CCIM: Cross-modal Cross-lingual Interactive Image Translation

Cong Ma[1,2], **Yaping Zhang**[1,2]*, **Mei Tu**[4], **Yang Zhao**[1,2], **Yu Zhou**[2,3], **Chengqing Zong**[1,2]
[1]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[2]State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS),
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]Fanyu AI Laboratory, Zhongke Fanyu Technology Co., Ltd, Beijing, China
[4]Samsung Research China - Beijing (SRC-B)
{cong.ma, yaping.zhang, yang.zhao, yzhou, cqzong}@nlpr.ia.ac.cn
mei.tu@samsung.com

Abstract

Text image machine translation (TIMT) which translates source language text images into target language texts has attracted intensive attention in recent years. Although the end-to-end TIMT model directly generates target translation from encoded text image features with an efficient architecture, it lacks the recognized source language information resulting in a decrease in translation performance. In this paper, we propose a novel Cross-modal Cross-lingual Interactive Model (CCIM) to incorporate source language information by synchronously generating source language and target language results through an interactive attention mechanism between two language decoders. Extensive experimental results have shown the interactive decoder significantly outperforms end-to-end TIMT models and has faster decoding speed with smaller model size than cascade models. [1]

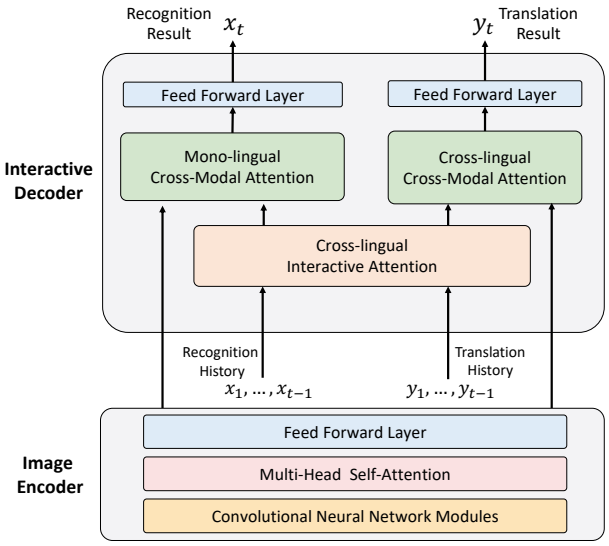
1 Introduction

Text image machine translation (TIMT) aims at translating text in images from the source language into the target language, which has been widely used in various applications such as photo translation, scene text translation, digital document translation, and so on. Existing research on TIMT is mainly divided into two categories of methods: cascade method and end-to-end method. Cascade method (Hinami et al., 2021; Shekar et al., 2021; Afli and Way, 2016; Chen et al., 2015; Du et al., 2011) takes a text image recognition (TIR) model for source language text recognition (Baek et al.; Shi et al., 2017, 2016; Zhang et al., 2021, 2019) and then translates them into target language texts with a machine translation (MT) model (Vaswani et al., 2017; Gehring et al., 2017a,b; Johnson et al., 2017; Bahdanau et al., 2015; Sutskever et al., 2014; Zhao et al., 2019, 2020). To explicitly recognize the source language embedded in text images, the cascade model combines TIR models and MT models for the TIMT task. However, two individual models in the cascade frame have double parameters and the decoding speed is slow. Meanwhile, errors in the TIR model are further propagated in the MT model leading to performance decrease. The end-to-end method directly translates the source language text image into target language through a unified encoder-decoder architecture, which is more parameter-efficient than cascade models with faster decoding speed (Ma et al., 2022; Su et al., 2021; Mansimov et al., 2020; Chen et al., 2020; Ma et al., 2023a,b,c).

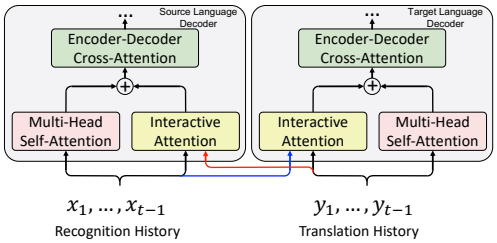
However, the performance of end-to-end models is limited because the translation process lacks explicit source language guidance from recognition texts. An intuitive solution is to incorporate the recognition history into the translation decoder to offer more efficient guidance. Recently, multi-source interaction has been studied to incorporate effective information into target model (Lu et al., 2022; Xu et al., 2021; He et al., 2021; Liu et al., 2020; Zhou et al., 2019a,b; Wang et al., 2019; Zoph and Knight, 2016). Although multi-source interaction is vital to enhance the encoding capacity of TIMT model through attending recognition information explicitly, it has not been explored yet.

To address the above issues, we propose a novel Cross-modal Cross-lingual Interactive Model (CCIM) for TIMT, which effectively incorporates source language recognition information into the TIMT decoder through interactive attention. The interactive decoder has two decoding modules, one for source language and the other one for target language generation. A cross-lingual interactive attention mechanism is introduced to bridge the two language decoders. When generating translation results, the target language decoder not only receives the hidden states from the encoder and previous decoded translation history but also attends to the

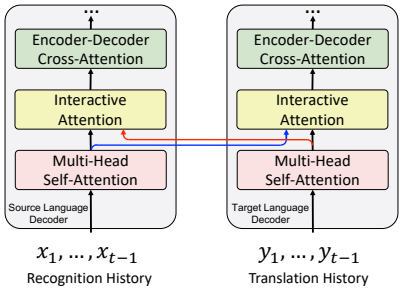
*Corresponding author.
[1]https://github.com/EriCongMa/CCIM



(a) Architecture of Cross-modal Cross-lingual Interactive Model for Text Image Translation



(b) Weighted Cross-lingual Interactive Attention



(c) Hierarchical Cross-lingual Interactive Attention

Figure 1: (a) illustrates the proposed cross-modal cross-lingual interactive model (CCIM). (b) illustrates the weighted cross-lingual interactive attention module. (c) illustrates the hierarchical cross-lingual interactive attention module.

decoded recognition history. Our contributions are summarized as follows:

- We propose a novel cross-modal cross-lingual interactive model (CCIM) for the TIMT task, which effectively enhances the translation decoder by incorporating recognition features.
- Weighted and hierarchical interactive decoding strategies have been studied to validate the effectiveness of interactive generation.
- Experimental results on three evaluation datasets have revealed the CCIM improves the translation quality of end-to-end TIMT models and outperforms cascade models with fewer parameters and faster decoding speed.

2 Methodology

2.1 Cross-modal Cross-lingual Interactive Model

The proposed CCIM model consists of an image encoder and an interactive decoder. As shown in Figure 1 (a), the image encoder first extracts image features given the source language text image, then two decoders are utilized for text image recognition and translation synchronously.

For image encoding, a convolutional neural network is utilized to extract image representation through multi-layer convolution and pooling operations (He et al., 2016). While for multi-head attention (MHA), the model collects information from different positions to update the hidden state of the current position (Vaswani et al., 2017):

$$\begin{aligned} &\mathrm{MHA}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1,...,\mathrm{head}_\mathrm{h})W_O \\ &\text{where } \mathrm{head}_\mathrm{i} = \mathrm{Attention}(QW_Q^i, KW_K^i, VW_V^i) \end{aligned} \tag{1}$$

where W_Q^i, W_K^i and W_V^i represent query, key, and value projection matrices for head i, respectively. W_O denotes the output projection matrix.

Self-attention (SA) The interactive decoder first calculates self-attention hidden states for both source language X and target language Y given the same query, key, and value:

$$\begin{aligned} H_X^{\mathrm{SA}} &= \mathrm{MHA}(X,X,X) \\ H_Y^{\mathrm{SA}} &= \mathrm{MHA}(Y,Y,Y) \end{aligned} \tag{2}$$

Then, cross-lingual interactive-attention (IA) hidden states are calculated through two language decoders. Two types of IA mechanisms are utilized to recognize and translate synchronously:

Weighted Interactive Attention (WIA) As shown in Figure 1 (b), the self-attention and interactive attention are calculated separately and then weighted summation:

$$\begin{aligned} H_X^{\mathrm{WIA}} &= H_X^{\mathrm{SA}} + \lambda \times \mathrm{MHA}(X,Y,Y) \\ H_Y^{\mathrm{WIA}} &= H_Y^{\mathrm{SA}} + \lambda \times \mathrm{MHA}(Y,X,X) \end{aligned} \tag{3}$$

where the query of WIA is from the corresponding language decoding history. Key and value are from the other language history.

Hierarchical Interactive Attention (HIA) To fuse self- and interactive-attention together, a hierarchical calculation mechanism is introduced to

	Synthetic TIMT			Subtitle TIMT	Street-view TIMT	MT Dataset	TIR Dataset
	#Train	#Valid	#Test	#Test	#Test	#Train	#Train
Zh⇒En	1,000,000	2,000	2,502	1,040	1,198	5,984,287	1,000,000
En⇒Zh	1,000,000	2,000	2,502	1,040	-	5,984,287	1,000,000
En⇒De	1,000,000	2,000	2,000	-	-	20,895,771	1,000,000

Table 1: Statistics of text image machine translation (TIMT), machine translation (MT), and text line image recognition (TIR) datasets.

obtain interactive information through serial computing as shown in Figure 1 (c):

$$\begin{aligned} H_X^{\text{HIA}} &= \text{MHA}(H_X^{\text{SA}}, H_Y^{\text{SA}}, H_Y^{\text{SA}}) \\ H_Y^{\text{HIA}} &= \text{MHA}(H_Y^{\text{SA}}, H_X^{\text{SA}}, H_X^{\text{SA}}) \end{aligned} \tag{4}$$

where the query is from the inner-lingual self-attention results, while the key and value are from the other language self-attention features.

Encoder-Decoder Cross-Attention (CA) Cross-lingual interactive-attention hidden states are fed into the encoder-decoder cross-attention mechanism to further incorporate encoder features into the decoder as in (Vaswani et al., 2017).

$$\begin{aligned} H_X^{\text{CA}} &= \text{MHA}(H_X^{\text{IA}}, H_I, H_I) \\ H_Y^{\text{CA}} &= \text{MHA}(H_Y^{\text{IA}}, H_I, H_I) \end{aligned} \tag{5}$$

where H_I represents the hidden states from the image encoder. $H_X^{\text{IA}}, H_Y^{\text{IA}}$ can be WIA or HIA for source and target languages. $H_X^{\text{CA}}, H_Y^{\text{CA}}$ denote the output of the encoder-decoder cross-attention module for source and target language, respectively.

The hidden states from the encoder-decoder cross-attention mechanism are then further encoded by the feedforward layer to obtain the interactive decoder layer outputs. Notes that the residual connections (He et al., 2016) and layer normalization (Ba et al., 2016) in standard transformer decoder are also utilized after self-attention, interactive attention, encoder-decoder cross attention and feedforward modules in interactive decoder (Zhao et al., 2023), which are not drawn in Figure 1 for simplification.

2.2 Loss Functions for Optimization

Since the interactive decoder has two decoders for the source language and target language respectively, TIR and TIMT tasks are optimized synchronously by multi-task learning. The training dataset contains triple paired samples as $D = \{I^i, X^i, Y^i\}_i^{|D|}$, where I^i is the i-th source language text image, X^i is the i-th source language texts and Y^i is the corresponding translated target language texts. The model is updated by optimizing both TIR and TIMT loss functions:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{TIR}} + \mathcal{L}_{\text{TIMT}} \\ \mathcal{L}_{\text{TIR}} &= -\sum_i^{|D|} \sum_j^M \log P(x_j^i | I^i, x_{<j}^i, y_{<j}^i) \\ \mathcal{L}_{\text{TIMT}} &= -\sum_i^{|D|} \sum_j^N \log P(y_j^i | I^i, x_{<j}^i, y_{<j}^i) \end{aligned} \tag{6}$$

where $x_{<j}$ and $y_{<j}$ denote the recognition and translation history. M and N represent the token length of the source language and target language. Note that the interactive decoder has three attention modules (self-attention, interactive attention from the other task, and encoder-decoder attention), the decoder generates tokens given the condition of both text image, recognition history, and translation history. Thus the interactive decoder has the potential to generate more accurate translation results.

2.3 Training and Inference

For each decoding step during training, the teacher-forcing decoding strategy is utilized to train the parameters in the decoder in a parallel computing way and each position in the decoder can attend to all positions in the decoder up to and including that position through the attention mask. During inference, the decoder generates both source and target language tokens by tokens in an auto-regressive way. For each step, the two sub-branches of the interactive decoder can attend encoder features, recognition history features, and translation history features, and predict both source and target language at the current step.

3 Experiments and Results

3.1 Datasets

The experiments have been conducted on a public TIMT corpus released by (Ma et al., 2022). The training set contains one million triple-paired samples of source language images, source language

Architecture	Synthetic			Subtitle		Street
	En⇒Zh	En⇒De	Zh⇒En	En⇒Zh	Zh⇒En	Zh⇒En
CLTIR (Chen et al., 2020)	18.02	15.55	10.74	16.47	9.04	0.43
+TIR	19.44	16.31	13.52	17.96	11.25	1.74
RTNet (Su et al., 2021)	18.91	15.82	12.54	17.63	10.63	1.07
+TIR	19.63	16.78	14.01	18.82	11.50	1.93
MTETIMT(Ma et al., 2022)	19.25	16.27	13.16	17.73	10.79	1.69
+MT	21.96	18.84	15.62	19.17	12.11	5.84
CCIM	**22.21**	**19.13**	**15.72**	**19.48**	**12.12**	**5.88**

Table 2: Performance of end-to-end models. All end-to-end models are trained with the same TIMT training dataset. External TIR and MT corpus are also kept the same among different architecture settings.

texts, and target language translation pairs for each translation direction. The source language text images in the training dataset are synthesized by using bilingual text sentences. To validate the generalization of models, one synthetic test set and two real-word (subtitle and street-view) test sets are utilized to evaluate the translation performance. The statistics of the dataset are shown in Table 1.

3.2 Experimental Settings

Image encoder in CCIM utilizes the same configuration in (Ma et al., 2022). The source language and target language decoder are 6-layer transformer decoder with 512-dimensional hidden sizes as in (Vaswani et al., 2017; Zhao et al., 2023). The maximum sentence length for English, German, and Chinese are set to 80, 80, and 40 respectively. The preprocessed image height is set to 32 and the channel is 3. To align the length of the image feature and text feature, preprocessed image width is resized to 320, 320, and 160. The batch size is set to 64, and the training step is 300,000. All models are initialized with Xavier initiation method (Glorot and Bengio, 2010) and optimized with Adam optimizer (Kingma and Ba, 2015) on a single NVIDIA V100 GPU. Sacre-BLEU[2] (Papineni et al., 2002) is utilized for evaluation metric.

3.3 Baseline Models

- CLTIR model is a vanilla multi-task learning based TIMT model with auxiliary TIR task training (Chen et al., 2020).
- RTNet bridges the TIR encoder and MT decoder through a feature transformer, which is also trained with TIR task (Su et al., 2021).
- MTETIMT is a machine translation enhanced TIMT model, which is trained with both auxiliary TIR and MT tasks (Ma et al., 2022).

[2]https://github.com/mjpost/sacrebleu

Architecture	BLEU↑	Param.↓	Speed↑
Cascade	20.46	195M	3.07
Our work: CCIM	22.21	147M	5.04

Table 3: Comparison of cascade and end-to-end CCIM models. The unit for parameters is million ($\times 10^6$), while the unit for speed is sentence per second.

3.4 Comparison with Different End-to-End TIMT Models

Table 2 shows the main results on three evaluation domains. As shown in Table 1, CCIM outperforms the existing best multi-task based MTETIT by 0.21 BLEU scores on average. Meanwhile, CCIM improves the translation performance on real-world domains by 0.12 BLEU scores on average, indicating the good generalization of our proposed method. Furthermore, CCIM can generate source language and target language synchronously, which can meet the requirement of both recognition and translation tasks in practical applications.

3.5 Model Size and Decoding Speed

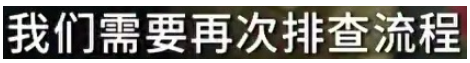
The Cascade model deploys TIR and MT models, leading to parameter redundancy and decoding delay. With an end-to-end architecture, CCIM outperforms the cascade model with fewer parameters and faster decoding speed as shown in Table 3. Specifically, CCIM decreases around 24.62% parameters and achieves 1.64x acceleration compared with the cascade model. Meanwhile, CCIM significantly outperforms the cascade model by 1.75 BLEU scores, which effectively alleviates the error propagation problem in the cascade model.

3.6 Comparison of Different Interactive Attention Types

To validate the effectiveness of interactive attention, an ablation study of replacing key and value

Interactive Attention Type	BLEU
Weighted Interactive Attention (Rand)	8.07
Hierarchical Interactive Attention (Rand)	11.23
Weighted Interactive Attention (WIA)	22.94
Hierarchical Interactive Attention (HIA)	24.18

Table 4: Comparison of Various Interactive Attention Types on English-to-Chinese validation set.



Recognition Ground Truth (Pinyin)	我们 需要 再次 排查 流程 (women xuyao zaici paicha liucheng)
End-to-end TIT	We need to check the process
Multi-task	We need to check the process
CCIM	We need to double-check the process
Translation Ground Truth	We need to double check the process

Figure 2: Case study of end-to-end TIMT models.

in interactive attention with random samples noise vector has been implemented. Experimental results in Table 4 show that random noise replaced interactive attention generates a poor translation, especially for weighted interactive attention. We attribute that weighted interactive attention incorporates noise signals through weighted summation, which severely disturbs the information flow. Furthermore, hierarchical interactive attention outperforms weighted interactive attention, which reveals that flexible calculation of hierarchical architecture is better than vanilla summation operation.

3.7 Case Study of CCIM Model

Fig. 2 shows an example of TIMT generated by end-to-end and CCIM models. Although the end-to-end model translates the general meaning of the sentence, it ignores the meaning of 'double-check' in the source language text image. Since there is no interaction during decoding, the multi-task based model also ignores this meaning. CCIM successfully translated this word through interactive attention with the source language decoder, indicating CCIM can effectively alleviate the problem of lacking source language information in vanilla end-to-end TIMT models.

3.8 Wait-k Strategy for CCIM

The wait-k strategy is commonly employed in speech translation, aiming at generating better translation given more recognition history. To validate the wait-k strategy in the TIMT task, we also conducted corresponding experiments as shown in Table 5. From the experimental result, the wait-k strategy makes the recognition task decode first, enabling the translation task to access more source language information for improved translation quality. While the wait-k strategy enhances translation quality, it does introduce some latency increase. The CCIM model achieves the best translation performance when $k = 4$ in our experiments.

Wait-k	WIA	HIA
Wait-0	22.94	24.18
Wait-1	22.99	24.32
Wait-2	23.07	24.75
Wait-3	23.64	24.91
Wait-4	**23.75**	**25.49**
Wait-5	23.36	25.18

Table 5: The performance of wait-k strategy for CCIM on English-to-Chinese Validation Set.

4 Conclusion

This paper proposes a novel interactive decoder based end-to-end TIMT model, which explicitly incorporates recognized hidden states into the translation process. Through the interactive attention mechanism, recognition and translation results are generated synchronously and mutually enhanced. By making full use of the source language recognition information, CCIM outperforms existing end-to-end and multi-task based TIMT models on both synthetic and real-world evaluation sets. Furthermore, with the end-to-end architecture, CCIM has fewer parameters and faster decoding speed than cascade models. Ablation study of different interactive attention types shows hierarchical interactive attention has stronger interactive ability across recognition and translation tasks. In the future, we will explore more interactive methods for end-to-end text image machine translation.

5 Acknowledgements

This work has been supported by the National Natural Science Foundation of China (NSFC) grants 62106265.

6 Limitations

Our method is now designed for text line images, which need preprocessing of text detection in images. In the future, we will consider optimizing the text detection and translation in images jointly to increase the scalability of our work.

References

Haithem Afli and Andy Way. 2016. Integrating optical character recognition and machine translation of historical documents. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities, LT4DH@COLING*, pages 109–116.

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4714–4722.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jinying Chen, Huaigu Cao, and Premkumar Natarajan. 2015. Integrating natural language processing with image document analysis: what we learned from two real-world applications. *Int. J. Document Anal. Recognit.*, 18(3):235–247.

Zhuo Chen, Fei Yin, Xu-Yao Zhang, Qing Yang, and Cheng-Lin Liu. 2020. Cross-lingual text image recognition via multi-task sequence to sequence learning. In *25th International Conference on Pattern Recognition (ICPR)*, pages 3122–3129.

Jun Du, Qiang Huo, Lei Sun, and Jian Sun. 2011. Snap and translate using windows phone. In *2011 International Conference on Document Analysis and Recognition (ICDAR)*, pages 809–813. IEEE Computer Society.

Jonas Gehring, Michael Auli, David Grangier, and Yann N. Dauphin. 2017a. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 123–135.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017b. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, pages 249–256.

Hao He, Qian Wang, Zhipeng Yu, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2021. Synchronous interactive decoding for multilingual neural machine translation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, pages 12981–12988.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Ryota Hinami, Shonosuke Ishiwatari, Kazuhiko Yasuda, and Yusuke Matsui. 2021. Towards fully automated manga translation. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc. Comput. Linguistics*, 5:339–351.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*.

Yuchen Liu, Jiajun Zhang, Hao Xiong, Long Zhou, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2020. Synchronous speech recognition and speech-to-text translation with interactive decoding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, pages 8417–8424.

Ziyao Lu, Xiang Li, Yang Liu, Chulun Zhou, Jianwei Cui, Bin Wang, Min Zhang, and Jinsong Su. 2022. Exploring multi-stage information interactions for multi-source neural machine translation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30:562–570.

Cong Ma, Xu Han, Linghui Wu, Yaping Zhang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023a. Modal contrastive learning based end-to-end text image machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–13.

Cong Ma, Yaping Zhang, Mei Tu, Xu Han, Linghui Wu, Yang Zhao, and Yu Zhou. 2022. Improving end-to-end text image translation from the auxiliary text translation task. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1664–1670.

Cong Ma, Yaping Zhang, Mei Tu, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023b. E2timt: Efficient and effective modal adapter for text image machine translation. In *Document Analysis and Recognition - ICDAR 2023*, pages 70–88, Cham. Springer Nature Switzerland.

Cong Ma, Yaping Zhang, Mei Tu, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023c. Multi-teacher knowledge distillation for end-to-end text image machine translation. In *Document Analysis and Recognition - ICDAR 2023*, pages 484–501, Cham. Springer Nature Switzerland.

Elman Mansimov, Mitchell Stern, Mia Chen, Orhan Firat, Jakob Uszkoreit, and Puneet Jain. 2020. Towards end-to-end in-image neural machine translation. In *Proceedings of the First International Workshop on Natural Language Processing Beyond Text*, pages 70–74, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318.

K. Chandra Shekar, Marilyn Cross, and Vignesh Vasudevan. 2021. Optical character recognition and neural machine translation using deep learning techniques. *Innovations in Computer Science and Engineering*.

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

Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2016. Robust scene text recognition with automatic rectification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4168–4176.

Tonghua Su, Shuchen Liu, and Shengjie Zhou. 2021. Rtnet: An end-to-end method for handwritten text image translation. In *16th International Conference on Document Analysis and Recognition (ICDAR)*, pages 99–113.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Yining Wang, Jiajun Zhang, Long Zhou, Yuchen Liu, and Chengqing Zong. 2019. Synchronously generating two languages with interactive decoding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3348–3353.

Weijia Xu, Yuwei Yin, Shuming Ma, Dongdong Zhang, and Haoyang Huang. 2021. Improving multilingual neural machine translation with auxiliary source languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3029–3041. Association for Computational Linguistics.

Yaping Zhang, Shuai Nie, Shan Liang, and Wenju Liu. 2021. Robust text image recognition via adversarial sequence-to-sequence domain adaptation. *IEEE Trans. Image Process.*, 30:3922–3933.

Yaping Zhang, Shuai Nie, Wenju Liu, Xing Xu, Dongxiang Zhang, and Heng Tao Shen. 2019. Sequence-to-sequence domain adaptation network for robust text image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2740–2749. Computer Vision Foundation / IEEE.

Yang Zhao, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2020. Knowledge graphs enhanced neural machine translation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4039–4045. ijcai.org.

Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2023. Transformer: A general framework from machine translation to others. *Mach. Intell. Res.*, 20(4):514–538.

Yang Zhao, Jiajun Zhang, Chengqing Zong, Zhongjun He, and Hua Wu. 2019. Addressing the under-translation problem from the entropy perspective. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 451–458. AAAI Press.

Long Zhou, Jiajun Zhang, and Chengqing Zong. 2019a. Synchronous bidirectional neural machine translation. *Trans. Assoc. Comput. Linguistics*, pages 91–105.

Long Zhou, Jiajun Zhang, Chengqing Zong, and Heng Yu. 2019b. Sequence generation: From both sides to the middle. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5471–5477.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 30–34. The Association for Computational Linguistics.