Cross-Lingual Text Image Recognition via Multi-Hierarchy Cross-Modal Mimic

Zhuo Chen, Fei Yin, Qing Yang, and Cheng-Lin Liu, *Fellow, IEEE*

***Abstract*—Optical character recognition and machine translation are usually studied and applied separately. In this paper, we consider a new problem named cross-lingual text image recognition (CLTIR) that integrates these two tasks together. The core of this problem is to recognize source language texts shown in images and transcribe them to the target language in an end-to-end manner. Traditional cascaded systems perform text image recognition and text translation sequentially. This can lead to error accumulation and parameter redundancy problems. To overcome these problems, we propose a multihierarchy cross-modal mimic (MHCMM) framework for end-to-end CLTIR, which can be trained with a massive bilingual text corpus and a small number of bilingual annotated text images. In this framework, a plug-in machine translation model is used as a teacher to guide the CLTIR model for learning representations compatible with image and text modes. Via adversarial learning and attention mechanisms, the proposed mimic method can integrate both global and local information in the semantic space. Experiments on a newly collected dataset demonstrate the superiority of the proposed framework. Our method outperforms other pipelines while containing fewer parameters. Additionally, the MHCMM framework can utilize a large-scale bilingual corpus to further improve the performance efficiently. The visualization of attention scores indicates that the proposed model can read text images in a fashion similar to the machine translation model reading text tokens.**

***Index Terms*—Cross-lingual text image recognition, cross-modal mimic, multihierarchy mimic.**

I. Introduction

Texts in images convey rich and high-level semantic knowledge. With the goal of obtaining textual information from images, text image recognition (TIR) has attracted enormous attention from both academia and industry. In recent years, benefiting from deep learning, TIR has been largely advancing through many new methods [1]–[6]. Most TIR methods process text images and output text transcripts constrained within the same language. These methods can be viewed as mono-lingual text image recognition (MLTIR).

Currently, communication among different languages has become popular. Obtaining textual information from text images in foreign languages is needed in applications such as travel guides, document image translation and network information retrieval. OCR converts document images into electronic texts in the original language. However, for users of foreign languages, to obtain the meaning of the document, it is necessary to translate it into the user's native language. Simply cascading OCR and machine translation modules do not provide a good solution because the OCR module can hardly output perfect texts. Thus, we consider a new problem called cross-lingual text image recognition (CLTIR), which aims to recognize source language text images and transcribe corresponding text to the target language, i.e., perform text image recognition and machine translation jointly.

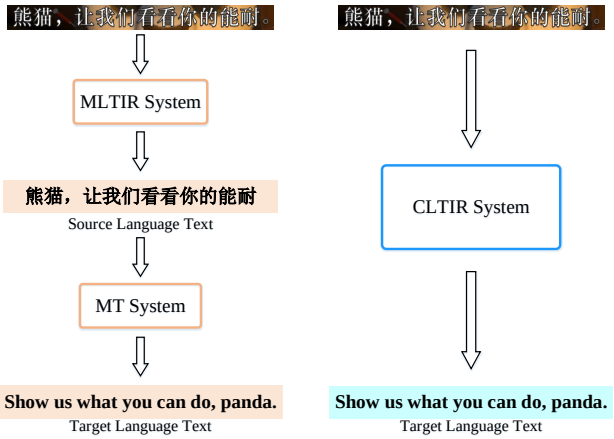

Fig. 1: The pipelines of (a) cascaded and (b) en-to-end cross-lingual text image recognition systems.

To the best of our knowledge, CLTIR has been handled mostly by two-stage cascaded modules: mono-lingual text image recognition (MLTIR) and machine translation (MT). In this manner (Fig. 1(a)), the text image is first fed into the MLTIR system, whose output text, in the source language, is then translated to the target language by the MT system. Much attention has been given to both MLTIR [1]–[6] and MT [7]–[13]. However, the cascaded scheme suffers from some drawbacks. First, the text recognition stage is error prone (especially for degraded or scene text images), and the cascaded system will accumulate errors in two stages. Second, the MLTIR and MT models are designed independently, leading to the problems of parameter redundancy, computational inefficiency, and suboptimal performance. Therefore, a simply assembled cascaded system is insufficient to meet the needs of CLTIR applications. This motivates the design of an end-to-end system that outputs text in the target language directly

This work has been supported by the National Key Research and Development Program Grant 2020AAA0108003 and the National Natural Science Foundation of China (NSFC) grants 61733007 and 61721004. (Corresponding author: Cheng-Lin Liu.)

Zhuo Chen, Fei Yin, Qing Yang and Cheng-Lin Liu are with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation of Chinese Academy of Sciences, 100190, P.R. China, and the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, P.R. China. (e-mail: {zhuo.chen, fyin, qyang, liucl}@nlpr.ia.ac.cn)

from images in the source language.

An end-to-end CLTIR system (Fig. 1(b)), without producing intermediate results, can better exploit the mutual information between source-language text images and target-language text. Nevertheless, it is a great challenge to integrate heterogeneous knowledge from text recognition and translation tasks, which are embedded in the different modes of image and text. Therefore, a roughly designed end-to-end method cannot produce satisfactory results, and this will be discussed in detail in Section V. To accomplish end-to-end CLTIR, a multitask learning framework is proposed in [14], where CLTIR and MLTIR tasks are trained simultaneously to exploit the information in both source and target languages. However, such end-to-end methods, either heavily rely on the amount of triplet data, i.e., triplets of text image - source-language label - target-language label. The acquisition of such triplet data is very laborious, which obstructs the development of CLTIR.

To overcome the insufficiencies of cascaded systems and the lack of annotated triplet data, we propose a novel multi-hierarchy cross-modal mimic (MHCMM) framework for end-to-end CLTIR, as shown in Fig. 2. In the MHCMM, an end-to-end CLTIR model is used as a student to mimic a text-to-text MT model, which can be pretrained and benefit from the availability of a large bilingual text corpus. The cross-modal mimic strategy, involving text and image modes, integrates both global and local features to learn multihierarchy knowledge. (1) In the global sense, we utilize adversarial learning to guide the student to match an overall distribution of the teacher to narrow down the representation gap between image and text modes. (2) In the local sense, with an attention mechanism, the student is forced to focus on the image region corresponding to the text region read by the teacher in every time step. In this multihierarchy manner, the student can learn the capacity of semantic comprehension from the teacher. Moreover, the system can be trained with a massive existing bilingual text corpus and only a small number of bilingual annotated text images.

To evaluate the performance of our proposed method, we collected a CLTIR dataset with text images and corresponding text labels in both source and target languages, named the bilingual annotation text image dataset (BLATID)[1]. Extensive experiments have been conducted to verify the superiority of the proposed method. The MHCMM method significantly outperforms baseline systems and can be further improved with a large-scale bilingual text corpus. The visualization of attention scores indicates that the student model in MHCMM has learned how to read text in source language images from the text-to-text MT model.

The major contribution of this paper lies in the following aspects:

- We study a new and challenging problem named cross-lingual text image recognition (CLTIR), which is encountered in many applications.
- We propose a cross-modal mimic framework to guide the end-to-end CLTIR model with a typical machine translation model as the teacher to benefit from a bilingual text corpus.
- Based on adversarial learning and attention mechanisms, the proposed MHCMM framework can integrate both global and local knowledge for multihierarchy mimics.
- We have constructed a dataset for CLTIR, named BLATID and performed extensive experiments on the dataset. The results demonstrate the superiority of the proposed method.

The remainder of this paper is organized as follows. Section II reviews related works. Section III presents preliminary knowledge for this study. Section IV introduces our proposed approach named MHCMM. Section V presents the experimental results and analysis. Section VI contains the concluding remarks.

II. Related Works

A. Text Image Recognition

Text image recognition has achieved enormous progress over the past decades [1]–[6], [15], [16]. Graves et al. [1] propose a handwritten text recognition method based on long short-term memory (LSTM) and connectionist temporal classification (CTC), which outperforms previous methods based on hidden Markov models (HMMs). Wu et al. [16] further extend the LSTM-based method to Chinese handwriting recognition. To address multiple scripts, Chen et al. [4], [15] propose networks for simultaneous text recognition and script identification in a multitask learning framework.

In addition to handwritten text recognition, many efforts have been devoted to scene text detection [17]–[22] and recognition [2], [3], [6], [23] in camera-captured images. For scene text recognition, Shi et al. [2] propose a convolutional recurrent neural network (CRNN), which uses convolutional layers and recurrent layers to exploit both spatial and temporal information. This framework has been adopted by many text recognition tasks and extended to many other sequence learning tasks. Yin et al. [3] propose a sliding convolutional character model for text recognition, which eliminates the low efficiency of sequential computation in RNNs. The 'Aster' in [23] is a method for irregular text recognition based on an attention mechanism in a two-stage manner, with two modules for rectification and recognition . Lin et al. [6] propose the 'STAN' for general scene text recognition by composing a sequential transformation network and an attention-based recognition network and achieve state-of-the-art performance on both regular and irregular text images.

Although these methods can achieve promising results, they are primarily designed for mono-lingual text image recognition. In other words, text image and output text are in the same language.

B. Machine Translation.

Recent machine translation methods are mostly based on recurrent neural networks (RNNs) and encoder-decoder frameworks [10]–[13]. The method in [24], [25] compresses all the necessary information of source sentences into fixed-length vectors. This limitation may make it difficult to cope

[1]The BLATID dataset is released at http://www.nlpr.ia.ac.cn/pal/Dataset/BLATID.html

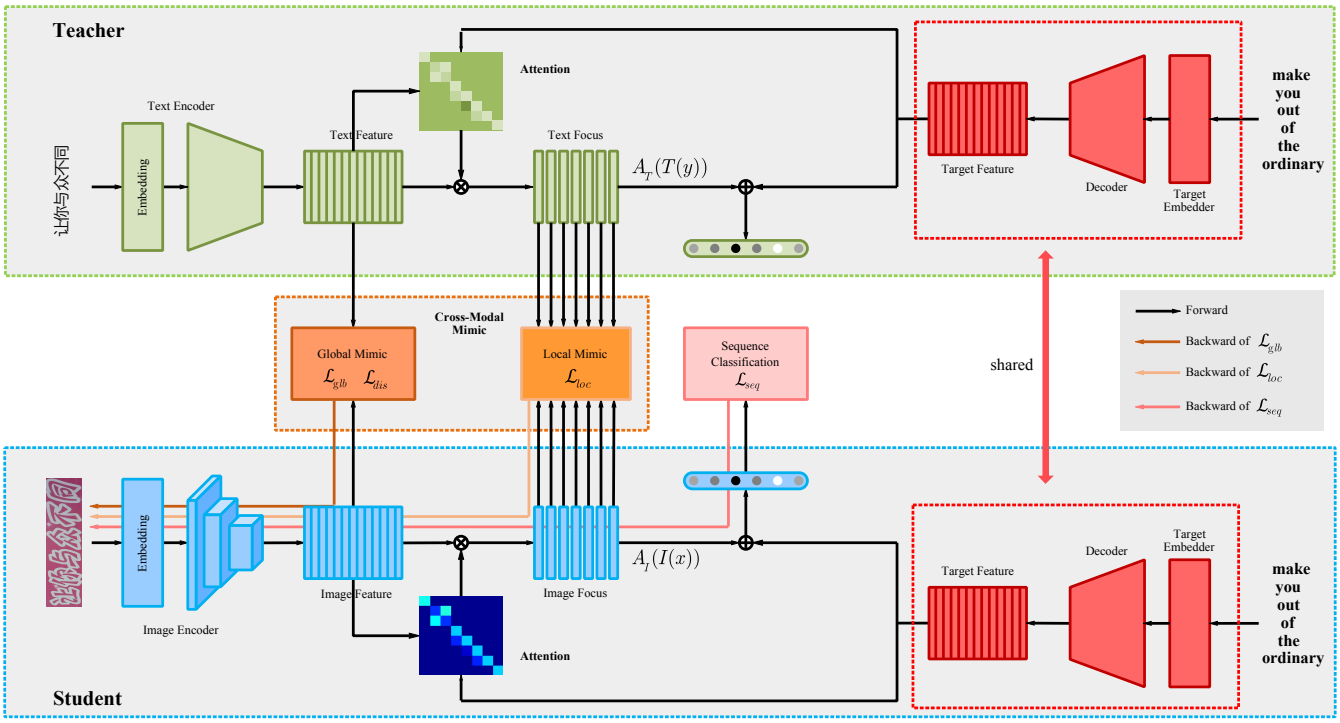

Fig. 2: Framework of the proposed MHCMM The student, encompassed by the blue box, is guided by the teacher within the green frame. The decoder is shared by both the teacher and the student. The optimization of students is supervised by global mimic ($\mathcal{L}_{glb}$, $\mathcal{L}_{dis}$), local mimic ($\mathcal{L}_{loc}$), and sequence classification ($\mathcal{L}_{seq}$). The decoder is used in the teacher model only in the pretraining stage.

with variable-length sequences. To address this issue, an attention model is proposed in [26] by adding an alignment module, named the attention mechanism, to the RNN-based encoder-decoder framework. However, due to the long-range temporal dependencies, RNN-based methods suffer from the low efficiency of sequential computation. To overcome this, Gehring et al. [7] propose a convolution-based attention model, which uses convolutional layers as basic building blocks and computes hidden representations in parallel for all input and output sequences. Meanwhile, Vaswani et al. [8] introduce the transformer, which solely relies on attention mechanisms rather than recurrent or convolutional units. Based on the work in [8], BERT is proposed in [9] to pretrain a deep language model under a self-supervision scheme. After pretraining, the model can be applied to a wide range of tasks. Similarly, Lin et al. [27] pretraine a universal multilingual neural machine translation model on dozens of language pairs jointly. Then, the model is fine-tuned for different language pairs to obtain specific MT models. Wang et al. [28] introduce adversarial noise to the output embedding layer. This regularization strategy improves the neural language model effectively and efficiently. To reduce the parameter and computational cost, Mehta et al. [29] propose a deep and lightweight transformer architecture.

Inspired by the above methods, we utilize the widely used attention mechanism in [7] to develop our CLTIR system. In addition, an extended version of our MHCMM framework will be developed based on Transformer [8] to verify its generalizability.

C. Mimic and Distillation.

Network mimic and knowledge distillation were originally proposed in [30], [31] to guide the compact student model to learn knowledge from the output of a large teacher model. Sometimes, the data of teacher and student models may be from different modalities. In this case, Gupta et al. [32] adopt mimics to transfer supervision between images from different modes. Similarly, the idea of utilizing paired samples to transfer knowledge between teachers and students has been widely used for cross-modal applications [33]–[37]. Albanie et al. [33] transfer knowledge from the image domain to the audio domain for emotion recognition. To perform unsupervised domain adaptation, Kundu et al. [36] adopt a cross-task distillation module to explore intertask coherency. In [37], knowledge is interpreted as priors on the parameters of the student model, and cross-modal knowledge is efficiently transferred between source and target datasets .

The above mimic methods, to the best of our knowledge, are limited to fixed-length feature representation. In this paper, we propose a novel multihierarchy mimic method for variable-length features in sequence learning.

D. Adversarial Learning

Goodfellow et al. [38] proposed the generative adversarial net (GAN), which consists of a generative Model G and a

discriminative Model D, to establish a minimax two-player game in an adversarial manner. Many variants of GANs have been successfully applied to many fields. In [39], a conditional GAN is proposed to generate samples conditioned on specific knowledge. To conduct unpaired generation, cycle consistency loss is introduced in CycleGAN [40] to establish a bidirectional mapping.

In addition to the generation task, the adversarial learning method has been adopted for various scenarios. To name a few, the approach introduced in [41] tries to learn joint multimodal representation by matching the posterior distribution of the representation to the given prior. Wu et al. [42] propose a novel mathematical expression recognition model based on a paired adversarial learning method to learn semantic-invariant representations. In [43], a multitask adversarial learning method is used to force the recognizer to explore the dependencies among multiple face analysis tasks from the label level.

Inspired by the effectiveness of adversarial learning, we adopt it to construct the MHCMM framework for learning representations compatible with image and text modes.

III. Preliminary Knowledge

A. Attention Mechanism

The attention mechanism involves three essential factors: *query*, *key*, and *value*. The key and value are in pairs, while the query is used for indexing. Usually, a weight vector is computed according to the similarity between the query and corresponding key. Then, the output is produced by weighting the values. In this manner, an attention mechanism is frequently adopted to align the source and target sequences.

Instead of relying on RNNs to compute intermediate encoder states $\boldsymbol{e}$ and decoder states $\boldsymbol{d}$, the convolutional sequence-to-sequence learning (ConvS2S) model [7] utilizes a fully convolutional architecture for achieving higher accuracy and reducing time consumption. In ConvS2S, both the encoder and the decoder are based on a stack of gated convolutional layers to compute intermediate states. We denote the outputs of the h-th encoder and the l-th decoder block as $\boldsymbol{e}^h = \left(\boldsymbol{e}_1^h, \boldsymbol{e}_2^h, ..., \boldsymbol{e}_m^h\right)$ and $\boldsymbol{d}^l = \left(\boldsymbol{d}_1^l, \boldsymbol{d}_2^l, ..., \boldsymbol{d}_n^l\right)$, respectively. Each element $\boldsymbol{e}_j^h$ integrates knowledge of k elements $\left(\boldsymbol{e}_i^{h-1}, \boldsymbol{e}_{i+1}^{h-1}, ..., \boldsymbol{e}_{i+k-1}^{h-1}\right)$ from the $(h-1)$-th layer when the kernel size is k. By stacking layers, the effective context size of the top feature is greatly increased. This allows the model to tackle the maximum length of the sequence.

The basic element of the attention mechanism is *attention score*, which denotes the relevancy between source states (key) and target elements (query). For the l-th decoder layer, the attention score α_{ij}^l can be formulated as:

$$\alpha_{ij}^l = \frac{exp\left(\hat{\boldsymbol{d}}_i^l \cdot \boldsymbol{e}_j^h\right)}{\sum_{t=1}^m exp\left(\hat{\boldsymbol{d}}_i^l \cdot \boldsymbol{e}_t^h\right)}, \tag{1}$$

where $(\cdot)$ represents the dot product, $\boldsymbol{e}_j^h$ represents the j-th output of the last encoder layer h, and $\hat{\boldsymbol{d}}_i^l$ is computed by combining the decoder output $\boldsymbol{d}_i^l$ and the previous target state $\boldsymbol{g}_i$: $\hat{\boldsymbol{d}}_i^l = \boldsymbol{W}_d^l \boldsymbol{d}_i^l + \boldsymbol{b}_d^l + \boldsymbol{g}_i$, where $\boldsymbol{W}_d^l$ and $\boldsymbol{b}_d^l$ represent the weight and bias for the learnable linear transformation.

To align encoder and decoder sequences, the attention score is utilized to integrate encoder states ($\boldsymbol{e}_j^h$) and input embedding ($\boldsymbol{m}_j$) by a weighted sum:

$$\boldsymbol{c}_i^l = \sum_{j=1}^m \alpha_{ij}^l \left(\boldsymbol{e}_j^h + \boldsymbol{m}_j\right). \tag{2}$$

In this manner, the method can determine which encoder states are of more importance to each decoder state. Then, $\boldsymbol{d}_i^L$ and $\boldsymbol{c}_i^L$, outputs of the decoder, are combined and fed to a fully connected layer with weight $\boldsymbol{W}_0$ and bias $\boldsymbol{b}_0$. After mapping to the category, the next word y_{i+1} can be predicted based on the previous results:

$$p\left(y_{i+1}|y_1, ..., y_i, \boldsymbol{x}\right) = Softmax\left(\boldsymbol{W}_0\left(\boldsymbol{d}_i^L \oplus \boldsymbol{c}_i^L\right) + \boldsymbol{b}_0\right). \tag{3}$$

Similar to ConvS2S, Transformer [8] also uses an attention mechanism for sequence alignment. For feature extraction, it performs a self-attention module, where each layer transcribes the input feature to the query, key, and value simultaneously. In this manner, each layer can capture the global representation of the sequence. Thus, it is easier for the transformer to learn long-range dependencies.

In our work, we design an attention-based framework for CLTIR, named MHCMM. For generalization, the MHCMM is developed for both the ConvS2S and transformer versions.

B. Adversarial Learning

Adversarial learning, originally proposed in [38] with GAN, has been applied to many learning tasks [44]–[46]. The basic algorithm of GAN is a two-player minimax game between a generator G and a discriminator D. D is designed to estimate whether the sample is from the real world or generated by G. On the other hand, G manages to transform the input data into a realistic sample. The overall objective of adversarial learning is formulated as:

$$\begin{aligned}\min_G \max_D V\left(G, D\right) =& \mathbb{E}_{\boldsymbol{x} \in p_{data}}\left[\log\left(D\left(\boldsymbol{x}\right)\right)\right] \\ &+ \mathbb{E}_{\boldsymbol{z} \in p_{noise}}\left[\log\left(1 - D\left(G\left(\boldsymbol{z}\right)\right)\right)\right],\end{aligned} \tag{4}$$

where $\boldsymbol{x} \in p_{data}$ and $\boldsymbol{z} \in p_{noise}$ are the real sample and noise, respectively.

For some nongenerative tasks, G will be replaced by other modules, while D is still utilized to guide the system to learn a distribution invariant knowledge. Based on this, extensions of adversarial learning have been applied to various fields involving text and image [47]–[49] to learn an overall distribution. Inspired by this, we use adversarial learning to learn the cross-modal feature representation in CLTIR.

IV. Methodology

Our end-to-end CLTIR model inputs text images in the source language and outputs transcripts in the target language. While in training, it allows us to leverage a mass of bilingual text corpus. In the following, we first illustrate the rationale of the cross-modal mimic framework and then describe the details of the CLTIR system.

A. Rationale

To formulate the CLTIR problem, we denote by $\boldsymbol{X}$, $\boldsymbol{Y}$ the image and text in the source language and $\boldsymbol{Z}$ denote the corresponding text in the target language. Under Bayesian decision theory, the goal of our system is to find a mapping function to maximize the conditional probability $P(\boldsymbol{Z}|\boldsymbol{X})$.

For cascaded systems, $P(\boldsymbol{Z}|\boldsymbol{X})$ can be decomposed as:

$$P(\boldsymbol{Z}|\boldsymbol{X}) = \sum_{\boldsymbol{y}\in\boldsymbol{Y}} P(\boldsymbol{y}|\boldsymbol{X})P(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{y}). \tag{5}$$

As the two modules, text recognizer $P(\boldsymbol{y}|\boldsymbol{X})$ and translator $P(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{y})$, are separate, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are conditionally independent when $\boldsymbol{Y}$ is given. Therefore, Eq. (5) can be simplified as:

$$\begin{aligned} P(\boldsymbol{Z}|\boldsymbol{X}) &= \sum_{\boldsymbol{y}\in\boldsymbol{Y}} P(\boldsymbol{y}|\boldsymbol{X})P(\boldsymbol{Z}|\boldsymbol{y}) \\ &= P(\boldsymbol{Y}|\boldsymbol{X})P(\boldsymbol{Z}|\boldsymbol{Y}). \end{aligned} \tag{6}$$

$P(\boldsymbol{Y}|\boldsymbol{X})$ and $P(\boldsymbol{Z}|\boldsymbol{Y})$ represent the MLTIR and MT models, respectively. Obviously, $P(\boldsymbol{Z}|\boldsymbol{X})$ heavily relies on both $P(\boldsymbol{Y}|\boldsymbol{X})$ and $P(\boldsymbol{Z}|\boldsymbol{Y})$, implying the error accumulation of cascaded systems. Ideally, the MLTIR module is perfect when $P(\boldsymbol{Y}|\boldsymbol{X}) = 1$. Then, the performance of the CLTIR totally depends on the MT model thereby eliminating the error accumulation problem. However, this is not realistic.

To alleviate the impact of $P(\boldsymbol{Y}|\boldsymbol{X})$, we design an end-to-end method without generating intermediate results and make the system learn compatible features from image and text modes.

Based on Bayesian modeling, by introducing latent features F in the MT system, we can represent the CLTIR system by $P(\boldsymbol{Z}|\boldsymbol{F},\boldsymbol{Y})$. As $\boldsymbol{F}$ is an intermediate representation, we can assume that modules $\mathcal{F}(\boldsymbol{Y}\rightarrow\boldsymbol{F})$ and $\mathcal{G}(\boldsymbol{F}\rightarrow\boldsymbol{Z})$ are separate. Thus, $\boldsymbol{Y}$ is irrelevant when calculating $\boldsymbol{Z}$, which only depends on $\boldsymbol{F}$:

$$P(\boldsymbol{Z}|\boldsymbol{F},\boldsymbol{Y}) \approx P(\boldsymbol{Z}|\boldsymbol{F}). \tag{7}$$

By analogy, denoting the feature representation of X as $\boldsymbol{F'}$, $P(\boldsymbol{Z}|\boldsymbol{F'},\boldsymbol{X})$ in the end-to-end CLTIR system can be formulated as:

$$P(\boldsymbol{Z}|\boldsymbol{F'},\boldsymbol{X}) \approx P(\boldsymbol{Z}|\boldsymbol{F'}). \tag{8}$$

According to Eq. (7) and Eq. (8), to make $P(\boldsymbol{Z}|\boldsymbol{F'},\boldsymbol{X})$ approximate $P(\boldsymbol{Z}|\boldsymbol{F'},\boldsymbol{Y})$, it is desired that the two latent representations are compatible:

$$P(\boldsymbol{Z}|\boldsymbol{F'}) = P(\boldsymbol{Z}|\boldsymbol{F}). \tag{9}$$

In this way, the end-to-end CLTIR system will reach the performance of the MT model as $P(\boldsymbol{Z}|\boldsymbol{F'},\boldsymbol{X}) \approx P(\boldsymbol{Z}|\boldsymbol{F},\boldsymbol{Y})$. Based on the analysis above, we propose a novel framework to learn feature representations that are compatible between image and text modes.

B. Framework

To be consistent with Eq. (9), MHCMM is designed for involving feature extraction for both image mode and text mode. As shown in Fig. 2, the text mode encoder network, within the green box, plays the role of teacher, while the image mode encoder, within the blue box, is a student model. The decoder module, in red, is shared by both the teacher and the student. The parameters of the decoder are first optimized during the pretraining of the teacher. Then, the student will inherit the parameters of the decoder and finetune them when training with text image data.

Teacher. The teacher, corresponding to $P(\boldsymbol{Z}|\boldsymbol{F},\boldsymbol{Y})$ in Eq. 7, guides the student. It is plug-in and can be a text machine translation model. During training, the teacher embeds text tokens utilizing a learnable lookup table and encodes the feature by 1D convolution layers to capture semantic knowledge. Specifically, the encoder of the teacher is designed with the pipeline in [7] with eight convolution layers and dimensionality of 256. Meanwhile, the decoder module will extract features of the target sequence. With text and target feature representations, the attention mechanism will compute the attention score, which is used to weight and sum the text feature sequence for each target state to obtain the attention region, named the *text focus*. Then, the states of the target feature and corresponding text focus are integrated to predict the result sequence.

Student. Corresponding to $P(\boldsymbol{Z}|\boldsymbol{F'},\boldsymbol{X})$ in Eq. 8, the student is an end-to-end CLTIR model. It inputs text images rather than text tokens. Meanwhile, the *image encoder* module, as shown in Table I, aims to encode images to the feature map that is compatible with text features from the teacher.

Before being fed into the network, all images are normalized to the same height with the aspect ratio unchanged. Then, image features are extracted by the image encoder, which is constructed by stacking convolutional layers inspired by VGG19 [50]. Finally, the image features are collapsed along the dimension of height to obtain feature sequence $\boldsymbol{f} \in \mathbb{R}^{t\times d}$, where t and d are the sequence length and feature dimensionality, respectively. In this manner, the input images are transformed into a high-level global representation. Furthermore, the image encoder is flexible and can be altered by advanced architectures.

After feature extraction, the image feature representation is fed into the attention module and aligned referring to the target feature in a similar way to the teacher. During training, the student is supervised under *cross-modal mimic* and *sequence classification* derived from the teacher and true labels, respectively. To guide the student to learn compatible features with the teacher, the cross-modal mimic is a multihierarchy that leverages both global and local knowledge. Meanwhile, the sequence classification objective is adopted to complete the overall optimization to achieve better recognition accuracy.

C. Global Mimic

For both MT and CLTIR tasks, the semantic knowledge is encoded in the high-level feature space. To guide the student to

TABLE I: Details of the image encoder. '#Out' represents the number of output feature maps, 'K', 'S', and 'P' stand for kernel size, stride, and padding size, respectively.

Type	Configurations	#Layers
Input	H(64)*W*1(gray image)	1
Convolution	#Out:64, K:3×3, S:1×1, P:1×1, ReLU	2
BatchNorm	-	
MaxPooling	K:2×2, S:2×2	1
Convolution	#Out:128, K:3×3, S:1×1, P:1×1, ReLU	2
BatchNorm	-	
MaxPooling	K:2×2, S:2×2	1
Convolution	#Out:256, K:3×3, S:1×1, P:1×1, ReLU	4
BatchNorm	-	
MaxPooling	K:2×2, S:2×2	1
Convolution	#Out:512, K:3×3, S:1×1, P:1×1, ReLU	4
BatchNorm	-	
MaxPooling	K:2×2, S:2×2	1
Convolution	#Out:512, K:3×3, S:1×1, P:1×1, ReLU	4
BatchNorm	-	
MaxPooling	K:2×2, S:2×2	1
AveragePooling	K:2×1, S:2×1	1
Linear	#Out:256	1

learn compatible knowledge with the teacher, a global mimic is established between text (teacher) and image (student) modes.

Both text images from the student and text tokens from the teacher are of variable length. Furthermore, as text image $\boldsymbol{x}$ and its corresponding text tokens $\boldsymbol{y}$ usually have different lengths and are without alignment, it is infeasible to adopt elementwise matching. Hence, we adopt adversarial learning to conduct a global mimic between the distributions of the two modes. Inspired by the generative adversarial network (GAN) [38], which can model complex data distributions efficiently, we utilize adversarial learning to encourage the image encoder of the student to learn fine and compatible semantic features.

Specifically, we construct a discriminator D, a multilayer perceptron (MLP) with one hidden layer of 512 neurons, to recognize whether each frame of the feature sequence is from the text encoder T or the image encoder I. Meanwhile, I tries to confuse D by producing images with a more compatible representation. Suppose that $\boldsymbol{f}_i$ is the i-th frame of the feature sequence derived from the text image and $\boldsymbol{g}_k$ is the k-th frame extracted from the text tokens. The loss functions for framewise adversarial learning can be formulated as:

$$\mathcal{L}_{dis}(T,I,D) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\in(\boldsymbol{X},\boldsymbol{Y})}\left[\mathbb{E}_{\boldsymbol{f}_i\in I(\boldsymbol{x})}\left[\log\left(1-D\left(\boldsymbol{f}_i\right)\right)\right] + \mathbb{E}_{\boldsymbol{g}_k\in T(\boldsymbol{y})}\left[\log\left(D\left(\boldsymbol{g}_k\right)\right)\right]\right], \tag{10}$$

$$\mathcal{L}_{glb}(I,D) = -\mathbb{E}_{\boldsymbol{x}\in\boldsymbol{X}}\left[\mathbb{E}_{\boldsymbol{f}_i\in I(\boldsymbol{x})}\left[\log\left(1-D\left(\boldsymbol{f}_i\right)\right)\right]\right]. \tag{11}$$

L_{dis} and L_{glb} are used in training for updating the discriminator and the image encoder, respectively.

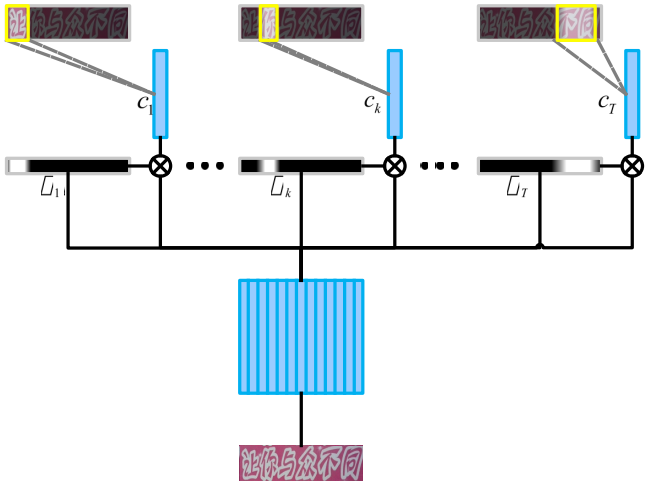

Fig. 3: The mechanism of attention in an end-to-end model for CLTIR.

D. Local Mimic

With the global mimic, the holistic similarity of image and text features are improved. However, the global mimic is only concerned with the overall distribution while neglecting details of features in semantic space. To balance the global distribution and local details, we also construct a local mimic for each weighted frame derived from the attention mechanism.

As illustrated in Section III-A, the feature sequence of the text image can be represented as $\boldsymbol{e} \in \mathbb{R}^{t\times d}$, where t and d are the length and feature dimensionality of the sequence. As shown in Fig. 3 and Eq. (2), the attention score α is utilized to obtain a merged feature $\boldsymbol{c}_i \in \mathbb{R}^d$ by weighting the feature sequence. The merged feature corresponds to a highlighted region of the original text image, which illustrates where the model pays more attention at the current time step. Thus, we call the feature *image focus* as it is similar to the focus of eyes when reading.

In this manner, we can capture the focus regions of the attention mechanism for both image and text modes at every time step. As the feature sequences from both image and text modes are mapped to fixed length and same dimensionality, a local mimic is established for elementwise matching. Suppose the input of image and text modes are $\boldsymbol{x}$ and $\boldsymbol{y}$. Then the image and text features fed into the attention module A_* can be represented as $I\left(\boldsymbol{x}\right)$ and $T\left(\boldsymbol{y}\right)$, where I and T are the text and image encoder, respectively. Thus, with mean squared error, the loss of local mimic can be formulated as:

$$\mathcal{L}_{loc}(I) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\in(\boldsymbol{X},\boldsymbol{Y})}\left[\left\|A_T\left(T(\boldsymbol{y})\right) - A_I\left(I(\boldsymbol{x})\right)\right\|^2\right], \tag{12}$$

where $A_I\left(I\left(\boldsymbol{x}\right)\right)$ and $A_T\left(T\left(\boldsymbol{y}\right)\right)$ produce the feature sequence $\boldsymbol{c}$ of the corresponding mode. In this manner, the student is guided to generate a series of local knowledge to be compatible with that from the teacher based on the attention mechanism.

E. Sequence Classification

The ultimate goal of CLTIR is to map text images to texts of the target language. Thus, in addition to the multihierarchy cross-modal mimic conducted in the semantic feature space, we also optimize the student model under the supervision of sequence classification. The loss can be expressed as the negative log-likelihood of prediction probability (also called cross-entropy, CE):

$$\mathcal{L}_{seq}(I) = -\mathbb{E}_{(\boldsymbol{x},\boldsymbol{z})\in(\boldsymbol{X},\boldsymbol{Z})}\left[\log\left(P\left(\boldsymbol{z}|A_I\left(I(\boldsymbol{x})\right)\right)\right)\right]. \quad (13)$$

F. Optimization

To train the whole model, we integrate all the loss functions involved in the encoder and decoder as:

$$\begin{aligned}(I^*, D^*) = \quad & \arg\min_I \max_D \{\lambda_{glb}[\mathcal{L}_{dis}(T, I, D) + \mathcal{L}_{glb}(I, D)] \\ & +\lambda_{loc}\mathcal{L}_{loc}(T, A_T, I, A_I) + \lambda_{seq}\mathcal{L}_{seq}(I, A_I)\}.\end{aligned} \quad (14)$$

During training, we optimize the model in an iterative process, as shown in Algorithm 1. In each training iteration, the procedure can be divided into three steps. First, with the global mimic of the teacher model, the image encoder I is optimized by playing the minimax game with D under $\mathcal{L}_{dis}$ and $\mathcal{L}_{glb}$. Meanwhile, the attention module will capture each key region to obtain the text focus and image focus. Next, local mimic is conducted to train I by minimizing MSE between features from text and image modes. Then, the sequence classification loss is computed based on the true labels and is used to update the parameters of the encoder and decoder. Under the supervision of the above three parts, MHCMM will aggregate all the gradients before updating the weights of the student model.

G. Inference

After training, the student model can be used for CLTIR while the teacher is no longer needed. The inference process is similar to that of typical MT systems. Taking a text image, the encoder will extract high-level semantic features. Then, with the attention mechanism, the decoder will pay attention to the image focus and predict a token for each time step. Sequentially, the model will leverage the overall image feature map and previously predicted token to output the results.

V. Experiments

Since there was no public dataset for CLTIR, we collected a new dataset for evaluating our proposed method. In the following, we first introduce the details of the dataset and then describe the experimental settings. We then present the results and their analysis.

A. Dataset

Our system is proposed for CLTIR, but there was no existing text recognition dataset with labels in another language. Hence, we constructed a bilingual annotation text image dataset (BLATID), as specified in Table II. In building the dataset, we took advantage of the existing machine translation corpus to save labor. The AI Challenger dataset (AIC) [51] consists of approximately 12M Chinese-English sentence pairs for training and 7.8k for validation. Inspired by [52], we synthesized Chinese text images referring to the corpus in AIC. In this manner, we could obtain plenty of triplet data containing triplets of Chinese text image - Chinese text label - English text label. We only generated 1M triplet data for training the student, while the remaining bilingual sentences can be used to supervise the teacher, as discussed in Section V-D. Considering the difficulty of collecting text images in practice, the scale of 1M text images in our dataset is significant. In addition, to increase the diversity of the validation set, we adopt 7.8K sentence pairs in the corpus to generate 100K triplet data with different graphic configurations.

On the other hand, the test set was derived from movies and their related bilingual subtitles, which are widely used in our daily life. We collected English-Chinese bilingual subtitles from 50 English animated films, where Chinese sentences

Algorithm 1 Optimization for MHCMM

Input: Dataset $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \in (\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, the initialized text encoder T, image encoder I and discriminator D are parameterized by $\theta_T, \theta_I, \theta_D$

Output: the optimized I, D and A_I are parameterized by $\hat{\theta}_I, \hat{\theta}_D, \hat{\theta}_{A_I}$

Initialize the image encoder I and discriminator D randomly.
Initialize the text encoder T and decoder with pretrained machine translation model.
repeat
  for number of training epochs **do**
    for number of mini-batches **do**
      // Global mimic for discriminator D
      $\theta_D \leftarrow \theta_D - \mu\lambda_{glb}\frac{\partial\mathcal{L}_{dis}(\boldsymbol{X},\boldsymbol{Y};\theta_T,\theta_I,\theta_D)}{\partial\theta_D}$
      // Global mimic for image encoder I
      $\theta_I \leftarrow \theta_I - \mu\lambda_{glb}\frac{\partial\mathcal{L}_{glb}(\boldsymbol{X};\theta_I)}{\partial\theta_I}$
      // Local mimic for image encoder I
      $(\theta_I, \theta_{A_I}) \leftarrow (\theta_I, \theta_{A_I}) - \mu\lambda_{loc}\frac{\partial\mathcal{L}_{loc}\left(\boldsymbol{X},\boldsymbol{Y};(\theta_I,\theta_{A_I})\right)}{\partial\left(\theta_I,\theta_{A_I}\right)}$
      // Sequence classification for image encoder I
      $(\theta_I, \theta_{A_I}) \leftarrow (\theta_I, \theta_{A_I}) - \mu\lambda_{sec}\frac{\partial\mathcal{L}_{seq}\left(\boldsymbol{X};\left(\theta_I,\theta_{A_I}\right)\right)}{\partial\left(\theta_I,\theta_{A_I}\right)}$
    end for
  end for
until convergence, got $\hat{\theta}_I = \theta_I$, $\hat{\theta}_D = \theta_D, \hat{\theta}_{A_I} = \theta_{A_I}$
return $\hat{\theta}_I, \hat{\theta}_D, \hat{\theta}_{A_I}$

TABLE II: Details of the newly collected BLATID dataset. '#corpus' means number of Chinese-English sentences, and '#samples' stands for number of triplets of "Chinese text image - Chinese text label - English text label".

Subset	AIC	BLATID	
		#corpus	#samples
Training	12M	1M	1M
Validation	7.8K	7.8K	100K
Test	—	55K	55K

are printed on corresponding frames based on timestamps. To increase the diversity of the dataset, we also randomly changed the subtitle setting, such as font, text size, shadow type, and position. In addition, with position information, subtitle regions can be cropped automatically, saving the efforts for text detection.

B. Implementation Details

As described in Section III-A, we adopt two dominating attention methods, ConvS2S and Transformer, to construct our system. We evaluate the variable configurations of MHCMM based on ConvS2S and then evaluate its generalizability by extending it to the transformer-based version.

All experiments in this paper are implemented on the PyTorch [53] platform. The cropped text images are converted to grayscale images and normalized to a fixed height of 64 with the aspect ratio unchanged. The MHCMM system is constructed with a text encoder and a decoder of 20 layers, all of which are configured with 256 hidden units. We train the networks using Adam [54] with a learning rate of 0.001 and a dropout rate of 0.1 and renormalize the gradients if their norm exceed 0.1. During training, datasets are distributed in parallel on 4 Titan XP GPUs for approximately 4 hours per epoch. Since the mimic procedure is skipped during inference, the student model of MHCMM can recognize over 250 English sentences from corresponding Chinese text images per second with a batch size of 24. Obviously, the running speed of MHCMM can meet the needs of daily use for ordinary people. To assess the performance, we calculate BLEU scores [55] for the CLTIR task.

C. Performance of MHCMM

To evaluate the performance of MHCMM, we conduct experiments on the following models (illustrated in Fig. 4) for comparison:

- Cascaded system: The system consists of separated MLTIR and MT models for two-stage CLTIR. In the first stage, the MLTIR model outputs the text string of the source language, which is then fed into the MT model in the second stage. For a fair comparison, the MT model is the same as the teacher in MHCMM.
- Single-task system: The system inputs images in the source language while outputting texts in the target language in an end-to-end flow. It takes the same architecture as the student in MHCMM and only differs in training strategy in that there is no multihierarchy cross-modal mimic.
- Multitask system: This is similar to the single-task system but adds a branch outputting source-language text in addition to the branch outputting target-language text.

For fair comparison, all the abovementioned CLTIR systems were trained using the BLATID dataset with the same configuration. The teacher in the MHCMM and the MT model in the cascaded system are the same translation model, which is developed with a 1M bilingual corpus corresponding to the generated images. The experimental results are shown in the top block of Table III.

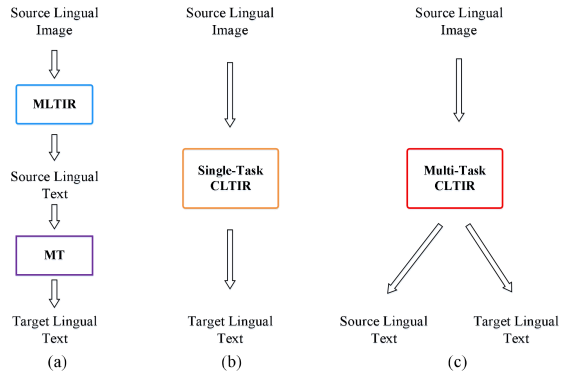

Fig. 4: The framework of comparison systems. (a) Cascaded system. (b) Single-task system. (c) Multitask system.

The first row in Table III shows the BLEUs of the translation model. When taking the source-language text as input, the model can achieve a BLEU of 22.61. In the cascaded system, the translation model takes the text recognition result as input. In this case, the obtained BLEU is only 19.92, which is much lower than that of machine translation. This is due to the accumulation of text recognition errors.

Although the end-to-end single-task system can alleviate the above problem, it is difficult to learn semantic features from text images directly. Thus, its performance is much worse than the cascaded baseline. The multitask system conducting MLTIR and CLTIR outperforms the single-task model by over 2 BLEU scores, which indicates that the CLTIR can benefit from the MLTIR task.

Obviously, the system with MHCMM achieves the best performance among all systems. It outperforms the cascaded system by approximately 1.0 BLEU. It is noteworthy that the MHCMM and single-task system have the same architecture during inference but learn in different ways. With the multi-hierarchy mimic, the BLEU of our method is improved by 4.08 and 2.46 for the validation and test sets, respectively. Thus, we can conclude that the proposed MHCMM can significantly improve the CLTIR system.

D. Extension Experiments with Vast Corpus

In practice, it is difficult to collect and annotate a large number of images, but large corpora of texts are available. Therefore, we have also conducted experiments on limited triplet data but a large amount of text corpus. In this scenario, the translation model is trained using all 12M sentence pairs of the AIC dataset. Meanwhile, the student in the MHCMM framework is developed with 1M tripled data in BLATID and the supervision derived from the semantic features of the translation model. For comparison, two models in the cascaded system are trained independently. The cascaded system has a strong MT model, but its text recognizer is trained using the images and corresponding mono-lingual text labels in the BLATID dataset. In addition, as the single-task and multitask systems cannot be further extended with an extra text corpus, we do not compare these two methods in this scenario.

The results are listed in Table III with '*'. Obviously, with an extra bilingual text corpus, both the cascaded and MHCMM

TABLE III: Experimental results on the BLATID dataset for CLTIR. All results are measured by BLEU scores. '#Params' means the number of parameters during inference. Results in the top block are established on the BLATID dataset, while the below results are produced by leveraging a large amount of text corpus in the AIC dataset.

Systems	Validation	Test	#Params
Translation	23.14	22.61	63M
Cascaded	21.72	19.92	96M
Single-Task	18.66	18.35	82M
Multi-Task	21.35	20.48	84M
MHCMM	**22.74**	**20.81**	82M
Translation*	31.49	30.94	63M
Cascaded*	27.82	24.64	96M
MHCMM*	**28.93**	**26.68**	82M

methods achieve better performance. For the cascaded system, it is intuitively obvious that a stronger MT model improves the final results. Compared with experiments in the above block, the performance of the MHCMM method can be further improved from 22.74 and 20.81 to 28.93 and 26.58, respectively. This indicates that the student in our framework can learn semantic knowledge from the teacher effectively. Furthermore, the system can be developed with a small number of triplet data, alleviating the burden of data collection and annotation. On the other hand, the proposed method is still superior to the baseline cascaded system while taking fewer parameters. Therefore, we can conclude that the MHCMM mechanism is data and parameter efficient.

E. Effects of Training Option

In this paper, the student model is optimized through three different paths: global mimic, local mimic, and sequence classification, as shown in Fig. 2. To verify the effects of different training options, we establish a series of experiments. As both the length of text and image features are variable and probably different, only adversarial learning is adopted as a feasible option for the global mimic. For a local mimic, the length of the weighted sum of the feature sequence is fixed. Thus, both adversarial learning and MSE are appropriate for guiding the student. In addition, we also try to utilize the teacher's logits (outputs of the last fully connected layer), rather than target text labels, to provide overall supervision for students in the manner of knowledge distillation (KD) [30] as a sequence classification objective.

The results are shown in Table IV. The first row corresponds to the abovementioned single-task system, which is developed without the supervision of teachers and promotion from a vast corpus. Thus, its performance is inferior to other models. By combining the global and local mimics, the proposed MHCMM method significantly outperforms other single-hierarchical mimic methods and achieves the best performance, as listed in the last row. Comparing the second row with the last row, we can see that the MSE loss for the local mimic can improve the CLTIR model by over 3.1 BLEU. Similarly, the model with adversarial learning for the global mimic can improve the performance from 27.03 and 25.84 to 28.93 and 26.68, respectively.

TABLE IV: Experimental results on the BLATID dataset based on different training options. All results are measured by BLEU scores. 'Adv' stands for adversarial learning, 'CE' means the cross entropy criterion, and 'MSE' is the mean square error criterion.

Global mimic	Local mimic	Sequence classification	Validation	Test
×	×	CE	18.66	18.35
Adv	×	CE	23.91	23.56
×	Adv	CE	24.51	21.70
×	MSE	CE	27.03	25.84
Adv	MSE	KD	23.63	23.47
Adv	MSE	CE	**28.93**	**26.68**

In addition, the results in the 2nd to 4th rows in Table IV reveal the effects of the two mimic algorithms. Obviously, the local mimic, whether with adversarial learning or MSE, is superior to the global mimic. This indicates that local knowledge is more significant than global knowledge in an attention-based sequence learning framework, as the essence of the task is elementwise classification.

For the local mimic, the MSE criterion dramatically outperforms the adversarial learning scheme by 4.14 BLEU on the test set. The underlying reason is that adversarial learning is concerned with the overall distribution, while MSE is sensitive to details. Thus, the MSE is more appropriate for local mimics in our method.

Furthermore, we also try to introduce the knowledge distillation method for sequence classification. However, its result is inferior at best. We conjecture that obtaining knowledge only from teachers is insufficient and real text labels are still indispensable.

F. Effects of Hyper-Parameters

To evaluate the role of loss weight λ_* in Eq. (14), we conduct experiments with different loss weights. The results are shown in Table V. It can be seen that for all λ_* from 0.1 to 10 times, there is only approximately $\pm$ 0.7 floating for BLEUs. This indicates that the accuracy is not sensitive to loss weights. Therefore, for generalization, we roughly set all loss weights to 1.0.

G. Extension to Transformer

In addition to the ConvS2S-based MHCMM, we also try to utilize the transformer architecture for the encoder and

TABLE V: Comparison results on the BLATID dataset for different loss weights.

λ_{glb}	λ_{loc}	λ_{seq}	Validation	Test
0.1	1.0	1.0	28.59	26.67
0.5	1.0	1.0	28.53	25.87
2	1.0	1.0	28.65	26.20
10	1.0	1.0	28.41	26.65
1.0	0.1	1.0	28.56	26.04
1.0	0.5	1.0	28.85	27.01
1.0	2	1.0	29.02	26.72
1.0	10	1.0	28.61	25.90
1.0	1.0	0.1	28.49	26.37
1.0	1.0	0.5	28.47	26.46
1.0	1.0	2	28.29	26.30
1.0	1.0	10	28.23	26.31

TABLE VI: Experiments of Transformer based systems on the BLATID dataset fro CLTIR.

System	Validation	Test
Translation	32.11	31.50
Cascade	28.35	25.95
MHCMM	**29.21**	**27.54**

decoder. Under the same pipeline, experiments on transformer-based MHCMM and homogeneous systems are established.

As shown in Table VI, the BLEU of the cascaded system is far below the translation model there by verifying the error accumulation problem. Meanwhile, the MHCMM still outperforms the cascaded system on both the validation and test sets. Thus, we can conclude that the MHCMM framework is robust for different attention schemes.

H. Visualization

In addition to the above quantitative results, we also show some qualitative examples. In the MHCMM method, the attention score is adopted to weight the sequence from the encoder. With this score, the sequences from the encoder and decoder can be aligned. Thus, it can indicate the essence of the CLTIR model. To study the effect of MHCMM, we visualize the attention scores α_{ij} in Eq. (1). As shown in Fig. 5, the two heatmaps in each subfigure represent the attention scores derived from the student (top) and teacher (bottom). The magnitude of each pixel indicates the value of score α_{ij} associated with the i-th element in the source sequence (x-axis) and the j-th element in the target sequence (y-axis). In each heatmap, brighter blocks indicate higher correspondence between the source and target sequences.

In Fig. 5(a), the length of the text is short, and the correspondence of the source and target sentences is ordinal. Thus, the corresponding blocks around the diagonal are bright, while the others are dim. This indicates that both the teacher and the student model can give reasonable attention scores. For Fig. 5(b) and Fig. 5(c), the alignment of sequences is complicated, yet the student model can still produce attention scores similar to those of the teacher model. For instance, in Fig. 5(c), the phrase 'get there soon' in the target sentence is aligned to a subsequence of reverse order in the source sentence in both heatmaps. In this case, the attention scores from both the teacher and the student have a similar distribution. From these results, we can conclude that the CLTIR model in the MHCMM framework can process the text images in the same manner as the conventional machine translation model.

VI. Conclusion and Future Work

In this paper, we consider a new problem, namely, cross-lingual text image recognition (CLTIR). To solve this problem, we propose a multihierarchy cross-modal mimic framework, where a machine translation model is used as a teacher to guide the student in the semantic feature space. Based on adversarial learning and attention mechanisms, we leverage both global and local knowledge to further improve the end-to-end CLTIR model. In addition, the framework can take advantage of a vast bilingual corpus to further improve performance. Comprehensive experiments conducted on a newly collected dataset demonstrate the superiority of the proposed method and justify the benefits of semantic feature guidance and the promise of attention alignment.

For further improvement and to target real-world applications, the proposed work can be extended in several ways in the future. First, text image data with bilingual annotation (called triplet data) are scarce. Although a large corpus of bilingual texts is shown to boost the performance in this paper, a method for training with a small amount of triplet data and plenty of image-text pair data is desired. Second, in the cross-modal mimic framework, better network architectures for the encoder and decoder can be exploited to yield better performance. Third, for daily life applications, the network models need to be compressed and accelerated for implementation on low-power computers.

References

[1] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 855–868, 2009.

[2] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, 2017.

[3] F. Yin, Y.-C. Wu, X.-Y. Zhang, and C.-L. Liu, "Scene text recognition with sliding convolutional character models," *arXiv preprint arXiv:1709.01727*, 2017.

[4] Z. Chen, F. Yin, X.-Y. Zhang, Q. Yang, and C.-L. Liu, "Multrenets: Multilingual text recognition networks for simultaneous script identification and handwriting recognition," *Pattern Recognition*, vol. 108, p. 107555, 2020.

[5] Z. Wan, J. Zhang, L. Zhang, J. Luo, and C. Yao, "On vocabulary reliance in scene text recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 425–11 434.

[6] Q. Lin, C. Luo, L. Jin, and S. Lai, "Stan: A sequential transformation attention-based network for scene text recognition," *Pattern Recognition*, vol. 111, p. 107692, 2021.

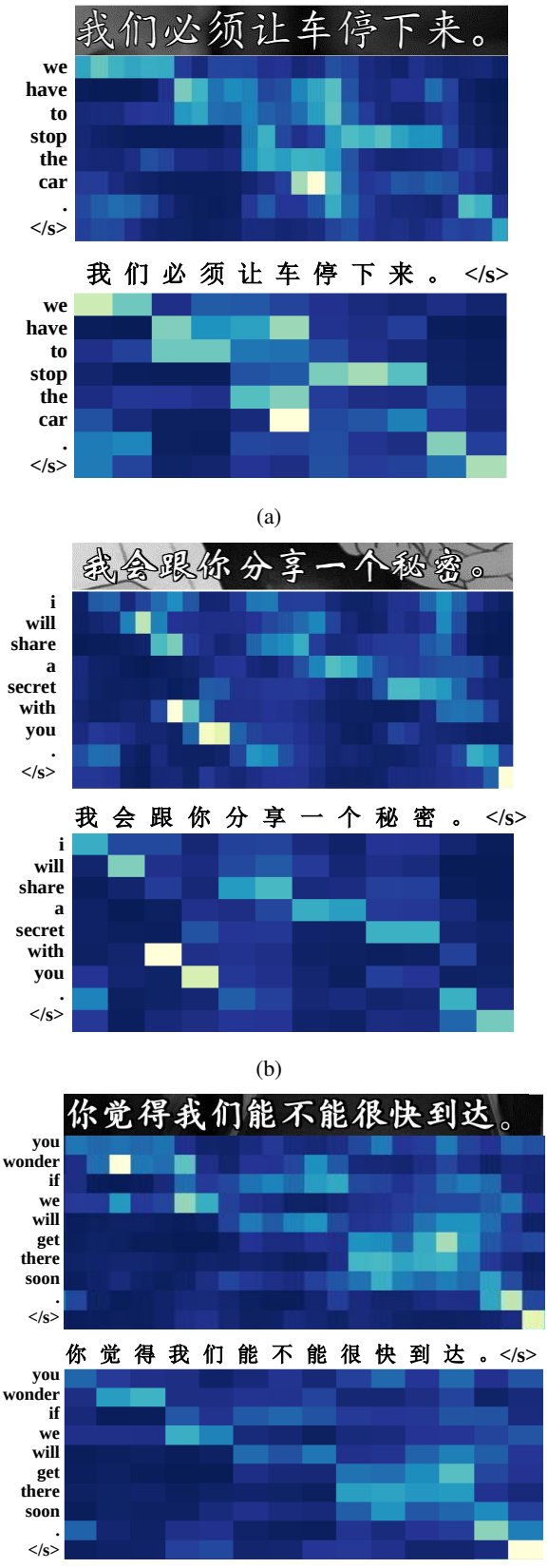

Fig. 5: Attention scores produced by the student (top) and teacher (bottom) in the proposed MHCMM framework.

[7] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proceedings of the International Conference on Learning Representations*, 2017, pp. 1243–1252.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[10] R. Dabre, C. Chu, and A. Kunchukuttan, "A survey of multilingual neural machine translation," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–38, 2020.

[11] F. Stahlberg, "Neural machine translation: A review," *Journal of Artificial Intelligence Research*, vol. 69, pp. 343–418, 2020.

[12] S. Yang, Y. Wang, and X. Chu, "A survey of deep learning techniques for neural machine translation," *arXiv preprint arXiv:2002.07526*, 2020.

[13] R. Dabre, C. Chu, and A. Kunchukuttan, "A comprehensive survey of multilingual neural machine translation," *arXiv preprint arXiv:2001.01115*, 2020.

[14] Z. Chen, F. Yin, X.-Y. Zhang, Q. Yang, and C.-L. Liu, "Cross-lingual text image recognition via multi-task sequence to sequence learning," in *Proceedings of the International Conference on Pattern Recognition*, 2020, pp. 3122–3129.

[15] Z. Chen, Y. Wu, F. Yin, and C.-L. Liu, "Simultaneous script identification and handwriting recognition via multi-task learning of recurrent neural networks," in *Proceedings of the IAPR International Conference on Document Analysis and Recognition*, vol. 1, 2017, pp. 525–530.

[16] Y.-C. Wu, F. Yin, Z. Chen, and C.-L. Liu, "Handwritten chinese text recognition using separable multi-dimensional recurrent neural network," in *Proceedings of the IAPR International Conference on Document Analysis and Recognition*, 2017, pp. 79–84.

[17] S. Long, X. He, and C. Yao, "Scene text detection and recognition: The deep learning era," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 161–184, 2021.

[18] S. Karaoglu, R. Tao, T. Gevers, and A. W. Smeulders, "Words matter: Scene text for image classification and retrieval," *IEEE Transactions on Multimedia*, vol. 19, no. 5, pp. 1063–1076, 2016.

[19] C. Yan, H. Xie, J. Chen, Z. Zha, X. Hao, Y. Zhang, and Q. Dai, "A fast uyghur text detector for complex background images," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3389–3398, 2018.

[20] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3111–3122, 2018.

[21] X. Ren, Y. Zhou, J. He, K. Chen, X. Yang, and J. Sun, "A convolutional neural network-based chinese text detection algorithm via text structure modeling," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 506–518, 2016.

[22] L. Wu, P. Shivakumara, T. Lu, and C. L. Tan, "A new technique for multi-oriented scene text line detection and tracking in video," *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1137–1152, 2015.

[23] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "Aster: An attentional scene text recognizer with flexible rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2035–2048, 2018.

[24] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1700–1709.

[25] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[26] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the International Conference on Learning Representations*, 2015.

[27] Z. Lin, X. Pan, M. Wang, X. Qiu, J. Feng, H. Zhou, and L. Li, "Pre-training multilingual neural machine translation by leveraging alignment information," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 2649–2663.

[28] D. Wang, C. Gong, and Q. Liu, "Improving neural language modeling via adversarial training," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2019, pp. 6555–6565.

[29] S. Mehta, M. Ghazvininejad, S. Iyer, L. Zettlemoyer, and H. Hajishirzi, "Delight: Very deep and light-weight transformer," in *Proceedings of the International Conference on Learning Representations*, 2021, pp. 1243–1252.

[30] D. J. Hinton Geoffrey, Vinyals Oriol, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
[31] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *Advances in Neural Information Processing Systems*, 2014, pp. 2654–2662.
[32] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2827–2836.
[33] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *Proceedings of the ACM International Conference on Multimedia*, 2018, pp. 292–301.
[34] F. M. Thoker and J. Gall, "Cross-modal knowledge distillation for action recognition," in *Proceedings of the International Conference on Image Processing*, 2019, pp. 6–10.
[35] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang, "Crdoco: Pixel-level domain transfer with cross-domain consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1791–1800.
[36] J. N. Kundu, N. Lakkakula, and R. V. Babu, "Um-adapt: Unsupervised multi-task adaptation using adversarial cross-task distillation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1436–1445.
[37] L. Zhao, X. Peng, Y. Chen, M. Kapadia, and D. N. Metaxas, "Knowledge as priors: Cross-modal knowledge generalization for datasets without superior knowledge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6528–6537.
[38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.
[39] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.
[40] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.
[41] F. Huang, X. Zhang, and Z. Li, "Learning joint multimodal representation with adversarial attention networks," in *Proceedings of the 26th ACM International Conference on Multimedia*, 2018, pp. 1874–1882.
[42] J.-W. Wu, F. Yin, Y.-M. Zhang, X.-Y. Zhang, and C.-L. Liu, "Handwritten mathematical expression recognition via paired adversarial learning," *International Journal of Computer Vision*, pp. 1–16, 2020.
[43] S. Wang, S. Yin, L. Hao, and G. Liang, "Multi-task face analyses through adversarial learning," *Pattern Recognition*, vol. 114, p. 107837, 2021.
[44] R. Li, L.-F. Cheong, and R. T. Tan, "Heavy rain image restoration: Integrating physics model and conditional adversarial learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
[45] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517–2526.
[46] S. Lin, R. Ji, C. Yan, B. Zhang, L. Cao, Q. Ye, F. Huang, and D. Doermann, "Towards optimal structured cnn pruning via generative adversarial learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2790–2799.
[47] Y. Zhang, S. Liang, S. Nie, W. Liu, and S. Peng, "Robust offline handwritten character recognition through exploring writer-independent features under the guidance of printed data," *Pattern Recognition Letters*, vol. 106, pp. 20 – 26, 2018.
[48] A. K. Bhunia, A. Das, A. K. Bhunia, P. S. R. Kishore, and P. P. Roy, "Handwriting recognition in low-resource scripts using adversarial learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4767–4776.
[49] M. Zhu, P. Pan, W. Chen, and Y. Yang, "Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[51] AIChallenger, "Ai challenger," https://github.com/AIChallenger/AI_Challenger_2018, 2018.
[52] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," *arXiv preprint arXiv:1406.2227*, 2014.
[53] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.
[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[55] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the Annual Meeting on Association for Computational Linguistics*, 2002, pp. 311–318.

Zhuo Chen received the B.S. degree in electronic information engineering from China Agricultural University, Beijing, China, in 2015. He pursued a Ph.D. degree in Pattern Recognition and Intelligent Systems at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation of Chinese Academy of Sciences, Beijing, China, during 2015-2021. His research interests include multilingual text recognition, multitask learning, and sequence pattern recognition.

Fei Yin is an associate professor at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation of Chinese Academy of Sciences, Beijing, China. He received a Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation of Chinese Academy of Sciences in 2010. He received his BS and MS from 'Xian University of Posts and Telecommunications in 1999 and Huazhong University of Science and Technology in 2002, respectively. His research interests include pattern recognition, document analysis and recognition. He has published more than 100 papers in international journals and conferences.

Qing Yang is a professor at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation of Chinese Academy of Sciences, Beijing, China. He received a Ph.D. degree in computer science from the Institute of Automation of Chinese Academy of Sciences, Beijing. His research interests include image processing, pattern recognition, and computer vision.

Cheng-Lin Liu (Fellow, IEEE) received a B.S. degree in electronic engineering from Wuhan University, Wuhan, China, an M.E. degree in electronic engineering from Beijing University of Technology, Beijing, China, and a Ph.D. degree in pattern recognition and intelligent control from the Institute of Automation of Chinese Academy of Sciences, Beijing, China, in 1989, 1992 and 1995, respectively. He was a postdoctoral fellow at the Korea Advanced Institute of Science and Technology (KAIST) and later at Tokyo University of Agriculture and Technology from March 1996 to March 1999. From 1999 to 2004, he was a research staff member and later a senior researcher at the Central Research Laboratory, Hitachi, Ltd., Tokyo, Japan. Since 2005, he has been a Professor at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation of Chinese Academy of Sciences, Beijing, China, and is now the director of the laboratory. His research interests include pattern recognition, machine learning, and the applications to character recognition and document analysis. He has published over 300 technical papers in prestigious international journals and conferences. He is an associate Editor-in-Chief of Pattern Recognition Journal and Acta Automatica Sinica and is on the editorial board of several international and domestic journals. He is a Fellow of the IEEE, the IAPR, the CAA and CAAI.