r, W_r \in \mathbb{R}^{d_{\text{in}} \times d_k^h}, P \in \mathbb{R}^{T \times d_{\text{in}}}$. Then the outputs of all the heads are concatenated and projected again as follows:

$$\text{MSAP}(Y) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W_o \tag{2}$$

where $W_o \in \mathbb{R}^{d_v \times d_v}$ is a learned linear transformation, $d_v = h * d_v^h$, and in this article, $d_k^h = d_v^h$.

2) Local Context Modeling: To encourage the model to capture both global and local contexts, we combine a convolution module with the self-attention module. The convolution module contains a pointwise convolution with an expansion factor of 2 with a gated linear unit activation layer [68]. Then a 1-D depthwise convolution operation is leveraged to reduce the computation and achieve local context modeling. Thus, the output could flow into a batch norm and a swish activation layer to help train and regularize deep models. In this manner, we place the self-attention module and the convolution module successively, endowing them with the ability of having the global and local perspective of the image. The feedforward module in the Transformer is split into two half-step feedforward layers, inspired by Conformer [58]. One is before the MSAP module and the other is after that. The output of an encoder block is

$$\begin{aligned} \widetilde{Y}_i &= Y_i + \frac{1}{2}\text{FFM}(Y_i), \quad \widehat{Y}_i = \widetilde{Y}_i + \text{MSAP}(\widetilde{Y}_i) \\ \overline{Y_i} &= \widehat{Y}_i + \text{CM}(\widehat{Y}_i), \quad O_i = \text{LN}\left(\overline{Y_i} + \frac{1}{2}\text{FFM}(\overline{Y_i})\right) \end{aligned} \tag{3}$$

where FFM refers to the feedforward module, CM refers to the convolution module, and LN refers to layer normalization. For input Y_i to the encoder block i, O_i represents the output of the block, where $i \in [1, N_e]$.

B. Transformer-Decoder-Based Attention Module

The text image is first fed into a two-layer subsampling module, and then a linear operation is applied to reshape it into an encoder-length vector. After that, the attention-augmented encoder is attached to transform the encoder-length vector into visual feature representation. Finally, the visual feature is put into two modules, one is the CTC module, aiming at efficient inference, and the other is the Transformer-decoder-based attention module, consisting of a Transformer decoder and an attention module, as shown in Fig. 3. The Transformer decoder contains a stack of N_d decoder blocks, each of which consists of three core modules: masked multihead self-attention, multihead self-attention, and a feedforward module. First, the input character is converted into a vector with d_v dimension using a character-level embedding layer. A unique position encoding with the same dimension is added to each embedding. The masked multihead self-attention prevents positions from participating in subsequent positions, that is to say, ensuring the prediction only relies on the information before the current position. The keys and values for the second multihead self-attention come from the output of the encoder and the queries come from the output of the masked multihead self-attention. The last is the feedforward module, which can be subdivided into two layers: a linear layer with a linear activation function and a linear layer with ReLU activation. The residual connection and layer normalization are used in each sublayer of the decoder block.

C. Hybrid CTC and Attention Decoder

CTC uses Markov assumptions to efficiently deal with sequential problems by dynamic programming [69], [70] and requires a series of conditional independence assumptions. Attention decoder does not need these extra assumptions. However, attention decoder is usually tough to train the encoder with proper alignment under conditions of noisy data and long sequences. Furthermore, the attention-based model predicts a character conditioned on the history of previous characters at each time step. The decoding frame by frame makes it difficult to apply in real-environment text recognition tasks. To overcome the above issues, we propose the hybrid CTC and attention decoder, consisting of the CTC module and Transformer-decoder-based attention module. The CTC module contains only a linear layer, which transforms the output of the encoder into CTC-form activation. The Transformer-decoder-based attention module is adopted to achieve attention-based probabilities, as illustrated above. Specifically, the attention objective is inserted into the CTC module as a regularization technique, which is only used for training. The forward–backward algorithm of CTC enforces monotonic alignment between input and output sequences. Besides, the CTC module can make the network converge fast. The attention module does not require any conditional independence assumptions, which can provide strong and effective supervision in training the recognizer.

Hybrid CTC and attention decoder integrates the attention module and the CTC module, where CTC and attention share the encoder network. This integrated method can not only make full use of the feature information before the current position through the encoding and decoding mechanism in the attention architecture but also leverage the feature information after the current position through the method of calculating the global probability in the CTC architecture. At the same time, hybrid CTC and attention decoder not only accelerates the convergence speed of the network but also improves the recognition performance. The proposed objective is represented as follows:

$$L = \alpha L_{\text{ctc}} + (1 - \alpha)L_{\text{attn}} \tag{4}$$

where α is a tunable parameter, with $0 \leq \alpha \leq 1$. L_{attn} is the attention loss, which adopts the cross-entropy loss function. L_{ctc} is the CTC loss, whose objective is to minimize the negative log-likelihood of conditional probability of ground truth

$$L_{\text{ctc}} = -\log P(G|S) \tag{5}$$

where G is the ground-truth label sequence, and $S = s_1, \ldots, s_T$ is the sequence produced by the encoder layer from the training image I. Each s_t is a probability distribution over the character set C, which contains all the labels in the task and a "blank" label. A sequence-to-sequence mapping function M is defined on sequence $\Pi \in C^T$. M maps Π into G by first removing the repeated labels, and then removing the "blank" label. For example, M maps "–hh-e-l-ll-oo–" ("-" represents "blank") into "hello." Then, the conditional probability is defined as the sum of probabilities of all π that are mapped

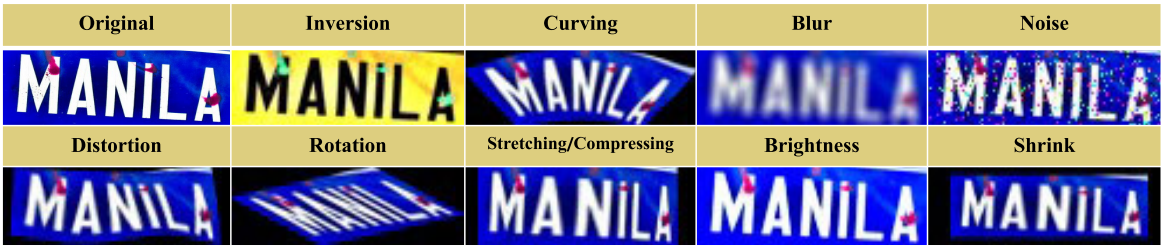


Fig. 4. Illustration of data augmentation methods designed for STR.

by M into G

$$P(G|S) = \sum_{\Pi:M(\Pi)=G} P(\Pi|S). \tag{6}$$

The probability of Π is defined as $P(\Pi|S) = \prod_{t=1}^{T} s_t^{\Pi_t}$, where $s_t^{\Pi_t}$ is the probability of having label Π_t at time stamp t. All the symbols used in the proposed method can be seen in the symbol table with meanings, as shown in Nomenclature.

IV. Experiments

A. Datasets

For training, we use the following datasets: MJSynth [83] and SynthText in the Wild [84]. The model is then evaluated on the following standard datasets consistent with [78]: IIIT5K-Words (IIIT5K) [85], Street View Text (SVT) [86], ICDAR 2003 (IC03) [87], ICDAR 2013 (IC13) [88], ICDAR2015 (IC15) [89], SVT Perspective (SVTP) [90], and CUTE80 (CUTE) [91].

Training Datasets: MJSynth [83] consists of nine million images produced by a synthetic text generation engine, with accurate world-level labeling. SynthText [84] is a synthetic dataset originally introduced for text detection in natural images, which contains 800 000 scene-text images, each with multiple instances of words in various styles. We train the proposed model using the combination of the two datasets.

IIIT5K contains words from both street scene texts and born-digital images, which has 2000 images for training and 3000 images for evaluation.

SVT consists of 257 images for training and 647 images for evaluation, some of which are noisy, blurry, or of low resolution.

IC03 contains 1156 images for training and 1110 images for evaluation. There are 243 ignored images whose words have nonalphanumeric characters or are less than three characters. The two different versions are widely used: IC03_867 and IC03_860. IC03_867 has 7 more text boxes than IC03_860.

IC13 consists of 848 images for training and 1095 for evaluation. Removing words with nonalphanumeric characters leaves us with 1015 images for evaluation. There are also two different versions: IC13_857 and IC13_1015. IC13_857 is the subset of IC13_1015 where words that are less than 3 characters are discarded.

IC15 consists of 4468 images for training and 2077 for evaluation. For a fair comparison, we also evaluate on two different versions: IC15_2077 and IC15_1811. IC15_1811 ignores many severely distorted, blurred, or non-alphanumeric character images.

TABLE I
Model Configurations

Model Version	Block N_e	Head h	Units w	Input
Small	6	4	1024	100×32
Base	12	4	1024	100×32
Big	12	4	2048	100×32
Big*	12	8	2048	256×64

SVTP contains 645 images taken from Google Street View, where words are of a great variety of viewpoints and orientations.

CUTE consists of 288 cropped images for evaluation. The images are mostly of high quality but have a lot of curved text instances.

B. Implementation Details

We use a union of MJSynth and SynthText as our training data and evaluate on the above real-world benchmarks. The Adam optimizer is adopted with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 1e-8$. The learning rate is varying over the course of training by $\text{warmup}^{0.5} \times \min(s^{-0.5}, s \times \text{warmup}^{-1.5})$. Here, s denotes the current training step, and warmup controls over the learning rate first increase and then decrease, which is set to 25 000. The training batch size is 192, and the number of epoch is 40. Gradient clipping is used at magnitude 5. The number of decoder blocks N_d is 6. d_{in} and d_v is set to 256. We train different versions of GLaLT for a thorough comparison. The detailed model configurations are summarized in Table I.

C. Comparison With State-of-the-Art

Table II shows comparison of the (Accuracy) result of our model on seven public benchmarks with a few state-of-the-art models. Our proposed GLaLT achieves a better tradeoff with regard to other existing models against the total accuracy and the number of model parameters. Concretely, compared with the scene text recognizer ViSTR-Small proposed in [14], our proposed GLaLT-Small obtains superior performance by a margin of 4.1% on IIIT5K, 2.2% on SVT, 3.6% on IC13_1015, and 2.3% on IC15_1811, only a little lower accuracy on CUTE (0.5%). Using a recipe of data augmentation specifically targeted for STR can significantly boost the accuracy of GLaLT. In Fig. 4, we illustrate some data augmentation methods designed for STR, such as inversion, curving, noise, blur, rotation, distortion, stretching/compressing, perspective, and shrinking. In the experiments, we randomly select three augmentation ways. We can see from Table II that applying data augmentation improves the total accuracy of GLaLT-Small by

TABLE II

PERFORMANCE OF EXISTING TEXT RECOGNITION MODELS AND OUR PROPOSED MODEL ON PUBLIC BENCHMARKS. THE LAST ROW SHOWS THE IMPROVEMENT OF THE PROPOSED BEST MODEL COMPARED WITH THE BASELINE METHOD VITSTR-BASE + AUG. "+AUG" MEANS ADDING DATA AUGMENTATION. "ACC." MEANS MODEL ACCURACY. THE TOP ACCURACY FOR EACH BENCHMARK IS SHOWN IN **BOLD**

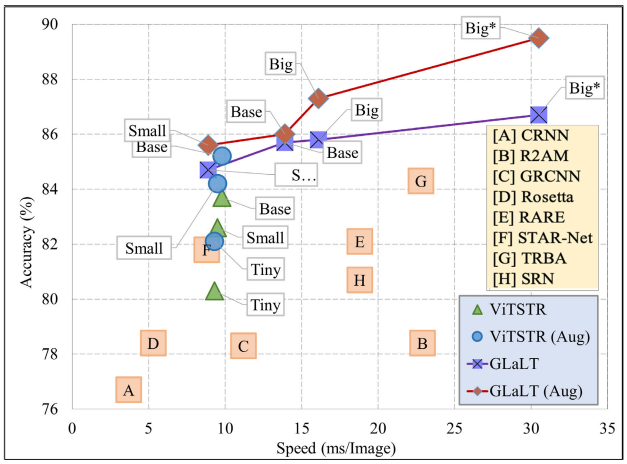
Methods	IIIT5K	SVT	IC03		IC13		IC15		SVTP	CUTE	Acc.	Params	CTC/
	3000	647	860	867	857	1015	1811	2077	645	288	%	$\times 10^6$	Attention
ATR [71]	-	-	-	-	-	-	-	-	75.8	69.3	-	-	Attention
FAN [72]	87.4	85.9	-	94.2	-	93.3	70.6	-	-	-	-	-	Attention
RARE [21]	86.0	85.4	93.5	93.4	92.3	91.0	73.9	68.3	75.4	71.0	82.1	10.8	Attention
STAR-Net [73]	85.2	84.7	93.4	93.0	91.2	90.5	74.5	68.7	74.7	69.2	81.8	48.9	CTC
Char-Net [45]	83.6	84.4	91.5	-	90.8	-	-	60.0	73.5	-	-	-	Attention
AON [46]	87.0	82.8	-	91.5	-	-	-	68.2	73.0	76.8	-	-	Attention
EP [74]	88.3	87.5	-	94.6	-	94.4	73.9	-	-	-	-	-	Attention
SSFL [75]	89.4	87.1	-	94.7	94.0	-	-	-	73.9	62.5	-	-	CTC
CRNN [15]	81.8	80.1	91.7	91.5	89.4	88.4	65.3	60.4	65.9	61.5	76.7	8.5	CTC
R2AM [76]	83.1	80.9	91.6	91.2	90.1	88.1	68.5	63.3	70.4	64.6	78.4	2.9	CTC
GCRNN [77]	82.9	81.1	92.7	92.3	90.0	88.4	68.1	62.9	68.5	65.5	78.3	4.8	CTC
Rosetta [13]	82.5	82.8	92.6	91.8	90.3	88.7	68.1	62.9	70.3	65.5	78.4	44.3	CTC
TRBA [78]	87.8	87.6	94.5	94.2	93.4	92.1	77.4	71.7	78.1	75.2	84.3	49.6	Attention
MORAN [22]	91.2	88.3	-	95.0	-	92.4	-	68.8	76.1	77.4	-	-	Attention
CCL [79]	91.1	85.9	-	93.5	-	92.8	-	72.9	-	-	-	-	CTC
TextScanner [23]	**93.9**	90.1	-	-	-	92.9	79.4	-	83.3	79.4	-	-	Attention
ASTER+AEG [80]	93.6	89.2	-	94.8	-	92.9	-	75.5	80.0	80.2	-	-	Attention
RobustScanner [24]	95.3	88.1	-	-	-	94.8	-	77.1	79.5	90.3	-	-	Attention
SEED [81]	93.8	89.6	-	-	-	92.8	-	80.0	81.4	**83.6**	-	-	Attention
SRN [82]	84.6	83.5	92.8	92.4	90.3	88.0	71.3	68.3	71.3	69.3	80.7	57.3	Attention
ViTSTR-Small [14]	85.6	85.3	93.9	93.6	91.7	90.6	75.3	69.5	78.1	71.3	82.6	21.5	CTC
ViTSTR-Small+Aug [14]	86.6	87.3	94.2	94.2	92.1	91.2	77.9	71.7	81.4	77.9	84.2	21.5	CTC
ViTSTR-Base [14]	86.9	87.2	93.8	93.4	92.1	91.3	76.8	71.1	80.0	74.7	83.7	85.8	CTC
ViTSTR-Base+Aug[14]	88.4	87.7	94.7	94.3	93.2	92.4	78.5	72.6	81.8	81.3	85.2	85.8	CTC
GLaLT-Small	89.7	87.5	94.1	94.0	94.6	94.2	77.6	71.3	75.5	69.8	84.7	11.3	CTC
GLaLT-Small+Aug	89.4	89.6	95.0	94.8	94.0	93.8	79.1	73.2	77.7	75.3	85.6	11.3	CTC
GLaLT-Base	90.5	88.6	95.2	95.3	95.2	95.1	78.3	72.4	78.3	70.5	85.7	21.7	CTC
GLaLT-Base +Aug	89.7	89.5	94.8	94.6	95.1	94.8	79.4	73.7	78.5	76.7	86.0	21.7	CTC
GLaLT-Big	90.2	88.3	95.1	95.0	95.1	94.7	79.2	73.0	76.3	73.6	85.8	34.3	CTC
GLaLT-Big+Aug	90.4	90.0	95.5	95.2	95.8	95.3	80.5	74.7	79.4	77.1	86.8	34.3	CTC
GLaLT-Big*	91.8	89.0	94.4	94.5	95.1	94.9	81.3	76.5	80.0	77.7	87.3	34.9	CTC
GLaLT-Big*+Aug	*93.1*	**90.9**	**96.3**	**95.6**	**96.4**	**96.2**	**83.7**	**80.1**	**83.2**	*83.3*	**89.5**	34.9	CTC
Improvement	*+4.7*	*+3.2*	*+1.6*	*+1.3*	*+3.2*	*+3.8*	*+5.2*	*+7.5*	*+1.4*	*+2.0*	*+4.3*	-	-

+0.9%, GLaLT-Base by +0.3%, GLaLT-Big by +1.0%, and GLaLT-Big* by +2.2%. Besides, GLaLT-Big* with data augmentation (GLaLT-Big* + Aug) can achieve an improvement of 4.3% in total average accuracy compared with ViSTR-Base + Aug, from 85.2% to 89.5%, with only 34.9M number of parameters, which requires less than half of the parameters for model inference. Without bells and whistles, our approach outperforms GCRNN [77], Rosetta [13], and TRBA [78]. GCRNN [77], Rosetta [13], and TRBA [78] achieve a total accuracy of 78.3%, 78.4%, and 84.3%, respectively, on public benchmarks, which is 11.2%, 11.1%, and 5.2% lower than our proposed GLaTE-Big* + Aug. CCL [79] is a novel character-level STR framework for simultaneously categorizing and localizing characters. Our method has an improvement of 2.0% on IIIT5K, 5.0% on SVT, 2.1% on IC03, 3.4% on IC13, and 7.2% on IC15 compared with it. Qiao et al. [81] propose the SEED method, which uses the global semantic information to enhance most encoder–decoder methods. It achieves 92.8% on IC13 and 81.4% on SVTP, which are 3.4% and 1.8% lower than our proposed GLaLT-Big* + Aug, respectively. We only have a little lower accuracy on IIIT5K and CUTE, 0.7% on IIIT5K, and 0.3% on CUTE. SRN [82] uses the global semantic reasoning module to capture global semantic context through multiway parallel transmission. Since the code for this article is not open source, we borrow the unofficial PyTorch implementation of SRN in the GitHub community for evaluating on the same standard datasets. With the same input size 100 × 32, GLaLT-Big + Aug leads to 6.1% increase in total accuracy. For the high resolution of input size, we can achieve a relative improvement of 8.8%.

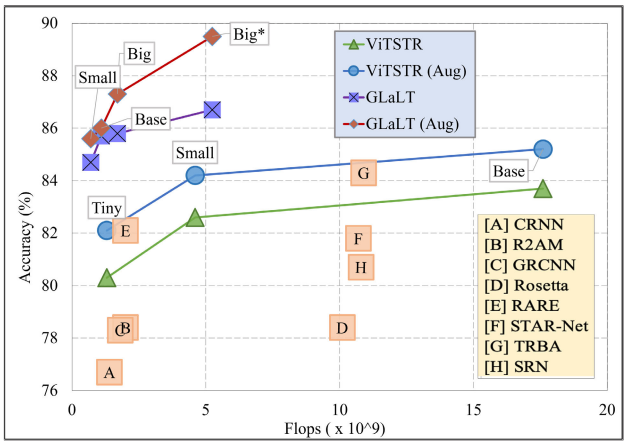
Fig. 5 shows the overall tradeoffs under three comparisons (accuracy *versus* speed, accuracy *versus* parameters, and accuracy *versus* flops). GLaLT-Small achieves 84.7% accuracy (85.6% with data augmentation), is fast at 8.9 ms/image, with a small number of 11.3M parameters. and requires much fewer

TABLE III
MODEL ACCURACY (ACC.), PARAMETERS (#PARA.), SPEED, AND COMPUTATIONAL REQUIREMENTS (FLOPS)

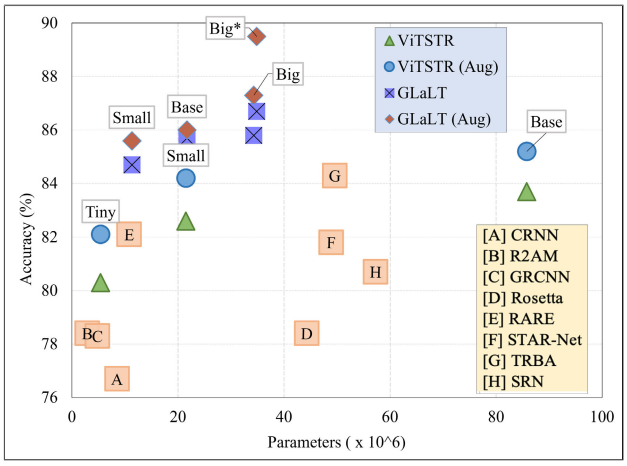
Methods	Acc. %	Speed ms/Image	#Para. ($\times 10^6$)	Flops ($\times 10^9$)
CRNN [15]	76.7	3.7	8.5	1.4
R2AM [76]	78.4	22.9	2.9	2.0
GCRNN [77]	78.3	11.2	4.8	1.8
Rosetta [13]	78.4	5.3	44.3	10.1
RARE [21]	82.1	18.8	10.8	2.0
STAR-Net [73]	81.8	8.8	48.9	10.7
TRBA [78]	84.3	22.8	49.6	10.9
SRN [82]	80.7	18.8	57.3	10.8
ViTSTR-Tiny [14]	80.3	9.3	5.4	1.3
ViTSTR-Tiny+Aug [14]	82.1	9.3	5.4	1.3
ViTSTR-Small [14]	82.6	9.5	21.5	4.6
ViTSTR-Small+Aug [14]	84.2	9.5	21.5	4.6
ViTSTR-Base [14]	83.7	9.8	85.8	17.6
ViTSTR-Base+Aug [14]	85.2	9.8	85.8	17.6
GLaLT-Small	84.7	8.9	11.3	0.7
GLaLT-Small+Aug	85.6	8.9	11.3	0.7
GLaLT-Base	85.7	13.9	21.7	1.1
GLaLT-Base+Aug	86.0	13.9	21.7	1.1
GLaLT-Big	85.8	16.1	34.3	1.7
GLaLT-Big+Aug	86.8	16.1	34.3	1.7
GLaLT-Big*	87.3	30.5	34.9	5.2
GLaLT-Big*+Aug	89.5	30.5	34.9	5.2



(a)



(b)



(c)

Fig. 5. Tradeoffs between accuracy versus (a) speed, (b) computational load (flops), and (c) number of parameters. Our proposed GLaLT is at or near the borders to maximize the performance on all the metrics.

computations at 0.7 $\times$ 10^9 flops. GLaLT-Base achieves a higher accuracy of 85.7% (86.0% with data augmentation) and is also fast at 13.9 ms/image while requiring 21.7M parameters and 1.1G flops. With data augmentation, GLaLT-Big achieves 87.4% accuracy at 16.1 ms/image and requires 34.3M parameters and 1.7G flops. GLaLT-Big* (with data augmentation) achieves the best accuracy and requires 34.9M parameters, which spends 30.5 ms in processing one image. As shown in Fig. 5, our proposed GLaLT is at or near the border of accuracy *versus* speed, accuracy *versus* parameters, and accuracy *versus* flops, which yields state-of-the-art performance with reasonably short running time and little computational overhead. Specifically, the best ViTSTR version, called ViTSTR-Base + Aug [14], achieves 85.2% accuracy, which is 0.8% lower than our proposed GLaLT-Base + Aug. Besides, it requires 85.8M parameters and 17.6G flops, which has 74.7% and 93.8% relative increase compared with ours, demanding higher computational requirements. SRN [82] is another variant of the Transformer. The inference speed has a 26.1% relative improvement compared with GLaLT-Base + Aug. Besides, our approach spends less than half of computational overheads and achieves a 5.3% improvement in accuracy. The performance of GLaLT-Big* + Aug method with a bigger input size can achieve a larger boost. More details can refer to Table III.

Some good representative results of GLaLT are illustrated in Fig. 6, As can be seen, GLaLT demonstrates excellent capability in recognizing extremely challenging text images. Specifically, Fig. 6 shows GLaLT has excellent capability on recognizing text images with low resolution and low visual quality, some of which are even hard to humans, like the image of "special" and the image of "perfect." Besides, we also demonstrate some results on handwriting texts with different forms, like the image of "concierge," the image of "mandarin," and the image of "tastes." The last two columns show that GLaLT is capable of recognizing a variety of text images with complex geometric appearance.

TABLE IV

DISENTANGLING THE ATTENTION-AUGMENTED ENCODER: 1) REMOVING THE CONVOLUTION MODULE; 2) REPLACING THE SYMMETRIC FFMS WITH A SINGLE FFM 1; 3) REPLACING THE SYMMETRIC FFMS WITH A SINGLE FFM 2; 4) REMOVING THE RELATIVE POSITIONAL EMBEDDING IN MULTIHEAD SELF-ATTENTION; AND 5) REPLACING THE ATTENTION-AUGMENTED ENCODER BLOCK WITH THE VANILLA TRANSFORMER ENCODER. ALL ABLATION STUDY RESULTS ARE EVALUATED WITHOUT DATA AUGMENTATION

Methods	IIIT5K	SVT	IC03		IC13		IC15		SVTP	CUTE	Acc.
	3000	647	860	867	857	1015	1811	2077	645	288	%
GLaLT-Big	**90.2**	**88.3**	**95.1**	**95.0**	**95.1**	**94.7**	**79.2**	**73.0**	76.3	73.6	**85.8**
- Convolution module	87.1	86.2	93.4	93.3	92.3	92.1	76.0	70.2	75.5	71.2	83.2
- Relative positional embedding	89.2	86.7	94.4	94.3	94.7	94.3	77.8	71.3	75.8	72.9	84.7
- FFM 1	88.6	86.6	94.8	94.6	93.9	93.3	76.0	69.7	**78.1**	72.2	84.0
-FFM 2	89.3	87.2	94.7	94.5	94.4	93.9	77.2	70.5	75.0	**74.7**	84.5
Vanilla Transformer	87.3	86.2	93.7	93.7	93.5	92.8	75.2	68.9	75.3	72.2	83.1

TABLE V

VARYING α FOR BALANCING THE CTC AND ATTENTION LOSS. "AUG" MEANS TRAINING THE MODEL WITH DATA AUGMENTATION

Methods	IIIT5K	SVT	IC03		IC13		IC15		SVTP	CUTE	Average
	3000	647	860	867	857	1015	1811	2077	645	288	%
0.0	88.9	85.5	93.9	94.0	94.6	94.5	76.6	67.6	75.7	70.1	85.3
0.0(Aug)	88.2	88.1	93.4	93.2	94.0	94.1	79.6	73.5	78.1	73.3	86.4
0.1	90.2	87.3	94.5	94.7	95.2	94.8	78.1	72.2	75.7	73.3	83.7
0.1(Aug)	90.3	88.4	94.8	94.8	95.1	94.8	80.2	74.3	78.9	77.1	85.1
0.3	90.2	88.3	95.1	95.0	95.1	94.7	79.2	73.0	76.3	73.6	85.8
0.3(Aug)	**90.4**	**89.9**	**95.5**	**95.2**	**95.8**	**95.3**	**80.5**	**74.7**	**79.4**	**77.1**	**86.8**
0.5	89.4	87.8	95.0	94.7	94.7	94.8	78.9	72.7	74.7	70.5	85.2
0.5(Aug)	90.2	89.6	94.5	94.2	95.2	94.7	79.9	74.3	78.6	76.0	86.3
0.7	89.3	87.9	94.3	93.9	94.3	94.1	76.9	70.9	75.8	71.2	84.5
0.7(Aug)	89.2	87.8	94.3	94.5	95.4	94.9	79.9	74.1	76.9	77.1	85.8
0.9	89.8	87.9	94.3	94.1	94.9	94.6	77.0	71.4	75.5	70.8	84.8
0.9(Aug)	89.1	87.9	94.3	94.7	94.9	94.6	78.9	73.2	75.9	73.6	85.3
1	90.2	87.3	93.6	93.7	94.6	94.3	78.0	68.8	75.9	66.7	84.4
1.0(Aug)	88.5	88.4	94.9	94.2	94.4	94.1	78.0	72.3	77.4	73.3	84.9

Fig. 6. Some good representative results of GLaLT. For example, the highly blurred and complicated geometric appearance are recognized very well when using GLaLT.

We also analyze some wrong cases, as shown in Fig. 7, which can be divided into three reasons. First, texts are severely disturbed by complex background, e.g., background stripes or color in the example of "gwaliab" and "litter." Second, the words which look very similar, like "n" in the image of "soon" and its fault result "v." Third, the curvature of text is too large, like a semielliptical shape in the image of "chathamkent." These wrong cases also point out the future research direction of the proposed GLaLT.

D. Ablation Study

In this section, we investigate the effects of the key modules of the proposed GLaLT, namely, the attention-augmented encoder and tradeoff between CTC and attention modules. All the experiments are conducted following consistent training strategies, and their performances are reported on the public benchmarks.

1) Encoder With Attention Augmentation: The attention-augmented encoder is different from the Transformer encoder

Fig. 7. Incorrect recognitions with their text labels in yellow and incorrect outputs in red.

TABLE VI
PERFORMANCE OF VARIOUS SETTINGS IN THE SUBSAMPLING MODULE

Methods	Accuracy	Parameters ($\times 10^6$)	Flops ($\times 10^9$)
GLaLT-Big	85.8	34.3	1.7
GLaLT-Big-4Conv	85.2	35.5	2.1
GLaLT-Big-6Conv	84.6	36.7	2.5
GLaLT-Big-VGG	85.1	34.5	1.9
GLaLT-Big-IncepNet	84.7	37.6	2.6
GLaLT-Big-ResNet	84.5	44.4	4.1

in several ways, in particular, the fusion of global range attention and local context modeling, two separate feedforward modules. In this series of experiments, we study the effects of these differences. First, we replace the attention-augmented encoder with the vanilla Transformer [92], [93]. Table IV shows that our proposed method achieves a significantly better accuracy compared with the vanilla Transformer. To validate the effectiveness of the GLRA module, we conduct another two experiments: removing the convolution module and substituting relative positional embedding for self-attention in the Transformer. Convolution module is the most essential part, which contributes to a great performance improvement. Specifically, the result of removing convolution module is 87.1% on IIIT5K, 86.2% on SVT, 93.4% on IC03 (860), 93.3% on IC03 (867), 92.3% on IC13 (857), 92.1% on IC13 (1015), 76.0% on IC15 (1811), and 70.2% on IC15 (2077), which are 3.1%, 2.1%, 1.7%, 1.7%, 2.8%, 2.6%, 3.2%, and 2.8% lower than the proposed GLaLT, respectively. Then, we compare the separate half-step feedforward module with the traditional single feedforward module. "- FFM 1" means removing the first feedforward module in each encoder block. "- FFM 2" means removing the last one in each encoder block. The results show that the accuracy of our proposed model achieves 88.3% on SVT, which is 1.7% higher than "- FFM 1" and 1.1% higher than "- FFM 2." Furthermore, our method significantly outperforms "- FFM 1" by a large margin of 3.2% and "- FFM 2" by a margin of 2.0% on IC15. The same improvement can also be shown on other datasets, and hence it is sufficient to demonstrate that the separate half-step feedforward module is more effective than the traditional single one.

2) Balance Between CTC and Attention Modules: The encoder network in our proposed model is shared with the CTC and attention modules. We leverage CTC and attention loss simultaneously in the training stage. In this section, we study how to balance the CTC and attention loss. Table V shows the detailed performance results with different tradeoff coefficient parameters on all the benchmark datasets. We can see that the best accuracy is achieved when $\alpha = 0.3$. For simplicity, we give the results on an irregular dataset SVTP with and without augmentation, as shown in the bar form of Fig. 8. The total performance curve on all the benchmark datasets is shown in the line form of Fig. 8. Unlike the sole attention module, the CTC module can enforce monotonic alignment between input and output sequences. Unlike the sole CTC module, the attention module does not rely on a series of conditional independence assumptions. From the line chart, it can be observed that the best performance is achieved when the value of α is about 0.3. The obvious performance boost is acquired when compared with the sole CTC module ($\alpha = 1.0$) and the sole attention module ($\alpha = 0.0$). Specifically, we achieve 1.9% absolute improvement over the sole CTC module and 0.4% absolute improvement over the sole attention module in total accuracy with data augmentation. Besides, it can also be seen from Table V that compared with the method without data augmentation, the method using a recipe of data augmentation targeted for STR can achieve a significant boost no matter what the parameter α is. For example, when α is set to 0.1, the method with data augmentation achieves 1.4% improvement in total accuracy compared with the method without data augmentation. The increase in accuracy on regular dataset SVT is 1.1% and the increase on irregular datasets is higher, such as SVTP (3.2%) and CUTE (3.8%).

3) Exploration of Various Subsampling Modules: In this section, we investigate the performance of various settings in the subsampling module. We begin to explore the module with different numbers of convolutional layers. As described above, the GLaLT-Big model consists of two convolutional layers with stride 2. As the depth of the convolutional layer increases, the accuracy of GLaLT-Big begins to decrease, as shown in Table VI. We also replace the two-layer subsampling module with recently popular CNN networks, such as VGG [59] and ResNet-34 [63], [72]. Performance slightly degrades as the number of layer increases. In particular, GLaLT-Big-VGG leveraging VGG to extract the encoder-length features is 1.0% worse and GLaLT-Big-ResNet is 1.6% worse in accuracy than its two-layer convolutional counterpart. Notably, this two-layer subsampling module has fewer parameters and is flops-efficient. The reason behind that may be although deeper

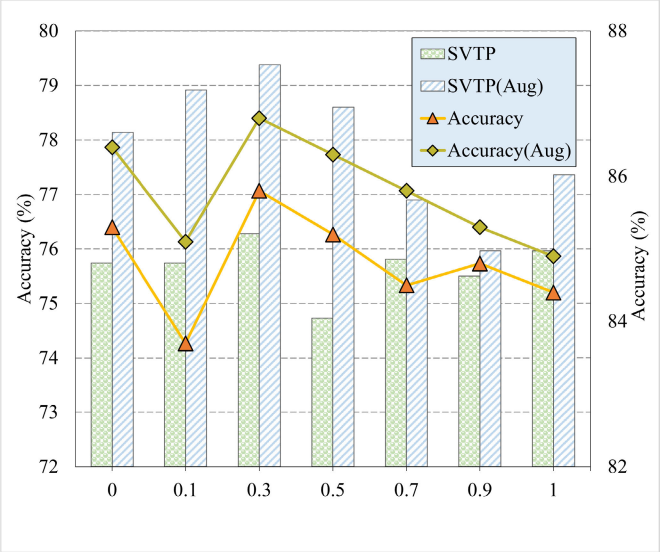


Fig. 8. Varying α for balancing the CTC and attention loss. The accuracy on an irregular dataset SVTP with and without augmentation is shown in the bar form. The total accuracy on all the benchmark datasets is shown in the line form.

convolutional layers can capture more high-level abstract semantic information from input, they often lead to loss of more fine-grained details due to resolution downsampling, especially for small text instances. Even if the subsequent GLRA module has strong feature extraction ability, the loss of information is irreversible and uncontrollable, so that we cannot benefit from the increased convolutional layers.

V. Conclusion

In this article, we propose the GLaLT method, which puts much emphasis on speed and computational efficiency. The GLRA in the attention-augmented encoder specializes in both local context modeling and long-term global dependencies, achieving remarkable improvement for the CTC-based methods. This work is the first attempt to introduce a multiloss constraint scheme into a Transformer in STR, which reduces the computing resource consumption and improves the accuracy of the CTC-based methods. Extensive experiments over several benchmarks demonstrate that the proposed GLaLT significantly outperforms the existing state-of-the-art approaches in both accuracy and inference efficiency.

In the future, we will focus more attention on nonhorizontal text recognition in natural scenes. Since a considerable portion of text in real-world scenarios is nonhorizontal, the weakness in capturing textual information in nonhorizontal texts severely restricts the practicality and applicability of the proposed method. Besides, combining zero-shot learning or few-shot learning algorithm with contextual semantic information in the process of text recognition is another problem we will concentrate on in the future. Especially in the recognition of ancient books, by integrating the visual model with contextual semantic information and combining various auxiliary information, it is possible to train and recognize samples of some categories (such as simplified Chinese characters) and extend to recognize samples of new categories (such as traditional Chinese characters), so that the machine achieves the effect of recognizing unknown characters.

References

[1] P. Staszewski, M. Jaworski, J. Cao, and L. Rutkowski, "A new approach to descriptors generation for image retrieval by analyzing activations of deep neural network layers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7913–7920, Dec. 2022.

[2] Z. Wei, X. Yang, N. Wang, and X. Gao, "Flexible body partition-based adversarial learning for visible infrared person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4676–4687, Sep. 2022.

[3] A.-A. Liu, H. Tian, N. Xu, W. Nie, Y. Zhang, and M. Kankanhalli, "Toward region-aware attention learning for scene graph generation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7655–7666, Dec. 2022.

[4] C. Zhou, Y. Liu, Q. Sun, and P. Lasang, "Vehicle detection and disparity estimation using blended stereo images," *IEEE Trans. Intell. Vehicles*, vol. 6, no. 4, pp. 690–698, Dec. 2021.

[5] K. Li, Z. Ding, K. Li, Y. Zhang, and Y. Fu, "Vehicle and person re-identification with support neighbor loss," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 826–838, Feb. 2022.

[6] J. Vargas Rivero, T. Gerbich, B. Buschardt, and J. Chen, "The effect of spray water on an automotive LIDAR sensor: A real-time simulation study," *IEEE Trans. Intell. Vehicles*, vol. 7, no. 1, pp. 57–72, Mar. 2022.

[7] K. Ma, Q. Xu, J. Zeng, G. Li, X. Cao, and Q. Huang, "A tale of HodgeRank and spectral method: Target attack against rank aggregation is the fixed point of adversarial game," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 14, 2022, doi: 10.1109/TPAMI.2022.3190939.

[8] M. Han, A. Wan, F. Zhang, and S. Ma, "An attribute-isolated secure communication architecture for intelligent connected vehicles," *IEEE Trans. Intell. Vehicles*, vol. 5, no. 4, pp. 545–555, Dec. 2020.

[9] H. Gao, Y. Qin, C. Hu, Y. Liu, and K. Li, "An interacting multiple model for trajectory prediction of intelligent vehicles in typical road traffic scenario," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 31, 2021, doi: 10.1109/TNNLS.2021.3136866.

[10] B. Weng, L. Capito, U. Ozguner, and K. Redmill, "Towards guaranteed safety assurance of automated driving systems with scenario sampling: An invariant set perspective," *IEEE Trans. Intell. Vehicles*, vol. 7, no. 3, pp. 638–651, Sep. 2022.

[11] Z. Zhang, L. Zhang, J. Deng, M. Wang, Z. Wang, and D. Cao, "An enabling trajectory planning scheme for lane change collision avoidance on highways," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 1, pp. 147–158, Jan. 2023, doi: 10.1109/TIV.2021.3117840.

[12] Z. Fan, L. Shi, Q. Liu, Z. Li, and Z. Zhang, "Discriminative Fisher embedding dictionary transfer learning for object recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 1, pp. 64–78, Jan. 2023, doi: 10.1109/TNNLS.2021.3089566.

[13] F. Borisyuk, A. Gordo, and V. Sivakumar, "Rosetta: Large scale system for text detection and recognition in images," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 71–79.

[14] R. Atienza, "Vision transformer for fast and efficient scene text recognition," in *Proc. Int. Conf. Document Anal. Recognit.* Cham, Switzerland: Springer, 2021, pp. 319–334.

[15] B. Shi, X. Bai, and C. 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.*, vol. 39, no. 11, pp. 2298–2304, Nov. 2016.

[16] W. Hu, X. Cai, J. Hou, S. Yi, and Z. Lin, "GTC: Guided training of CTC towards efficient and accurate scene text recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11005–11012.

[17] H. Fei, Y. Zhang, Y. Ren, and D. Ji, "Optimizing attention for sequence modeling via reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3612–3621, Aug. 2022.

[18] L. Xie, M. Zhang, Y. Li, W. Qin, Y. Yan, and E. Yin, "Vision–language navigation with beam-constrained global normalization," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 1, 2022, doi: 10.1109/TNNLS.2022.3183287.

[19] J. Zhang, Z. Fang, H. Sun, and Z. Wang, "Adaptive semantic-enhanced transformer for image captioning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 29, 2022, doi: 10.1109/TNNLS.2022.3185320.

[20] X. Zhao, X. Feng, and H. Chen, "A background knowledge revising and incorporating dialogue model," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Nov. 18, 2021, doi: 10.1109/TNNLS.2021.3123128.

[21] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4168–4176.

[22] C. Luo, L. Jin, and Z. Sun, "MORAN: A multi-object rectified attention network for scene text recognition," *Pattern Recognit.*, vol. 90, pp. 109–118, Jun. 2019.

[23] Z. Wan, M. He, H. Chen, X. Bai, and C. Yao, "TextScanner: Reading characters in order for robust scene text recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12120–12127.

[24] X. Yue, Z. Kuang, C. Lin, H. Sun, and W. Zhang, "RobustScanner: Dynamically enhancing positional clues for robust text recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, 2020, pp. 135–151.

[25] F. Sheng, Z. Chen, and B. Xu, "NRTR: A no-recurrence sequence-to-sequence model for scene text recognition," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 781–786.

[26] L. Yang, P. Wang, H. Li, Z. Li, and Y. Zhang, "A holistic representation guided attention network for scene text recognition," *Neurocomputing*, vol. 414, pp. 67–75, Nov. 2020.

[27] T. Wang et al., "Implicit feature alignment: Learn to convert text recognizer to text spotter," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5973–5982.

[28] C. Yao, X. Bai, and W. Liu, "A unified framework for multioriented text detection and recognition," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4737–4749, Nov. 2014.

[29] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Zürich, Switzerland: Springer, 2014, pp. 512–528.

[30] Y. Song, J. Si, S. Coleman, and D. Kerr, "Editorial biologically learned/inspired methods for sensing, control, and decision," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 5, pp. 1820–1824, May 2022.

[31] Q. Chen, Y. Wang, and Y. Song, "Tracking control of self-restructuring systems: A low-complexity neuroadaptive PID approach with guaranteed performance," *IEEE Trans. Cybern.*, early access, Nov. 8, 2021, doi: 10.1109/TCYB.2021.3123191.

[32] X. Liu, T. Kawanishi, X. Wu, and K. Kashino, "Scene text recognition with CNN classifier and WFST-based word labeling," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3999–4004.

[33] V. Goel, A. Mishra, K. Alahari, and C. V. Jawahar, "Whole is greater than sum of parts: Recognizing scene text words," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 398–402.

[34] J. Rodriguez and F. Perronnin, "Label embedding for text recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2013, pp. 1–5.

[35] Q. Deng and D. Söffker, "A review of the current HMM-based approaches of driving behaviors recognition and prediction," *IEEE Trans. Intell. Vehicles*, vol. 7, no. 1, pp. 21–31, Mar. 2022.

[36] I. Ahmed, S. Din, G. Jeon, F. Piccialli, and G. Fortino, "Towards collaborative robotics in top view surveillance: A framework for multiple object tracking by detection using deep learning," *IEEE/CAA J. Automatica Sinica*, vol. 8, no. 7, pp. 1253–1270, Jul. 2021.

[37] C. Wang, F. Li, Y. Wang, and J. R. Wagner, "Haptic assistive control with learning-based driver intent recognition for semi-autonomous vehicles," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 1, pp. 425–437, Jan. 2023, doi: 10.1109/TIV.2021.3137805.

[38] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Word spotting and recognition with embedded attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2552–2566, Dec. 2014.

[39] A. Gordo, "Supervised mid-level features for word image representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2956–2964.

[40] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *Int. J. Comput. Vis.*, vol. 116, no. 1, pp. 1–20, 2016.

[41] T.-H. Chen and T. S. Chang, "RangeSeg: Range-aware real time segmentation of 3D LiDAR point clouds," *IEEE Trans. Intell. Vehicles*, vol. 7, no. 1, pp. 93–101, Mar. 2022.

[42] K. Liu, Z. Ye, H. Guo, D. Cao, L. Chen, and F.-Y. Wang, "FISS GAN: A generative adversarial network for foggy image semantic segmentation," *IEEE/CAA J. Automatica Sinica*, vol. 8, no. 8, pp. 1428–1439, Aug. 2021.

[43] X. Li, H. Duan, Y. Tian, and F.-Y. Wang, "Exploring image generation for UAV change detection," *IEEE/CAA J. Automatica Sinica*, vol. 9, no. 6, pp. 1061–1072, Jun. 2022.

[44] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[45] W. Liu, C. Chen, and K.-Y. Wong, "Char-Net: A character-aware neural network for distorted scene text recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–8.

[46] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, and S. Zhou, "AON: Towards arbitrarily-oriented text recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5571–5579.

[47] M. Liao et al., "Scene text recognition from two-dimensional perspective," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 8714–8721.

[48] P. Xu, C. K. Joshi, and X. Bresson, "Multigraph transformer for free-hand sketch recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5150–5161, Oct. 2022.

[49] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 494–514, Feb. 2022.

[50] J. Chen, N. Zhao, R. Zhang, L. Chen, K. Huang, and Z. Qiu, "Refined crack detection via LECSFormer for autonomous road inspection vehicles," *IEEE Trans. Intell. Vehicles*, early access, Sep. 6, 2022, doi: 10.1109/TIV.2022.3204583.

[51] L. Jiao et al., "New generation deep learning for video object detection: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3195–3215, Aug. 2022.

[52] S. Zhao, M. Hu, Z. Cai, Z. Zhang, T. Zhou, and F. Liu, "Enhancing Chinese character representation with lattice-aligned attention," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 5, 2021, doi: 10.1109/TNNLS.2021.3114378.

[53] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022.

[54] N. Lu et al., "MASTER: Multi-aspect non-local network for scene text recognition," *Pattern Recognit.*, vol. 117, Sep. 2021, Art. no. 107980.

[55] P. Cai, Y. Sun, H. Wang, and M. Liu, "VTGNet: A vision-based trajectory generation network for autonomous vehicles in urban environments," *IEEE Trans. Intell. Vehicles*, vol. 6, no. 3, pp. 419–429, Sep. 2021.

[56] P. Huang, J. Han, N. Liu, J. Ren, and D. Zhang, "Scribble-supervised video object segmentation," *IEEE/CAA J. Automatica Sinica*, vol. 9, no. 2, pp. 339–353, Feb. 2022.

[57] K. Zhang, Y. Su, X. Guo, L. Qi, and Z. Zhao, "MU-GAN: Facial attribute editing based on multi-attention mechanism," *IEEE/CAA J. Automat. Sinica*, vol. 8, no. 9, pp. 1614–1626, Sep. 2020.

[58] A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," 2020, *arXiv:2005.08100*.

[59] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[60] Y. Yang, Z. Ni, M. Gao, J. Zhang, and D. Tao, "Collaborative pushing and grasping of tightly stacked objects via deep reinforcement learning," *IEEE/CAA J. Automatica Sinica*, vol. 9, no. 1, pp. 135–145, Jan. 2022.

[61] K. Samal, H. Kumawat, P. Saha, M. Wolf, and S. Mukhopadhyay, "Task-driven RGB-lidar fusion for object tracking in resource-efficient autonomous system," *IEEE Trans. Intell. Vehicles*, vol. 7, no. 1, pp. 102–112, Mar. 2022.

[62] C.-H. Yeh et al., "Lightweight deep neural network for joint learning of underwater object detection and color conversion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6129–6143, Nov. 2022.

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[64] J. Zhang, L. Pan, Q.-L. Han, C. Chen, S. Wen, and Y. Xiang, "Deep learning based attack detection for cyber-physical system cybersecurity: A survey," *IEEE CAA J. Autom. Sin.*, vol. 9, no. 3, pp. 377–391, Mar. 2022.

[65] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," 2019, *arXiv:1901.02860*.

[66] M. Kaselimi, A. Voulodimos, I. Daskalopoulos, N. Doulamis, and A. Doulamis, "A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 2, 2022, doi: 10.1109/TNNLS.2022.3144791.

[67] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, 2017, pp. 1–11.

[68] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 933–941.

[69] Q. Wei, L. Han, and T. Zhang, "Spiking adaptive dynamic programming based on Poisson process for discrete-time nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 5, pp. 1846–1856, May 2022.

[70] X. Hu, H. Zhang, D. Ma, R. Wang, T. Wang, and X. Xie, "Real-time leak location of long-distance pipeline using adaptive dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 31, 2021, doi: 10.1109/TNNLS.2021.3136939.

[71] X. Yang, D. He, Z. Zhou, D. Kifer, and C. L. Giles, "Learning to read irregular text with attention mechanisms," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, vol. 1, no. 2, p. 3.

[72] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, and S. Zhou, "Focusing attention: Towards accurate text recognition in natural images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5076–5084.

[73] W. Liu, C. Chen, K.-Y. Wong, Z. Su, and J. Han, "STAR-Net: A SpaTial attention residue network for scene text recognition," in *Proc. Brit. Mach. Vis. Conf.*, vol. 2, 2016, p. 7.

[74] F. Bai, Z. Cheng, Y. Niu, S. Pu, and S. Zhou, "Edit probability for scene text recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1508–1516.

[75] Y. Liu, Z. Wang, H. Jin, and I. Wassell, "Synthetically supervised feature learning for scene text recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 435–451.

[76] C.-Y. Lee and S. Osindero, "Recursive recurrent nets with attention modeling for OCR in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2231–2239.

[77] J. Wang and X. Hu, "Gated recurrent convolution neural network for OCR," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 334–343.

[78] J. Baek et al., "What is wrong with scene text recognition model comparisons? Dataset and model analysis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4715–4723.

[79] X. Qi, Y. Chen, R. Xiao, C.-G. Li, Q. Zou, and S. Cui, "A novel joint character categorization and localization approach for character-level scene text recognition," in *Proc. Int. Conf. Document Anal. Recognit. Workshops (ICDARW)*, Sep. 2019, pp. 83–90.

[80] X. Chen, T. Wang, Y. Zhu, L. Jin, and C. Luo, "Adaptive embedding gate for attention-based scene text recognition," *Neurocomputing*, vol. 381, pp. 261–271, Mar. 2020.

[81] Z. Qiao, Y. Zhou, D. Yang, Y. Zhou, and W. Wang, "SEED: Semantics enhanced encoder–decoder framework for scene text recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13528–13537.

[82] D. Yu et al., "Towards accurate scene text recognition with semantic reasoning networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12113–12122.

[83] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," 2014, *arXiv:1406.2227*.

[84] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2315–2324.

[85] A. Mishra, K. Alahari, and C. Jawahar, "Scene text recognition using higher order language priors," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 1–12.

[86] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1457–1464.

[87] S. M. Lucas et al., "ICDAR 2003 robust reading competitions: Entries, results, and future directions," *Int. J. Document Anal. Recognit.*, vol. 7, no. 2, pp. 105–122, 2005.

[88] D. Karatzas et al., "ICDAR 2013 robust reading competition," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 1484–1493.

[89] D. Karatzas et al., "ICDAR 2015 competition on robust reading," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 1156–1160.

[90] T. Q. Phan, P. Shivakumara, S. Tian, and C. L. Tan, "Recognizing text with perspective distortion in natural scenes," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 569–576.

[91] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. Tan, "A robust arbitrary text detection system for natural scene images," *Expert Syst. Appl.*, vol. 41, no. 18, pp. 8027–8048, 2014.

[92] Y. Tian, J. Wang, Y. Wang, C. Zhao, F. Yao, and X. Wang, "Federated vehicular transformers and their federations: Privacy-preserving computing and cooperation for autonomous driving," *IEEE Trans. Intell. Vehicles*, vol. 7, no. 3, pp. 456–465, Sep. 2022.

[93] J. Ma, J. Liu, Q. Lin, B. Wu, Y. Wang, and Y. You, "Multitask learning for visual question answering," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 30, 2021, doi: 10.1109/TNNLS.2021.3105284.

Hui Zhang (Member, IEEE) received the B.S. degree in automation from Beijing Jiaotong University, Beijing, China, in 2015, and the Ph.D. degree in control theory and control engineering from the University of Chinese Academy of Sciences (UCAS), Beijing, in 2020.

From August 2018 to October 2019, she was supported by UCAS as a joint-supervision Ph.D. Student with The University of Rhode Island, Kingston, RI, USA. She is currently a Lecturer with the School of Computer and Information Technology, Beijing Jiaotong University. Her research interests include computer vision, pattern recognition, and intelligent transportation systems.

Guiyang Luo received the Ph.D. degree in computer science and technology from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2020.

He is currently an Associate Researcher with the State Key Laboratory of Networking and Switching Technology, BUPT. His current research interests include machine-type communications and intelligent transportation systems.

Jian Kang received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2013, and the Ph.D. degree from the Department of Electronic Engineering, Tsinghua University, in 2018.

His current research interests include speech recognition and optical character recognition.

Shan Huang received the M.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2015.

She is currently a Researcher with Tencent, Beijing. Her current research interests include optical character recognition and multimodal machine learning.

Xiao Wang (Senior Member, IEEE) received the B.E. degree in network engineering from the Dalian University of Technology, Dalian, China, in 2011, and the M.E. and Ph.D. degrees in social computing from the University of Chinese Academy of Sciences, Beijing, China, in 2013 and 2016, respectively.

She is currently the President of the Qingdao Academy of Intelligent Industries, Qingdao, China, and an Associate Professor with the School of Artificial Intelligence, Anhui University, Hefei, China. She has authored or coauthored more than 60 publications in international refereed journals and conferences. Her research interests include cyber-physical-social systems, social computing, social transportation, and cognitive intelligence.

Fei-Yue Wang (Fellow, IEEE) received the Ph.D. degree in computer and systems engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990.

He joined The University of Arizona, Tucson, AZ, USA, in 1990, and became a Professor and the Director of the Robotics and Automation Laboratory (RAL) and Program in Advanced Research for Complex Systems (PARCS). In 1999, he founded the Intelligent Control and Systems Engineering Center, Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, under the support of the Outstanding Oversea Chinese Talents Program from the State Planning Council and "100 Talent Program" from CAS, and in 2002, he was appointed as the Director of the Key Laboratory of Complex Systems and Intelligence Science, CAS. In 2011, he became the State Specially Appointed Expert and the Director of the State Key Laboratory for Management and Control of Complex Systems, CAS. His research interests include methods and applications for parallel systems, social computing, and knowledge automation.

Dr. Wang is an Elected Fellow of INCOSE, IFAC, ASME, and AAAS. In 2007, he received the 2nd Class National Prize in Natural Sciences of China and awarded the Outstanding Scientist by ACM for his work in intelligent control and social computing. He received the IEEE ITS Outstanding Application and Research Awards in 2009 and 2011, and the IEEE SMC Norbert Wiener Award in 2014. Since 1997, he has been serving as the General or Program Chair for more than 20 IEEE, INFORMS, ACM, and ASME conferences. He was the President of the IEEE Intelligent Transportation Systems (ITS) Society from 2005 to 2007, the Chinese Association for Science and Technology (CAST, USA) in 2005, and the American Zhu Kezhen Education Foundation from 2007 to 2008, and the Vice President of the ACM China Council from 2010 to 2011. Since 2008, he has been the Vice President and the Secretary General of the Chinese Association of Automation. He was the Founding Editor-in-Chief of the *International Journal of Intelligent Control and Systems* from 1995 to 2000 and *IEEE ITS Magazine* from 2006 to 2007, and the Editor-in-Chief (EiC) of the IEEE INTELLIGENT SYSTEMS from 2009 to 2012 and the IEEE TRANSACTIONS ON ITS from 2009 to 2016. He is currently the EiC of China's *Journal of Command and Control*.