Modal Contrastive Learning Based End-to-End Text Image Machine Translation

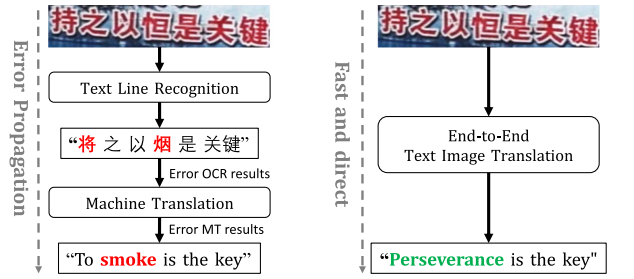
Cong Ma, Xu Han, Linghui Wu, Yaping Zhang, Yang Zhao, Yu Zhou, and Chengqing Zong, *Fellow, IEEE*

***Abstract*—Text image machine translation (TIMT) aims at directly translating text in the source language embedded in images into the target language. Most existing systems follow the cascaded pipeline diagram from recognition to translation, which suffers from the problem of error propagation, parameter redundancy, and information reduction. The end-to-end model has the potential to alleviate these issues via bridging the recognition and translation models. However, the challenge is the data limitation and modality gap between text and image. In this paper, we propose a novel end-to-end model, namely Modal contrastive learning based End-to-end Text Image Machine Translation (METIMT), which alleviates these issues through end-to-end text image machine translation architecture and modal contrastive learning. Specifically, an image encoder is designed to encode images into the same feature space of corresponding text sentences, with the guidance of an intra-modal and inter-modal contrastive learning module. To further promote the research of text image machine translation, we have constructed one synthetic and two real-world datasets. Extensive experiments show that our lighter, faster model outperforms not only existing pipeline methods but also state-of-the-art end-to-end models on both synthetic and real-world evaluation sets. Our code and dataset will be released to the public.**

***Index Terms*—Text image machine translation, contrastive learning, text image recognition, machine translation.**

I. Introduction

TEXT image machine translation (TIMT) is widely used in daily life like photo translation, which translates text embedded in the image into another language. The current solution for TIMT utilizes the pipeline diagram as shown in Fig. 1(a), which first recognizes the source text with text line recognition (TIR) module and then translates the text into target language with machine translation (MT) module [1], [2], [3], [4], [5]. However, the pipeline systems need to deploy two independent modules without any information sharing, causing error propagation, parameter redundancy, and information reduction. End-to-end TIMT model is potential to alleviate these issues as shown in Fig. 1(b).



(a) Pipeline Text Image Translation (b) End-to-End Text Image Translation

Fig. 1. Comparison of pipeline and end-to-end text image translation. End-to-end text image translation model outperforms pipeline diagram by alleviating error propagation with fewer parameters and faster decoding speed.

Existing end-to-end methods on TIMT explore various architectures [6] and multi-task learning with text line recognition task [7], [8]. However, end-to-end models still perform worse than pipeline models. We attribute the challenges of end-to-end TIMT research to two major aspects.

- *Modality gap:* TIMT is a cross-modal generation task, where the modality representation gap between text and image is huge. Existing end-to-end models encode image and text features independently without any information interaction. Recently, several multimodal learning tasks show promising improvements in both cross-modal understanding and generation tasks via reducing the modality gap [9], [10], [11]. However, there is rarely research on text image machine translation, where texts are embedded in images.
- *Data limitation:* As we know, end-to-end models usually rely on large-scale annotated datasets. However, it is an extremely high cost to collect large-scale data for end-to-end TIMT research and development. To the best of our knowledge, there is no public dataset designed for TIMT. Although there is some existing research exploring end-to-end TIMT with text line recognition datasets [6], [7], [8], they didn't incorporate the text parallel corpus to alleviate the problem of data limitation. Meanwhile, none of these data is publicly released, which limits the research and applications of TIMT.

Manuscript received 27 December 2021; revised 30 November 2022; accepted 30 September 2023. Date of publication 13 October 2023; date of current version 4 April 2024. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62106265. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yue Zhang. *(Corresponding author: Chengqing Zong.)*

The authors are with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with Institute of Automation Chinese Academy of Sciences, Beijing 100190, China (e-mail: cong.ma@nlpr.ia.ac.cn; xu.han@nlpr.ia.ac.cn; linghui.wu@nlpr.ia.ac.cn; yaping.zhang@nlpr.ia.ac.cn; yang.zhao@nlpr.ia.ac.cn; yzhou@nlpr.ia.ac.cn; cqzong@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TASLP.2023.3324540

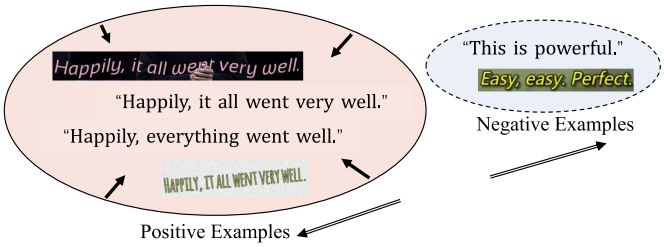


Fig. 2. Modal contrastive learning is utilized to make positive examples attracted, and negative examples separated in feature space.

To address the above challenges in TIMT, we propose a novel modal contrastive learning based end-to-end TIMT model, where a transformer based backbone for end-to-end TIMT is designed. Image and text features encoded by corresponding encoders are then aligned through a modal contrastive learning module, which attracts multimodal positive examples together, and separates negative examples as shown in Fig. 2. Specifically, for sentences with the same/similar semantics, the feature representation of these text sentences and text images should be similar. Meanwhile, images of the same sentence with different backgrounds, fonts, font colors, and blurring levels should also have similar hidden features due to the same text information in images. Sentences with similar semantics but different expressions should also have similar hidden features. Sentences and images with different semantics should have different feature representations to obtain the corresponding translation results. Furthermore, to alleviate the problem of data scarcity, a text image synthesis method is utilized to synthesize text images based on the large-scale bilingual text parallel corpus. We further annotate subtitle and street view test sets to evaluate our method. Experimental results show our proposed method outperforms the existing pipeline diagram and end-to-end text image machine translation models significantly.

The main contributions of our work are summarized as follows:

- We propose a novel end-to-end text image machine translation model named Modal contrastive learning based End-to-end Text Image Machine Translation (METIMT), which utilizes intra- and inter-modal contrastive learning to alleviate the modality gap by encoding image and text features into shared semantic feature space.
- A synthetic and two real-world datasets are constructed to train and evaluate text image translation models. The synthetic training dataset contains one million text-image pairs for each translation direction. We believe these datasets can effectively address the data limitation problem and further promote the research of text image translation. All these datasets will be released to the public.
- Extensive experimental results show our proposed METIMT model outperforms pipeline systems through alleviating the error propagation problem and achieves new state-of-the-art among existing end-to-end text image machine translation models.

II. Methodology

Our proposed modal contrastive learning based end-to-end text image machine translation model is composed of two primary parts: a transformer based end-to-end TIMT architecture and a modal contrastive learning module. The transformer based end-to-end TIMT model is the backbone of transforming source language text images into target language text sentences. Modal contrastive learning, including both intra- and inter-modal contrastive learning, is designed to improve the representation learning of text images.

A. Problem Definition

TIMT is to translate source language text images into the target language. Let $D_{\text{TIMT}} = \{I, Y\}$ be the dataset for TIMT, which consists of source language text image I, and corresponding target language translation Y. The end-to-end model optimizes the translation loss function of translating target language sequence Y given image input I:

$$\mathcal{L}_E(\theta_E) = -\sum_{(I,Y)\in D_{\text{TIMT}}} \log P(Y|I;\theta_E) \tag{1}$$

where $\mathcal{L}_E$ and θ_E are the loss function and estimated parameters of end-to-end TIMT model respectively.

B. Transformer Based TIMT Architecture

In order to translate source language text images into target language text sentences, we propose a transformer based architecture as shown in Fig. 3(a). In our proposed end-to-end TIMT model, two modalities of inputs are considered during model training: text images and text sentences. As a result, we adopt an image encoder before the transformer encoder to encode images, whereas the embedding layer is utilized to encode text sentences.

a) Image encoder: To represent image features from text image input, we extend the transformer with an external image encoder. as shown in Fig. 3(b). Image encoder is composed of TPS module [12], [13], ResNet [14] and a transformer encoder [15].

For image input, text images in natural scenes have diverse directions, which increases the difficulty of image feature encoding. To reduce this burden, the image encoder first normalizes the text image direction. As shown in Fig. 3(b), given a text image I, thin-plate spline (TPS), a variant of the spatial transformation network (STN) [16], is applied at the beginning of image encoder $\mathbf{I}(\cdot)$ to normalize input image by transforming the tilted texts in the image into the horizontal direction [12], [13], which is essential for robust training and prediction. Specifically, TPS consists of a sequence of processes: finding a text boundary, linking the location of the pixels in the boundary to those of the normalized image, and generating a normalized image by using the values of pixels and the linking information. As shown in Fig. 3(b), TPS could transform text images of diverse directions into the horizontal position and employs a smooth spline interpolation, which releases the burden of extracting the text image representation.

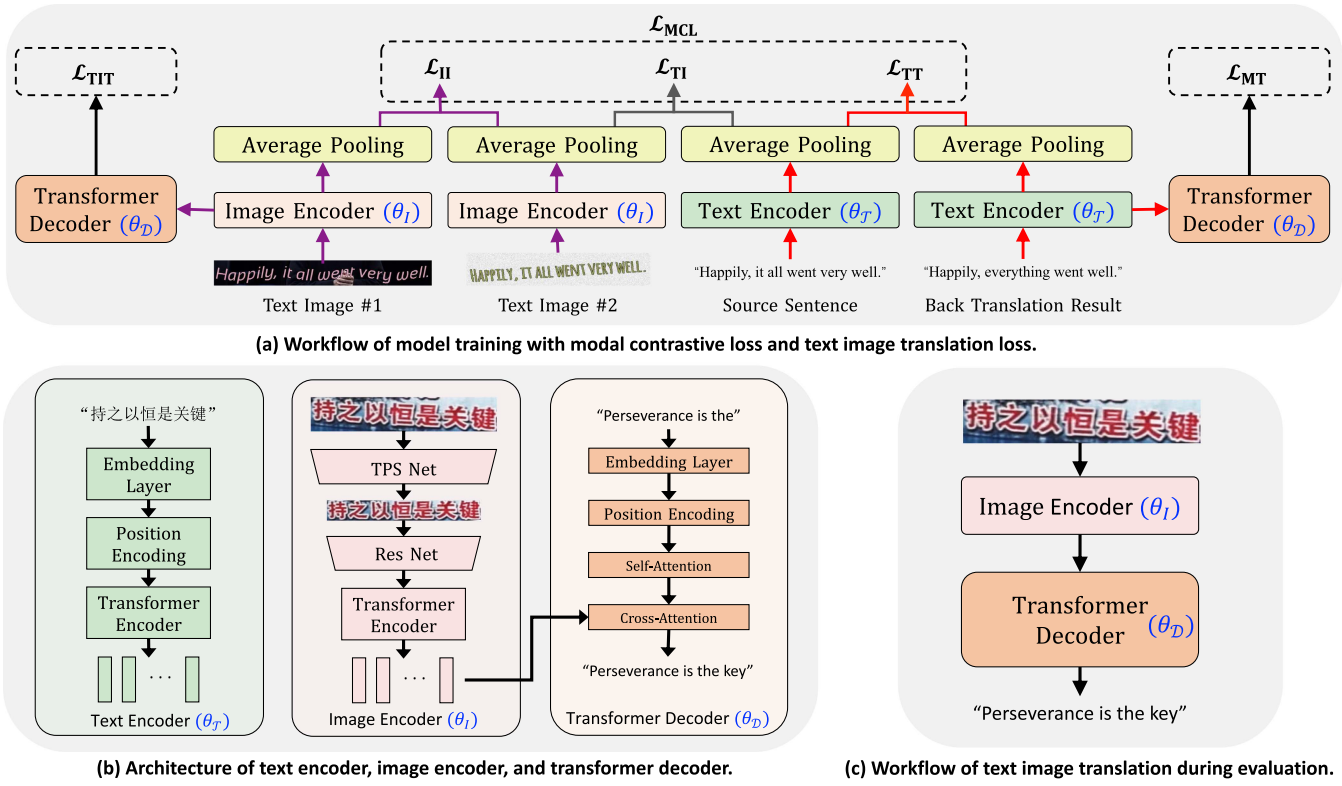


Fig. 3. Diagram overview of our proposed method. (a) depicts the training procedure of modal contrastive learning based end-to-end text image translation. (b) illustrates the details of the individual modules. (c) shows the evaluation procedure of our proposed end-to-end text image translation model.

ResNet, which is composed of stacked convolution, pooling, and residual connection layers, is utilized to extract feature maps of the given image input. Image features are extracted from the final convolutional (29-th) layer of ResNet, which we implement the same architecture as in [17]. Then, a transformer encoder $\mathbf{E}(\cdot)$ is utilized to encode semantic features:

$$F_{\mathbf{I}} = \mathbf{I}(I;\theta_{\mathbf{I}}) = \mathbf{E}(\text{ResNet}(\text{TPS}(I))) \tag{2}$$

where $F_{\mathbf{I}} \in \mathbb{R}^{l_{\mathbf{I}} \times h_{\mathbf{E}}}$ is the encoded image feature. $l_{\mathbf{I}}$ and $h_{\mathbf{E}}$ represent the length of image feature sequence and the hidden size of transformer encoder. $\theta_{\mathbf{I}}$ denotes parameters in image encoder.

b) Text encoder: Text encoder $\mathbf{T}(\cdot)$ is composed of a standard transformer encoder as shown in Fig. 3(b):

$$F_{\mathbf{T}} = \mathbf{T}(X;\theta_{\mathbf{T}}) = \mathbf{E}(X) \tag{3}$$

where $F_{\mathbf{T}} \in \mathbb{R}^{l_{\mathbf{T}} \times h_{\mathbf{E}}}$ is the encoded text feature of source language input X. $l_{\mathbf{T}}$ represents the length of text feature sequence. Notice we utilize the same transformer hidden size $h_{\mathbf{E}}$ in the image encoder and text encoder. $\theta_{\mathbf{T}}$ denotes parameters in text encoder.

Global representation is calculated after the average pooling operation:

$$\overline{F_{\mathbf{I}}} = \text{AvgPool}(F_{\mathbf{I}}); \quad \overline{F_{\mathbf{T}}} = \text{AvgPool}(F_{\mathbf{T}}) \tag{4}$$

where $\overline{F_{\mathbf{I}}} \in \mathbb{R}^{h_{\mathbf{E}}}$ and $\overline{F_{\mathbf{T}}} \in \mathbb{R}^{h_{\mathbf{E}}}$ represent global representation of image and text features, which are utilized to calculate the contrastive loss in modal contrastive learning module.

c) Decoder: A transformer decoder $\mathbf{D}(\cdot)$ is utilized to decode target language auto-regressively given text image feature $F_{\mathbf{I}}$:

$$F_{\mathbf{D}}^{(I)} = \mathbf{D}(Y_h, F_{\mathbf{I}}; \theta_{\mathbf{D}}) \tag{5}$$

where $F_{\mathbf{D}}^{(I)}$ represents transformer decoder output based on text image features. Y_h denotes decoded history. $\theta_{\mathbf{D}}$ denotes parameters in decoder.

Parameters in the TIMT task are optimized with the cross entropy loss function:

$$\mathcal{L}_{\text{TIMT}} = -\sum_{(I,Y)\in D_{\text{TIMT}}} \log P(Y|I;\theta_{\mathbf{I}},\theta_{\mathbf{D}})$$

$$P(Y|I) \propto \exp(W_o F_{\mathbf{D}}^{(I)}) \tag{6}$$

where W_o is a linear transformation mapping decoder features into target language space.

C. Modal Contrastive Learning

In the text image machine translation task, text images and text sentences with similar semantic meaning should be close in semantic feature space as shown in Fig. 2. Meanwhile, images with the same text sentence content and different background images should also have similar semantic features. In order to learn semantic invariant representation, we adopt both intra-modal and inter-modal contrastive learning to improve TIMT performance as shown in Fig. 3(a). Intra-modal contrastive learning includes image-image and text-text contrastive learning, while

inter-modal contrastive learning denotes text-image contrastive learning.

a) Text-text contrastive learning: In order to learn semantic invariant text representation, we implement Text-Text Contrastive Learning (TT-CL) to align text features of similar sentences. Positive pairs of text-text contrastive learning are generated by back translation [18], which translates the source sentence into the target language and then translates back to the source language. Let $\overline{F_{\mathbf{T}}^{i}}$ be the global feature of i-th source text X^i, $\overline{F_{\mathbf{T}}^{i+}}$ be the global feature of back translation result, and $\overline{F_{\mathbf{T}}^{j}}$ be the global feature of j-th source text X^j in the mini-batch. Text-text contrastive loss is calculated as:

$$\mathcal{L}_{\text{TT}}^{i} = -\log \frac{\exp(d(\overline{F_{\mathbf{T}}^{i}}, \overline{F_{\mathbf{T}}^{i+}})/\tau)}{\sum_{j}^{K} \exp(d(\overline{F_{\mathbf{T}}^{i}}, \overline{F_{\mathbf{T}}^{j}})/\tau)} \tag{7}$$

where K represents the size of the mini-batch. τ represents temperature parameter and $d(q, k)$ denotes similarity metric.

b) Image-image contrastive learning: Given the same sentence, its text image may have various formats due to different fonts, background images, and other image augmentation operations. However, the semantic representation of these text images should be similar. Based on this motivation, we implement Image-Image Contrastive Learning (II-CL) to learn image format invariant features of text images. Let $\overline{F_{\mathbf{I}}^{i}}$ be the global feature of i-th text image I^i, $\overline{F_{\mathbf{I}}^{i+}}$ be the global feature of text image with the same content as I^i, and $\overline{F_{\mathbf{I}}^{j}}$ be the global feature of j-th text image I^j in the mini-batch. Image-image contrastive loss is calculated as:

$$\mathcal{L}_{\text{II}}^{i} = -\log \frac{\exp(d(\overline{F_{\mathbf{I}}^{i}}, \overline{F_{\mathbf{I}}^{i+}})/\tau)}{\sum_{j}^{K} \exp(d(\overline{F_{\mathbf{I}}^{i}}, \overline{F_{\mathbf{I}}^{j}})/\tau)} \tag{8}$$

c) Text-image contrastive learning: Considering the semantic representation of text image should be similar to its text sentence, inter-modal Text-Image Contrastive Learning (TI-CL) is incorporated by aligning image features and text features. Let $\overline{F_{\mathbf{T}}^{i}}$ be the global text feature of i-th source text X^i, $\overline{F_{\mathbf{I}}^{i}}$ be the global image feature of i-th text image I^i. $\overline{F_{\mathbf{I}}^{j}}$ represents the global text feature of j-th text image I^j in the mini-batch. Text-image contrastive loss is calculated as:

$$\mathcal{L}_{\text{TI}}^{i} = -\log \frac{\exp(d(\overline{F_{\mathbf{T}}^{i}}, \overline{F_{\mathbf{I}}^{i}})/\tau)}{\sum_{j}^{K} \exp(d(\overline{F_{\mathbf{T}}^{i}}, \overline{F_{\mathbf{I}}^{j}})/\tau)} \tag{9}$$

After weighted summation of different contrastive losses, the final modal contrastive learning loss is:

$$\mathcal{L}_{\text{MCL}} = \sum_{i}^{|D_{\text{TIMT}}|} (\lambda_{\text{II}} \cdot \mathcal{L}_{\text{II}}^{i} + \lambda_{\text{TT}} \cdot \mathcal{L}_{\text{TT}}^{i} + \lambda_{\text{TI}} \cdot \mathcal{L}_{\text{TI}}^{i}) \tag{10}$$

where $\lambda_{\text{II}}, \lambda_{\text{TT}}$ and λ_{TI} represent the weight of corresponding loss functions. With different weighted summations of contrastive losses, we evaluate the effectiveness of intra-modal and inter-modal contrastive learning for TIMT. Notice we utilize the same temperature parameter $\tau = 0.2$ and cosine distance based similarity function $d(q, k) = q^T k / \|q\| \|k\|$ for all contrastive loss functions.

D. Multi-Task Learning With Text Translation

In order to fully utilize the text parallel corpus, we incorporate text translation auxiliary task during training. Specifically, the transformer decoder $\mathbf{D}(\cdot)$ is utilized to decode target language auto-regressively given source language text features $F_{\mathbf{T}}$:

$$F_{\mathbf{D}}^{(T)} = \mathbf{D}(Y_h, F_{\mathbf{T}}; \theta_{\mathbf{D}}) \tag{11}$$

where $F_{\mathbf{D}}^{(T)}$ represents transformer decoder output based on source language text features. Y_h denotes decoded history. $\theta_{\mathbf{D}}$ denotes parameters in decoder. The final text translation loss function is:

$$\mathcal{L}_{\text{MT}} = \sum_{(X,Y) \in D_{\text{MT}}} \log P(Y|X; \theta_{\mathbf{T}}, \theta_{\mathbf{D}})$$

$$P(Y|X) \propto \exp(W_o F_{\mathbf{D}}^{(T)}) \tag{12}$$

where D_{MT} represents the text parallel corpus. X, Y denote the source language sentence and target language sentence respectively. W_o is a linear transformation mapping decoder features into target language space as introduced in the text image translation task.

E. Training and Inference

a) Training: Training procedure of our proposed method optimizes text image translation loss $\mathcal{L}_{\text{TIMT}}$, modal contrastive loss $\mathcal{L}_{\text{MCL}}$, and text translation loss $\mathcal{L}_{\text{MT}}$ simultaneously as shown in Fig. 3(a):

$$\mathcal{L} = \lambda_{\text{TIMT}} \cdot \mathcal{L}_{\text{TIMT}} + \lambda_{\text{MCL}} \cdot \mathcal{L}_{\text{MCL}} + \lambda_{\text{MT}} \cdot \mathcal{L}_{\text{MT}} \tag{13}$$

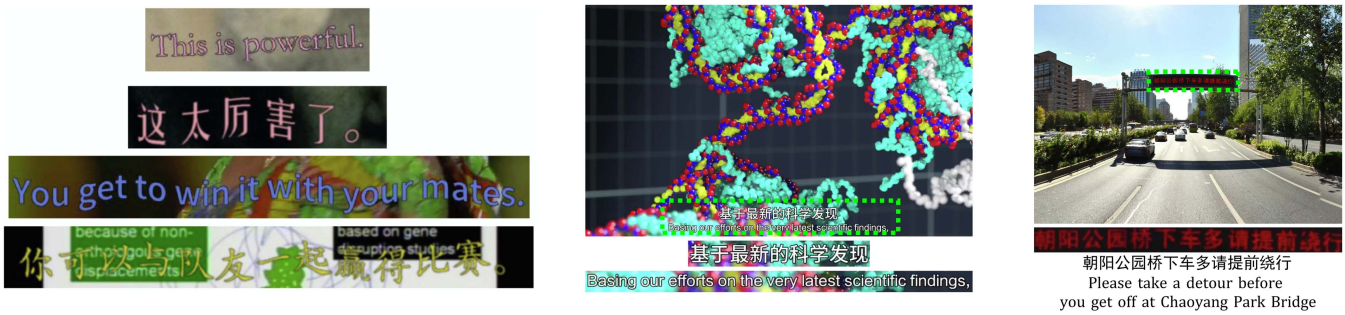
where λ_{TIMT}, λ_{MCL} and λ_{MT} denotes the weight of end-to-end text image translation loss, modal contrastive learning loss, and text translation loss functions respectively.

b) Inference: The end-to-end TIMT model is evaluated after removing the text encoder, which means the text information is only utilized during training. As shown in Fig. 3(c), our proposed model translates the target language based on source language text image input without any text input or modal contrastive information during evaluation.

III. Experiments

A. Dataset

To alleviate data limitation in text image translation, we construct a synthetic TIMT training dataset based on the bilingual parallel corpus. Meanwhile, a synthetic test set, a subtitle test set, and a street view test set are annotated to evaluate our proposed method as shown in Fig. 4. Three translation directions are utilized in this work: English-to-Chinese (En $\Rightarrow$ Zh), English-to-German (En $\Rightarrow$ De), and Chinese-to-English (Zh $\Rightarrow$ En). For synthetic dataset construction, we first prepare text parallel corpus, font types, and background images. Then a text image generator is utilized to synthesize images with source language text in them.



(a) Synthetic Samples (b) Subtitle Samples (c) Street View Samples

Fig. 4. Examples of text image translation datasets.

1) Preprocessing of Text Data: Text parallel data from Workshop of Machine Translation[1] is utilized as the text content of the synthetic text images. Before pre-processing, we remove illegal sentences in the corpus which include duplicate sentences and sentences in different languages other than the source or target language (filtered by our language detection tools). After this step, we obtain 5,984,287 (around 6 M) En ⇔ Zh parallel sentences and 20,895,771 (around 21 M) En ⇒ De parallel sentences.

Preprocessing steps include escape character transformation, text normalization, and tokenization. First, all escape characters are transformed into corresponding marks with a well-designed rule-based replacement method. Second, numbers and punctuations are normalized into single-byte characters (SBC), and non-print symbols or marks are removed. Third, we tokenize sentences and obtain subword granularity tokenized sentences. English and German sentences are tokenized by the toolkit in Moses.[2] Chinese sentences are tokenized by Jieba.[3] We then split tokens into subword units.[4] The number of merge operations in byte pair encoding (BPE) is set to 32,000 for all languages [19]. All these preprocessed text sentences are used for text machine translation with multi-task training in our proposed method.

For the text sentence filtering, we discard Chinese sentences longer than 40 characters and English, German sentences longer than 80 characters. This length filter threshold is set after analysis of the length distribution of real-world text images. The validation set is constructed with the same filtering method. Sentences in the IWSLT test16, IWSLT test17, WMT test17, WMT test18, and WMT test19 are used to synthesize synthetic TIMT test set with the same method as the training set. Statistics of the synthetic dataset are shown in Table I.

In summary, we finally obtain 1,000,000 parallel sentences of each language direction for text image machine translation and also keep large-scale text parallel sentences for text machine translation.

2) Text Image Synthesis: Text images are composed of text contents and background images. text contents are collected and processed as introduced in **Preprocessing of Text Data**. For text image effects, font rendering, text skewing, projective distortion, noise, font color, and background images are mainly considered as shown in Fig. 5.

- *Font Rendering:* 20 Chinese font types and 100 English font types are collected to render texts in images. Given one text sentence, one font is randomly sampled from the font set and utilized as the font in the corresponding synthetic text image. Examples of different font types are shown in Fig. 5(a) and (b).
- *Text Skewing:* The Text line in the image is randomly skewed by minus five to five degrees from the horizontal as shown in Fig. 5(c) and (d).
- *Projective Distortion:* The text line in the image is distorted with a random, full-projection transformation, simulating the 3D world. Sine and Cosine curve based distortion are utilized during text image synthesis. Examples are shown in Fig. 5(e) and (f).
- *Noise:* Gaussian noise, random blur effects, and image compression effects are introduced to the text images as shown in Fig. 5(g).
- *Font Color:* Font color is randomly sampled from all RGB color combinations as shown in Fig. 5(h).
- *Background Images:* Video frames are utilized as the background images of synthetic text images. Various genres of videos like cartoons, movies, and soap operas from Youtube[5] and TED[6] video websites are collected after splitting into frames. Finally, 51,454 frame images (22,570 YouTube video frames and 28,884 TED video frames) are utilized as the background images of synthetic text images. Examples are shown in Fig. 5(i) and (j).

TABLE I
STATISTICS OF SAMPLES IN SYNTHETIC, SUBTITLE, AND STREET VIEW TEXT IMAGE TRANSLATION DATASETS

	Synthetic			Subtitle	Street View
	#Train	#Valid	#Test	#Test	#Test
Zh⇒En	1,000,000	2,000	2,502	1,040	1,198
En⇒Zh	1,000,000	2,000	2,502	1,040	-
En⇒De	1,000,000	2,000	2,000	-	-

[1]http://www.statmt.org/wmt18/
[2]hppts://www.statmt.org/moses/
[3]https://github.com/fxsjy/jieba
[4]https://github.com/rsennrich/subword-nmt
[5]https://www.youtube.com/
[6]https://www.ted.com/

(a) Text image with *Lato-Light* font.

(b) Text image with *Pacifico* font.

(c) Text image with minus five degrees skewing from horizontal.

(d) Text image with five degrees skewing from horizontal.

(e) Text image with sine curve distortion.

(f) Text image with cosine curve distortion.

(g) Text image with Gaussian blur noise.

(h) Text image with red font color.

(i) Text image with video frame background image.

(j) Text image with random combination of all effects.

Fig. 5. Examples of text image effects.

To synthesize text images efficiently, *Text Recognition Data Generator*[7] toolkit is used to synthesize source language text images with prepared background images, randomly selected fonts, and various text shapes as shown in Fig. 4(a). The final format of the synthetic training dataset forms as triple tuples **{source language image, source language text, target language text}**.

3) Real-World Text Image Annotation: To evaluate the effectiveness of our proposed model, we also annotate real-world evaluation datasets.

- *Subtitle Dataset:* This dataset takes real-world video to generate text images. First, Videos with bilingual subtitles are collected and split into frames. The subtitle part of the video frame is detected with text detection toolkit [20]. Human translators then wrote down the subtitle transcripts and the corresponding translated results as shown in Fig. 4(b). Finally, 1,040 text images and corresponding translations are obtained in the subtitle evaluation dataset.
- *Street View Dataset:* Real-world street view images from Tencent Chinese Street View Dataset[8] is used to construct street view dataset [21]. We discard the image by considering the image size smaller than 6 KB (around 2000 pixels) because too small texts in the image are hard to recognize by humans and are seldom occurred in real-world applications. From filtered text images, human translators translate text images into the target language English as shown in Fig. 4(c). Finally, 1,198 text images and corresponding translations are obtained in the street view evaluation dataset.

B. Experimental Settings

For model settings, the image encoder which includes TPS and ResNet utilizes the same configuration in [22]. The transformer encoder and decoder utilize the configuration of transformer_base in [15], which contains 6-layer encoders and 6-layer decoders with 512-dimensional hidden sizes. The maximum length for Chinese, English, and German is set to 40, 80, and 80 characters respectively. Preprocessed image height is set to 32 pixels and the input channel is 3. To align the length of image features and text features, preprocessed image width is resized to 160, 320, and 320 pixels respectively. The batch size is set to 64 for all model settings. All models are trained with Adam optimizer [23] for 300,000 steps on 2 NVIDIA V100 GPUs. To offer a fair comparison, all models are trained with the same dataset introduced in Section III-A.

For evaluation metric, we report detokenized BLEU [24] using sacre-BLEU[9] for text image translation task on synthetic, subtitle, and street view test sets.

C. Baseline Models

We compare our proposed method with the following end-to-end text image machine translation baseline models:

- *TRBA:* This model represents the best model setting in [22], Which includes TPS net, ResNet for image feature extraction, BiLSTM for sequential modeling, and attention-based RNN for text image recognition.[10] We train this model with our constructed synthetic training set under the end-to-end text image translation protocol.

[7]https://github.com/Belval/TextRecognitionDataGenerator
[8]https://ctwdataset.github.io/
[9]https://github.com/mjpost/sacrebleu
[10]https://github.com/clovaai/deep-text-recognition-benchmark

TABLE II
PERFORMANCE OF END-TO-END TEXT IMAGE TRANSLATION ON SYNTHETIC DATASET

Index	Architecture	TI-CL	II-CL	TT-CL	En⇒Zh		En⇒De		Zh⇒En	
					Valid	Test	Valid	Test	Valid	Test
Baseline Models										
(1)	TRBA [23]	-	-	-	12.14	9.61	7.84	7.36	6.57	4.77
(2)	CLTIR [7]	-	-	-	20.49	18.02	17.16	15.55	12.93	10.74
(3)	ItNet [6]	-	-	-	21.07	18.43	17.19	15.71	13.56	11.38
(4)	RTNet [8]	-	-	-	21.44	18.91	17.24	15.82	14.08	12.54
(5)	CLTIR w/ OCR Multi-task [7]	-	-	-	22.71	19.44	17.80	16.31	15.25	13.52
(6)	RTNet w/ OCR Multi-task [8]	-	-	-	22.87	19.63	17.86	16.78	15.87	14.01
Our Models										
(7)	TPS+ResNet+Transformer	-	-	-	22.25	19.25	16.51	16.27	14.96	13.16
(8)	w/ Intra-modal MCL	-	-	✓	22.89	19.73	17.94	17.08	16.01	14.21
(9)		-	✓	-	23.80	20.74	18.69	18.25	16.42	14.45
(10)		-	✓	✓	24.21	21.39	19.31	19.16	16.65	14.69
(11)	w/ Inter-modal MCL	✓	-	-	24.23	22.56	20.01	19.42	16.83	14.97
(12)	w/ Intra- and Inter-modal MCL	✓	-	✓	25.44	22.95	20.48	19.58	17.19	15.06
(13)		✓	✓	-	25.67	23.10	21.15	20.04	17.64	15.23
(14)	METIMT	✓	✓	✓	**26.00**	**23.21**	**21.74**	**20.15**	**17.96**	**15.74**

In order to provide a fair comparison, all these models are trained with our constructed synthetic text image translation training dataset. Different settings of modal contrastive learning are determined by contrastive loss weight mentioned in (10), and the best model setting is named as METIMT.

- *CLTIR:* A cross-lingual text image recognition (CLTIR) architecture is proposed in [7], which contains a convolutional image encoder, a semantic encoder, a target language decoder, and an alignment model. It further trained the end-to-end TIMT model and text line recognition model simultaneously with a shared convolutional image encoder. For text line recognition multi-task setting, bidirectional long short term memory (BiLSTM) and connectionist temporal classification (CTC) are combined to generate source language recognition results.
- *ItNet:* This model utilizes a convolutional neural network for image encoding, and a transformer decoder for target language generation [6]. Specifically, five different encoder settings (like variants of ResNet and DarkNet) and four different decoder settings (variants of the transformer) are studied how encoder and decoder size affect model accuracy. To offer a faire comparison, ResNet and transformer_Base setting is utilized in this paper due to the similar model size to other baseline models.
- *RTNet:* Feature transformer is proposed to bridge the semantic gaps between image encoder and text decoder [8]. Specifically, a pre-trained text image recognition encoder is connected with a pre-trained text machine translation decoder through a feature transformer module, and the feature transformer is trained on synthetic TIMT data with multi-task training with text line recognition task.

D. Results and Analysis

a) Results of end-to-end TIMT on synthetic dataset: Table II shows the results of end-to-end TIMT on synthetic dataset. Experimental results show our proposed end-to-end TIMT backbone ResNet+Transformer_Base (line 7) outperforms previous vanilla end-to-end architecture in all language directions (line 1-6). Results of intra-modal contrastive learning reveal that image-image contrastive learning (II-CL) is more effective for end-to-end TIMT than text-text contrastive learning (TT-CL) as shown in line 8-10 in Table II. Inter-modal text-image contrastive learning (TI-CL) (line 11) improves translation performance in all translation directions compared with intra-modal contrastive learning (line 8-10). Furthermore, the translation performance is improved after incorporating both intra-modal and inter-modal contrastive learning (line 12, 13), which indicates that intra-modal and inter-modal contrastive learning are complementary. Finally, our best model (line14) named METIMT, which is composed of text-image, image-image, and text-text contrastive learning achieves new state-of-the-art among all end-to-end text image machine translation models and multi-task enhanced models.

b) Results on real-world evaluation dataset: To evaluate the generalization of models trained with the synthetic dataset, we also evaluate models on real-world datasets as shown in Table IV. With both intra-modal and inter-modal contrastive learning, our proposed method achieves the best translation performance among all models, which is consistent with the results in the synthetic evaluation dataset. Text images in the subtitle dataset are always shown with standard fonts and clean background images. As a result, the translation performance of the end-to-end model is even comparable with text machine translation. The street view test set is extracted from real-world images on the street and the texts in it are strongly influenced by light, background, and various fonts, which is still a big challenge for further improvements.

c) Comparison with pipeline system: We compare the performance of the end-to-end TIMT model with the pipeline diagram. For text line recognition models, we evaluate recognition performance on English synthetic text image recognition datasets, and Table III shows the word error rate (WER) of recognition results. Compared with the recurrent neural network based VGG+BiLSTM+CTC [25] and TPS+ResNet+BiLSTM+Attn [22], ResNet+Transformer achieves lower WER on text line image recognition, which

TABLE III
COMPARISON OF VARIOUS PIPELINE AND METIMT ARCHITECTURES ON ENGLISH-TO-CHINESE SYNTHETIC VALIDATION SET

Pipeline System				METIMT		
Text Image Recognition Model	MT Model	WER (↓)	BLEU (↑)	Encoder	Decoder	BLEU (↑)
VGG+BiLSTM+CTC [26]	Transf.Base	21.06%	17.62	VGG+BiLSTM	Transf.Base Decoder	19.78
	Transf.Big		20.01		Transf.Big Decoder	22.64
ResNet+BiLSTM+Attention [23]	Transf.Base	18.91%	19.95	ResNet+BiLSTM	Transf.Base Decoder	22.17
	Transf.Big		21.36		Transf.Big Decoder	23.94
ResNet+Transformer	Transf.Base	16.25%	23.07	ResNet+Transf.Base Encoder	Transf.Base Decoder	26.00
	Transf.Big		25.92		Transf.Big Decoder	27.15

'Transf.' is the abbreviation of transformer. METIMT utilizes the similar encoder-decoder to corresponding pipeline system.

TABLE IV
PERFORMANCE OF TEXT IMAGE TRANSLATION ON REAL-WORLD TEST DATASETS

Architecture	Subtitle		Street View
	En⇒Zh	Zh⇒En	Zh⇒En
TRBA [23]	12.12	5.18	0.36
CLTIR [7]	16.47	9.04	0.43
ItNet [6]	16.91	10.08	0.94
RTNet [8]	17.63	10.63	1.07
CLTIR+OCR [7]	17.96	11.25	1.74
RTNet+OCR [8]	18.82	11.50	1.93
METIMT	**19.49**	**12.13**	**5.91**

TABLE V
COMPARISON BETWEEN PIPELINE SYSTEM AND OUR BEST END-TO-END TEXT IMAGE TRANSLATION MODEL ON SYNTHETIC TEST SET

Direction	Metrics	Pipeline	METIMT	Δ
En⇒Zh	BLEU	20.46	23.21	↑ 2.75
	Params.	195.1	121.9	↓ 37.5%
	Speed	3.07	5.21	↑ 1.70x
En⇒De	BLEU	16.48	20.15	↑ 3.67
	Params.	178.7	113.7	↓ 36.4%
	Speed	2.63	4.32	↑ 1.64x
Zh⇒En	BLEU	15.12	15.74	↑ 0.62
	Params.	224.7	136.7	↓ 39.2%
	Speed	4.79	8.96	↑ 1.87x

Params. represents the amount of trainable parameters in the model, and unit for parameters is million ($\times 10^6$). The unit of speed is sentence per second.

means incorporating transformer is better for text line recognition task. For the machine translation task, Transformer_Base and Transformer_Big models [15] are evaluated, and Transformer_Big achieves higher BLEU score than Transformer_Base in pipeline text image translation task. METIMT models represent various end-to-end models trained with modal contrastive learning. From this experiment, stronger encoders and target language decoders bring further improvements for text image translation task. As a result, ResNet and transformer combined TIMT model is taken as the principle architecture in our experiments. From the comparison between various pipeline and METIMT models, better text line image recognition and machine translation model of pipeline system improves the TIMT performance. Meanwhile, corresponding architecture based METIMT, which utilizes a similar encoder as the TIR encoder and a similar decoder as the MT decoder, outperforms pipeline systems, indicating our proposed METIMT has good generalization among different encoder-decoder architectures.

Table V shows the comparison between the ResNet+Transformer_Base pipeline system and the corresponding end-to-end METIMT model. BLEU score, amount of trainable parameters, and decoding speed are evaluated to compare the various aspects of the pipeline system and METIMT based end-to-end model. The Δ is calculated as:

$$\begin{aligned} \Delta_{\text{BLEU}} &= \text{BLEU(METIMT) - BLEU(Pipeline)} \\ \Delta_{\text{Params.}} &= \frac{\text{Params.(Pipeline)-Params.(METIMT)}}{\text{Params.(Pipeline)}} \\ \Delta_{\text{Speed}} &= \frac{\text{Speed(METIMT)}}{\text{Speed(Pipeline)}} \end{aligned} \tag{14}$$

Our end-to-end METIMT model outperforms pipeline system in all three translation directions, with 2.75, 3.67, and 0.62 BLEU improvements on English-to-Chinese, English-to-German, and Chinese-to-English translation directions respectively. For model size, our end-to-end model decrease 37.5%, 36.4%, and 39.2% trainable parameters compared with the pipeline system respectively. Furthermore, the METIMT model outperforms the pipeline system with 1.70x, 1.64x, and 1.87x times faster decoding speed. This comparison shows our proposed method could effectively alleviate the error propagation, parameter redundancy, and decoding delay problems in pipeline systems.

To further analyze the improvements of METIMT, text images grouped by word error rate in the text line recognition model are decoded by pipeline and METIMT respectively. Fig. 7 shows the BLEU score calculated by different WER groups. When the text image recognition model in pipeline system makes few recognition errors (as shown in group $WER \in [0\%, 25\%)$ of Fig. 7), the MT model can translate good results, and METIMT just achieves slight improvements. Fig. 6(a) and (b) show cases in $WER \in [0\%, 25\%)$ group. METIMT performs better than the pipeline system in Fig. 6(a), where METIMT translates most of the meaning, but the TIR model made a recognition error of 'brain' into 'bird' leading further translation error. In Fig. 6(b), the pipeline system performs well due to the perfect recognition results of the TIR model, but METIMT makes minor errors in translation results. In group $WER \in [25\%, 50\%)$ and $WER \in [50\%, 75\%)$ of Fig. 7, METIMT performs much better than pipeline system and Fig. 6(c) shows one case in $WER \in [25\%, 50\%)$ group. TIR model in the pipeline system recognizes

Recognition Prediction	Remember, when you're under stress, the **bird** releases **sortisol**.
Recognition Ground Truth	Remember, when you're under stress, the **brain** releases **cortisol**.
Pipeline Result (Pinyin)	记得， 当 你 承受 压力 时， **鸟儿** 会 释放 **布里 蒂尔** 。 (jide, dang ni chengshou yali shi, **niaoer** hui shifang **buli dier**.)
METIMT (Pinyin)	记住，当 你 受 压力 时，**大脑** 释放 了 **皮肤** 。 (jizhu, dang ni shou yali shi, **danao** shifang le **pifu**.)
Translation Ground Truth (Pinyin)	记住， 当 你 有 压力 时， 大脑 会 释放 皮质醇 。 (jizhu, dang ni you yali shi, **danao** hui shifang **pizhichun**.)

(a) WER∈ [0%, 25%) | METIMT Performs Better Than Pipeline

Recognition Prediction	And the side effects in that case occur in 50 percent of the patients.
Recognition Ground Truth	And the side effects in that case occur in 50 percent of the patients.
Pipeline Result (Pinyin)	在 这种 **情况** 下， 副作用 发生 在 50% 的 患者 。 (zai zhezhong **qingkuang** xia, fuzuoyong fasheng zai 50% de huanzhe.)
METIMT (Pinyin)	这个 **案件** 的 副作用 在 50% 的 病人 中 出现 . (zhege **anjian** de fuzuoyong zai 50% de bingren zhong chuxian.)
Translation Ground Truth (Pinyin)	而 这种 **情况** 的 副作用 会 发生 在 50% 的 患者 身上 。 (er zhezhong **qingkuang** de fuzuoyong hui fasheng zai 50% de huanzhe shenshang.)

(b) WER∈ [0%, 25%) | Pipeline Performs Better Than METIMT

Recognition Prediction	**T hal mas** the day I fell in love with the ocean.
Recognition Ground Truth	That was the day I fell in love with the ocean.
Pipeline Result (Pinyin)	我 **不 喜欢** 海洋 的 日子 。 (wo **bu xihuan** haiyang de rizi.)
METIMT (Pinyin)	那 是 我 **爱上 了** 大海 的 那 一天 。 (na shi wo **aishang le** dahai de na yitian.)
Translation Ground Truth (Pinyin)	从 那天 起， 我 深深 **爱上 了** 海洋 。 (cong natian qi, wo shenshen **aishang le** haiyang.)

(c) WER∈ [25%, 50%) | METIMT Performs Better Than Pipeline

Recognition Prediction	**An AIDST FOR work wars or FITTING, MINURSING TO PIN WIND while 15.**
Recognition Ground Truth	AND AFTER **FOUR WEEKS** OF FEEDING, THEY WERE READY TO **SPIN WITH US.**
Pipeline Result (Pinyin)	在 **15 岁** 的 时候，**一个 为 工作 而 工作 的 人** . (zai **15 sui** de shihou, **yige wei gongzuo er gongzuo de ren**.)
METIMT (Pinyin)	在 吃 了 **四周** 的 午餐 之后， **我们 就 不再 去 玩 了** 。 (zai chi le **sizhou** de wucan zhihou, **women jiu buzai qu wan le**.)
Translation Ground Truth (Pinyin)	喂养 **四周** 之后， 它们 开始 **为 我们 吐丝** 。 (weiyang **sizhou** zhihou, tamen kaishi **wei women tusi**.)

(d) WER∈ [75%, 100%) | Both Pipeline and METIMT Perform Poor

Fig. 6. Case study of text image translation in various WER groups. Red color means mistakes, and green color means correct translation.

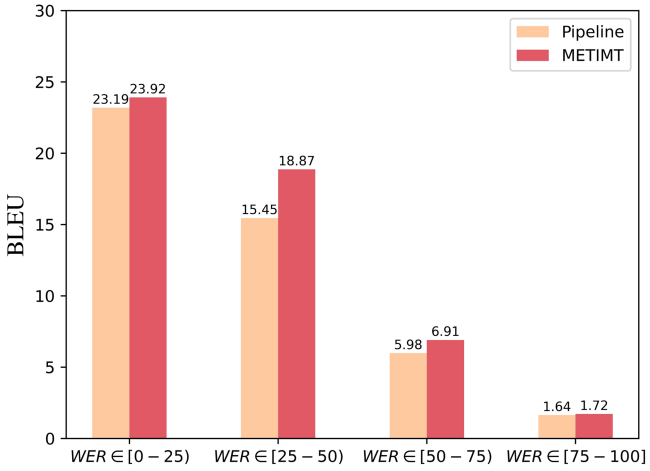


Fig. 7. Comparison of experimental results grouped by word error rate in pipeline system. $WER \in [25\%, 50\%)$ represents sentences of which word error rate is greater than or equal to 20% and less than 50%.

"That was" as "T hal mas" by mistake, which is further propagated by the MT model. METIMT effectively addresses the error propagation problem and translates the correct target sentence. When the text image is too difficult to recognize as shown in Fig. 6(d), which is a case in $WER \in [75\%, 100\%)$ group, both pipeline and METIMT models perform poor due to the difficulty of recognizing texts in such images. As a result, METIMT can alleviate the error propagation problems in pipeline system when the TIR model makes several recognition errors.

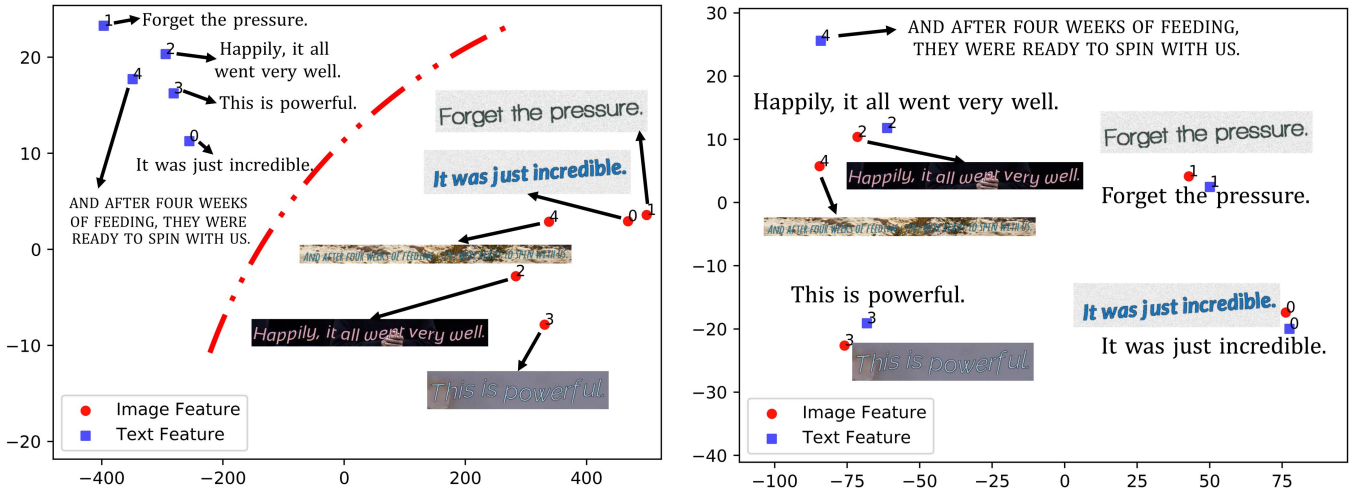
d) Comparison with production OCR and MT system: To compare with the production OCR system followed by a translation system, METIT is compared with Google Tesseract-OCR [11] and Google MT .[12] Table VI shows the results of the Google Tesseract-OCR&MT pipeline system, METIT, and OCR Ground Truth+Google MT. OCR Ground Truth+Google MT represents the inputs of the MT system are ground truth of text images, which have no recognition error. As a result, OCR Ground Truth+Google MT is the current upper-bound system for text image translation task. Since synthetic test images are rendered with various effects as introduced in Section III-A, the performance of Google Tesseract-OCR performs is limited with an average word error rate of 54.41% on synthetic test sets. Although the text direction in the subtitle test set is horizontal and the font is standard, tesseract OCR performs 33.91% word error rate due to the various background images. Recognition errors are further propagated in the Google MT system, and the translation performance is worse than METIMT. From this experiment, the existing production OCR and MT pipeline system performs not well enough on text image translation task and the main reason for the performance drop is the error made by OCR models. As a result, our proposed end-to-end METIMT can effectively address the error propagation problem in OCR and MT pipeline systems.

TABLE VI
COMPARISON WITH COMMERCIAL OCR AND MT PIPELINE SYSTEM

Architectures	Synthetic En⇒Zh	Synthetic En⇒De	Subtitle En⇒Zh
Word Error Rate of Recognition Results			
Google Tesseract-OCR	55.66%	53.15%	33.91%
BLEU Score of Translation Results			
Google Tesseract-OCR&MT	15.41	13.21	15.47
METIMT	23.21	20.15	19.49
OCR Ground Truth + Google MT	39.12	33.01	26.30

e) Effect of modal contrastive learning: To analyze the effect of modal contrastive learning, we visualize the image and text features without and with modal contrastive learning. Fig. 8 shows the text and image features after t-SNE dimension reduction. Both text and image features are reduced to 2-dimensional feature space, and the axes represent the feature values of two dimensions. Without modal contrastive learning, image and text features are encoded into separated feature subspaces as shown in Fig. 8(a), even if they are the same sentence with just modality

[11] https://github.com/tesseract-ocr/tesseract
[12] https://translate.google.com



(a) Without Modal Contrastive Learning. (b) With Modal Contrastive Learning.

Fig. 8. Visualization of image and text features after t-SNE dimensional reduction. Red round points ● denote image features while blue square points ■ represent text features. Image and text features are encoded into separated subspace when training independently as shown in (a). Modal contrastive learning aligns image and text features of the same sentence as shown in (b). Axes represent two dimensions of features after t-SNE dimension reduction.

Recognition Prediction	When we **trust** met,
Recognition Ground Truth	When we **first** met,
Pipeline Result (Pinyin)	当 我们 **信任** 时 , (dang women **xinren** shi ,)
METIMT (Pinyin)	当 我们 **第一次** 见面 时 , (dang women **diyici** jianmian shi ,)
Translation Ground Truth (Pinyin)	我们 **第一次** 见面 的 时候 , (women **diyici** jianmian de shihou ,)

(a) Case study of Subtitle Translation

Recognition Prediction (Pinyin)	**将** 之 以 **烟** 是 关键 (**jiang** zhi yi **yan** shi guanjian)
Recognition Ground Truth (Pinyin)	**持之以恒** 是 关键 (**chizhiyiheng** shi guanjian)
Pipeline Result	To **smoke** is the key
METIMT	It is key to be **patient**
Translation Ground Truth	**Perseverance** is the key

(b) Case study of Street-View Translation

Fig. 9. Case study of text image translation. Red color means mistakes, and green color means correct translation. These examples show our proposed METIT outperforms pipeline system by alleviating error propagation.

difference. Modal contrastive learning attracts different modality features of the same sentence together as shown in Fig. 8(b), indicating text images and text sentences are mapped into the same semantic feature space. Furthermore, when the text image is difficult to recognize, like the example in Fig. 6(d) and No.4 example in Fig. 8, the image and text features are not aligned well, leading to translation errors. While for text images and text sentences that have the same semantic content and similar feature vectors after modal contrastive learning, the translation performance is significantly improved by aligning image and text features before decoding.

f) Hyper-parameter analysis: Modal contrastive loss weight λ_{MCL} is the key parameter during model training. We evaluate several hyper-parameter settings as shown in Fig. 10. From this evaluation, the optimal value of λ_{MCL} is 0.3. With the increment of λ_{MCL}, the performance drops due to the main task in our work is text image translation. In order to have a good translation performance, the weight of text image translation λ_{TIMT} and weight of text translation λ_{MT} are set to 1.0 in our experiments.

TABLE VII
ABLATION STUDY OF ENGLISH-TO-CHINESE TEXT IMAGE TRANSLATION ON SYNTHETIC VALIDATION AND TEST SET

Architecture	Validation	Test
METIMT	**26.00**	**23.21**
– **TI-CL**	24.21 (↓ 1.79)	21.39 (↓ 1.82)
– **II-CL**	22.89 (↓ 1.32)	19.73 (↓ 1.66)
– **TT-CL**	22.25 (↓ 0.64)	19.25 (↓ 0.48)

g) Ablation study of various contrastive losses: Table VII shows the ablation study of components in our proposed methods. From the ablation study on the synthetic validation set, 1.79 BLEU drops without inter-modal text-image contrastive learning (TI-CL). Removing image-image contrastive learning (II-CL) hurts performance by 1.32 BLEU and 0.64 BLEU drops

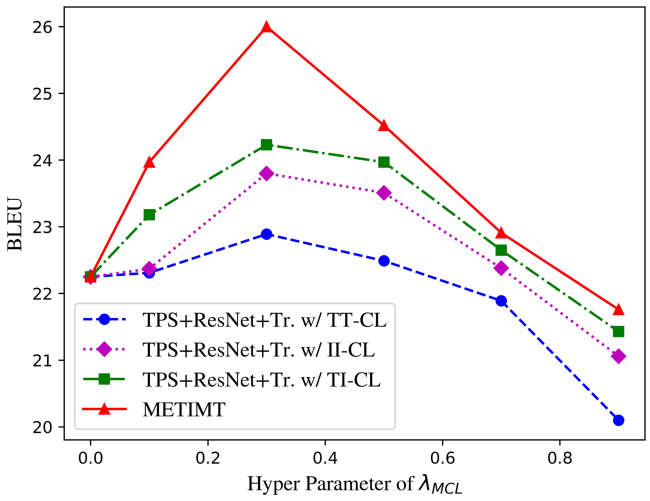


Fig. 10. Hyper-parameter evaluation of modal contrastive loss weight λ_{MCL} on English-to-Chinese synthetic valid set.

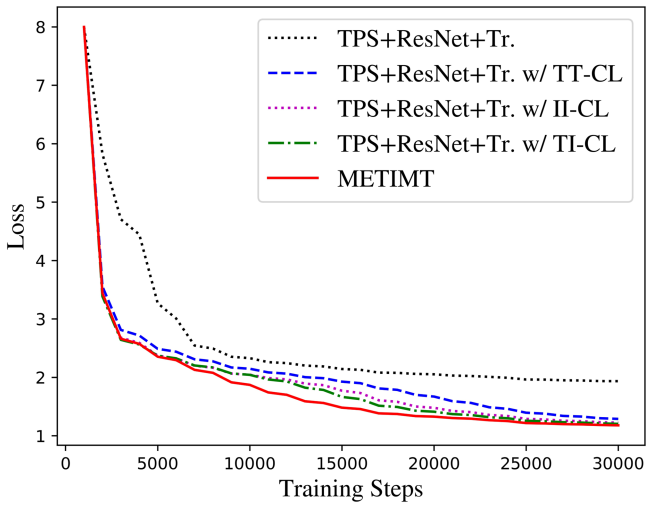


Fig. 11. Training losses of various model settings. TPS+ResNet+Tr. represents our proposed end-to-end text image translation model, which is composed of TPS, ResNet for image feature extraction, and a transformer for target language generation. TT-CL, II-CL, and TI-CL represent text-text, image-image, and text-image contrastive learning, respectively.

when removing text-text contrastive learning (TT-CL). Ablation study indicates training with modal contrastive learning is vital for text image translation.

h) Case study on real-world test sets: Fig. 9 shows the case study of text image translation on real-world test sets. In these two examples, the pipeline diagram makes mistakes of recognition, which further leads to failed translation. Our proposed model, which is trained with both intra-modal and inter-modal contrastive learning, generates good translation without errors, indicating end-to-end model could effectively alleviate the error propagation in the pipeline system.

i) Convergence analysis: Fig. 11 shows the training loss over time steps for different model settings. The vanilla end-to-end model (TPS+ResNet+Transformer, TPS+ResNet+Tr.) is difficult to optimize. Incorporating modal contrastive learning improves the convergence speed, indicating modal contrastive learning is crucial for the optimization procedure of text image translation. Furthermore, intra- and inter-modal contrastive learning could improve the training speed complementarily.

IV. Related Work

A. Text Image Machine Translation

Text image machine translation has traditionally been approached through a pipeline system which consists of a text line recognition model [22], [25], [26], [27], [28] and a text MT model [15], [29], [30]. The pipeline system recognizes texts in images with a recognition model, and then translates to the target language with a text translation model, causing error propagation, parameter redundancy, and decoding delay problems [1], [3], [4]. Recent work explores translating text images with the end-to-end model. [31] took a preliminary step for image-to-image translation by transforming source language images into target images directly without considering any text information, but the experimental result shows vanilla end-to-end image-to-image translation performs much worse than pipeline models. [7] proposed to train end-to-end TIMT by multi-task training with text image recognition task, which performs comparably with pipeline models. ItNet was designed for the TIMT task, which studied various CNN-based encoder and transformer-based decoder architecture combinations [6]. RTNet was proposed to connect the text line recognition encoder and MT decoder with a feature transformer module, and fine-tune the feature transformer module by multi-task training of end-to-end TIMT task and text image recognition task [8].

B. Contrastive Learning

Contrastive learning is an effective method for representation learning [32], [33]. Significant improvements have been shown in cross-lingual pre-training, language understanding, and text representation learning [34], [35], [36], [37], [38], [39]. Multimodal contrastive learning was proven effective to bridge text and image representation learning [9], [10], [11]. Although research in multimodal contrastive learning explores both intra-modal and inter-modal contrastive learning, they are limited in the area of multimodal understanding tasks, like visual question answering, image caption, text-image retrieval, and so on. Different from multimodal contrastive learning between images and corresponding captions, our proposed end-to-end TIMT model with modal contrastive learning aims to learn semantic invariant, background image invariant, and font invariant features for text images through both intra-modal and inter-modal contrastive learning. To the best of our knowledge, this is the first time to improve TIMT performance with modal contrastive learning.

C. Multimodal Machine Translation

Transforming visual modal data into target language textual strings, multi-modal machine translation [40], [41], [42] takes

bi-modal inputs including source language sentences and semantic aligned images to predict target language sentences. Research on MMT proposes various model architectures to incorporate visual and textual information together to guide translation. Various datasets for multimodal machine translation are proposed to improve the image-guided translation [43], [44]. Research of multimodal machine translation mainly focuses on translating source language with the help of semantically related images rather than translating the text contents in images.

V. Conclusion

In this paper, we propose a novel modal contrastive text image machine translation model to align text image and text sentence representation learning through both intra-modal and inter-modal contrastive learning. Experimental results show our end-to-end TIMT model achieves new state-of-the-art among end-to-end TIMT models. Meanwhile, our proposed method effectively alleviates error propagation problems in pipeline systems with fewer parameters and faster decoding speed. Furthermore, one synthetic and two real-world datasets are constructed to alleviate the text image translation data limitation problem. Analysis shows models trained with our synthetic dataset also have good generalization on the real-world test sets.

In the future, we will incorporate text line detection module into end-to-end text image translation and construct more scenario datasets to further promote the research of text image translation.

References

[1] R. Hinami, S. Ishiwatari, K. Yasuda, and Y. Matsui, "Towards fully automated manga translation," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, 12998–13008.

[2] K. C. Shekar, M. A. Cross, and V. Vasudevan, "Optical character recognition and neural machine translation using deep learning techniques," in *Innovations in Computer Science and Engineering*. Singapore: Springer, 2021, pp. 277–283. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-33-4543-0_30

[3] H. Afli and A. Way, "Integrating optical character recognition and machine translation of historical documents," in *Proc. Workshop Lang. Technol. Resour. Tools Digit. Humanities*, 2016, pp. 109–116.

[4] J. Chen, H. Cao, and P. Natarajan, "Integrating natural language processing with image document analysis: What we learned from two real-world applications," *Int. J. Document Anal. Recognit.*, vol. 18, no. 3, pp. 235–247, 2015.

[5] J. Du, Q. Huo, L. Sun, and J. Sun, "Snap and translate using windows phone," in *Proc. IEEE Int. Conf. Document Anal. Recognit.*, 2011, pp. 809–813.

[6] P. Jain, O. Firat, Q. Ge, and S. Liang, "Image translation network," in *Image Translation Model*, 2021. [Online]. Available: https://vigilworkshop.github.io/static/papers-2021/5.pdf

[7] Z. Chen, F. Yin, X. Zhang, Q. Yang, and C. Liu, "Cross-lingual text image recognition via multi-task sequence to sequence learning," in *Proc. IEEE 25th Int. Conf. Pattern Recognit.*, 2020, pp. 3122–3129.

[8] T. Su, S. Liu, and S. Zhou, "RTNet: An end-to-end method for handwritten text image translation," in *Proc. 16th Int. Conf. Document Anal. Recognit.*, 2021, pp. 99–113.

[9] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst. 32: Annu. Conf. Neural Inf. Process. Syst.*, 2019, pp. 13–23.

[10] X. Yuan et al., "Multimodal contrastive training for visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6995–7004.

[11] W. Li et al., "UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 2592–2607.

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

[13] W. Liu, C. Chen, K. K. Wong, Z. Su, and J. Han, "STAR-Net: A spatial attention residue network for scene text recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 7–20.

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

[15] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. 30: Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[16] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst. 28: Annu. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[17] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, and S. Zhou, "Focusing attention: Towards accurate text recognition in natural images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5086–5094.

[18] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 489–500.

[19] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 1715–1725.

[20] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9365–9374.

[21] T. Yuan, Z. Zhu, K. Xu, C. Li, T. Mu, and S. Hu, "A large chinese text dataset in the wild," *J. Comput. Sci. Technol.*, vol. 34, no. 3, pp. 509–521, 2019.

[22] J. Baek et al., "What is wrong with scene text recognition model comparisons? Dataset and model analysis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4714–4722.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.

[24] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[25] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, Nov. 2017.

[26] Y. Zhang, S. Nie, S. Liang, and W. Liu, "Robust text image recognition via adversarial sequence-to-sequence domain adaptation," *IEEE Trans. Image Process.*, vol. 30, pp. 3922–3933, 2021.

[27] Y. Zhang, S. Nie, W. Liu, X. Xu, D. Zhang, and H. T. Shen, "Sequence-to-sequence domain adaptation network for robust text image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2740–2749.

[28] 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 Recognit. Lett.*, vol. 106, pp. 20–26, 2018.

[29] Y. Zhao, J. Zhang, Y. Zhou, and C. Zong, "Knowledge graphs enhanced neural machine translation," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2020, pp. 4039–4045.

[30] Y. Zhao, L. Xiang, J. Zhu, J. Zhang, Y. Zhou, and C. Zong, "Knowledge graph enhanced neural machine translation via multi-task learning on sub-entity granularity," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 4495–4505, doi: 10.18653/v1/2020.coling-main.397.

[31] E. Mansimov, M. Stern, M. Chen, O. Firat, J. Uszkoreit, and P. Jain, "Towards end-to-end in-image neural machine translation," in *Proc. 1st Int. Workshop Natural Lang. Process. Beyond Text*, 2020, pp. 70–74.

[32] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3733–3742.

[33] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[34] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst. 26: 27th Annu. Conf. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[35] N. Saunshi, O. Plevrakis, S. Arora, M. Khodak, and H. Khandeparkar, "A theoretical analysis of contrastive unsupervised representation learning," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 5628–5637.

[36] D. Iter, K. Guu, L. Lansing, and D. Jurafsky, "Pretraining with contrastive sentence objectives improves discourse performance of language models," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4859–4870.

[37] Z. Chi et al., "InfoXLM: An information-theoretic framework for cross-lingual language model pre-training," in *Proc. Conf. North Amer. Chap. Assoc. Comput. Linguist.: Hum. Lang. Technol.*, 2021, pp. 3576–3588.

[38] H. Fang and P. Xie, "CERT: Contrastive self-supervised learning for language understanding," 2020, *arXiv:2005.12766*.

[39] J. Giorgi, O. Nitski, B. Wang, and G. Bader, "DeCLUTR: Deep contrastive learning for unsupervised textual representations," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguist. 11th Inte. Joint Conf. Nat. Lang. Process.*, Aug. 2021, vol. 1, pp. 879–895.

[40] I. Calixto and Q. Liu, "Incorporating global visual features into attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 992–1003.

[41] O. Caglayan et al., "Cross-lingual visual pre-training for multimodal machine translation," in *Proc. 16th Conf. Eur. Chap. Assoc. Comput. Linguist.: Main Vol.*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds., Apr. 2021, pp. 1317–1324.

[42] X. Huang, J. Zhang, and C. Zong, "Entity-level cross-modal learning improves multi-modal machine translation," in *Proc. Findings the Assoc. Comput. Linguistics: EMNLP*, 2021, pp. 1067–1080. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.92

[43] X. Wang, J. Wu, J. Chen, L. Li, Y. Wang, and W. Y. Wang, "VATEX: A large-scale, high-quality multilingual dataset for video-and-language research," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4580–4590.

[44] R. Sanabria et al., "How2: A large-scale dataset for multimodal language understanding," in *Proc. NeurIPS*, Montréal, Canada, 2018.