Born a BabyNet with Hierarchical Parental Supervision for End-to-End Text Image Machine Translation

Cong Ma[1,2], Yaping Zhang[1,2*], Zhiyang Zhang[1,2], Yupu Liang[1,2], 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
{cong.ma, yaping.zhang, zhiyang.zhang, yupu.liang, yang.zhao, yzhou, cqzong}@nlpr.ia.ac.cn

Abstract

Text image machine translation (TIMT) aims at translating source language texts in images into another target language, which has been proven successful by bridging text image recognition encoder and text translation decoder. However, it is still an open question of how to incorporate fine-grained knowledge supervision to make it consistent between recognition and translation modules. In this paper, we propose a novel TIMT method named as BabyNet, which is optimized with hierarchical parental supervision to improve translation performance. Inspired by genetic recombination and variation in the field of genetics, the proposed BabyNet is inherited from the recognition and translation parent models with a variation module of which parameters can be updated when training on the TIMT task. Meanwhile, hierarchical and multi-granularity supervision from parent models is introduced to bridge the gap between inherited modules in BabyNet. Extensive experiments on both synthetic and real-world TIMT tests show that our proposed method significantly outperforms existing methods. Further analyses of various parent model combinations show the good generalization of our method.

Keywords: Text image translation, knowledge inheritance, hierarchical parental supervision, contrastive learning

1. Introduction

Research of Machine translation (MT) has been achieved significant progress in recent years (Vaswani et al., 2017; Gehring et al., 2017a,b; Johnson et al., 2017; Bahdanau et al., 2015; Sutskever et al., 2014), which translates the source language texts into another target language. Text image machine translation (TIMT) is one of the vital branches of MT research by translating source language texts in images to target language texts, which has been used in various real-world applications, such as photo translation, scanned document image translation, and screen-shot translation.

Existing research on TIMT is mainly divided into two types: (1) Cascade methods combine text image recognition (TIR) and MT models to recognize-then-translate source language text images (Hinami et al., 2021; Shekar et al., 2021; Afli and Way, 2016; Chen et al., 2015; Du et al., 2011). Cascade methods have the advantage of utilizing well-trained TIR and MT models. However, when the TIR model has recognition errors, MT models will expand these errors and cause error propagation problems. Meanwhile, deploying separated TIR and MT models leads to parameter redundancy and decoding delay issues. (2) End-to-end methods utilize an image encoder to obtain source language text image features and generate the target language directly (Mansimov et al., 2020), which has a more efficient architecture and faster decoding speed compared with the cascade method. However, end-to-end methods have the shortcomings of data scarcity and the difficulty of cross-modal optimization, which limits the performance. Existing research (Ma et al., 2023c, 2022; Chen et al., 2020c) explores to incorporate external TIR or MT datasets to alleviate the problem of data limitation.

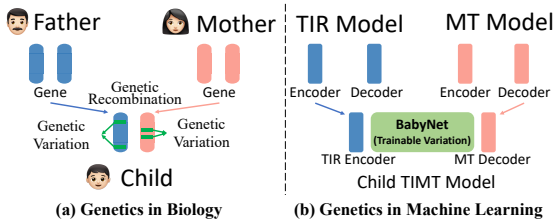


Figure 1: Diagram of Genetics in Biology and Machine Learning.

From the above-mentioned analysis, both cascade and end-to-end methods have advantages and shortcomings. Thus, how to combine the advantages to achieve a better TIMT model has attracted extensive attention recently. Modal adapter and RTNet are studied to bridge TIR encoder and MT decoder for TIMT model (Ma et al., 2023b; Su et al., 2021). However, guidance in these methods is still coarse-grained, how to alleviate the gaps between the TIR encoder and MT de-

*Corresponding author.

LREC-COLING 2024, pages 2468–2479

coder with fine-grained supervision signal remains an unsolved problem.

In this paper, we propose to bridge the pre-trained TIR and MT models with a trainable BabyNet as shown in Figure 1 (b). Inspired by genetic recombination in genetics as shown in Figure 1 (a), the TIR and MT models are regarded as parent models, which provide half modules respectively to born a child model for the TIMT task. The inherited modules from parent models are recombined together and the parameters are fixed, while a trainable BabyNet is inserted to bridge the task gap between TIR and MT modules. To better adapt BabyNet features into the MT feature space, hierarchical parental supervision is introduced to improve the optimization of BabyNet parameters. Embedding features, sequential features, and decoding probability distribution from the MT model are utilized to provide parental guidance for BabyNet. Furthermore, both local and global granularities are incorporated in hierarchical parental supervision, which is proven effective and complementary to improve translation performance. Our contributions are summarized as follows:

- We propose a novel BabyNet optimized with hierarchical parental supervision for TIMT task, which can take advantage of both cascade and end-to-end TIMT methods.
- Global and local fine-grained supervision is jointly utilized to bridge the task gap between the TIR encoder and MT decoder.
- Extensive experiment results on various parent model combinations show the effectiveness and good generalization of our method.[1]

2. Methodology

2.1. TIMT Task Formulation

TIMT task aims at translating source language text lines embedded in images into target language texts. Assume the source language text line in the image is $\mathbf{I}$ and its corresponding translation ground truth is $\mathbf{Y} = \{y_1, y_2, ..., y_t, ...y_L\}$, where the y_t denotes the t-th token and L represents the length of the target language sentence. The optimization loss function of the TIMT task is:

$$\mathcal{L}_{\text{TIMT}} = -\sum_{i=1}^{|\mathbf{D}_{\text{TIMT}}|}\sum_{t=1}^{L} \log P(\hat{y}_t^{(i)}|\mathbf{I}^{(i)}, \hat{\mathbf{Y}}_{<t}^{(i)}) \quad (1)$$

where $P(\hat{y}_t^{(i)}|\mathbf{I}^{(i)}, \hat{\mathbf{Y}}_{<t}^{(i)})$ denotes the generation probability at decoding step t. $\hat{y}_t$ represents the predicted target token and $\hat{\mathbf{Y}}_{<t}$ denotes the translation history before t-th decoding step. $\mathbf{D}_{\text{TIMT}}$ denotes the dataset of text image machine translation task containing source language image and target translation parallel pairs.

2.2. Architecture of Parent and Child Models

2.2.1. Parent Models

TIR Model Text image recognition model encodes text line images containing source language with an image encoder:

$$F_I^{\text{TIR}} = \text{Patch\_Embedding}(\mathbf{I}; \theta_I^{\text{TIR}}) \quad (2)$$

where Patch_Embedding$(\cdot)$ denotes the image patch embedding based encoder as in ViT (Dosovitskiy et al., 2021) and θ_I^{TIR} denotes the parameters of image encoder. $F_I^{\text{TIR}} \in \mathbb{R}^{l_I \cdot d_I}$ represents the image feature sequence, while l_I, d_I denote the length and dimension of image feature sequence, respectively. To model the contextual information, a sequential encoder is utilized to further encode the image features by considering the whole feature sequences. The transformer based sequential encoder (Vaswani et al., 2017; Zhao et al., 2023) is formulated as:

$$\begin{aligned} F_S^{\text{TIR}} &= \text{TransformerEncoder}(F_I^{\text{TIR}}; \theta_S^{\text{TIR}}) \\ &= \text{FFN}(\text{MultiHead}(F_I^{\text{TIR}}, F_I^{\text{TIR}}, F_I^{\text{TIR}})) \end{aligned} \quad (3)$$

where θ_S^{TIR} denotes the parameters of TIR sequential encoder. MultiHead(Q, K, V) represents the multi-head self-attention function which takes the same image feature F_I^{TIR} for query matrix Q, key matrix K, and value matrix V. FFN$(\cdot)$ denotes the position-wise fully connected feed-forward network. Note that the residual connection and layer normalization are used in each sub-layer, which are omitted in the presentation for simplicity. $F_S^{\text{TIR}} \in \mathbb{R}^{l_I \cdot d_S}$ represents the image sequential feature and d_S denotes the feature dimension.

Finally, the source language decoder generates recognized texts for text line images:

$$\begin{aligned} F_D^{\text{TIR}} &= \text{Src\_TransformerDecoder}(F_S^{\text{TIR}}; \theta_D^{\text{TIR}}) \\ P(\hat{x}_t|\mathbf{I}, \hat{\mathbf{X}}_{<t}) &\propto \exp(W_o^{\text{TIR}} F_D^{\text{TIR}}) \end{aligned} \quad (4)$$

where Src_TransformerDecoder$(\cdot)$ represents the source language transformer decoder, which consists of multi-head self-attention, cross-attention, and feed-forward sub-modules in each layer. $F_D^{\text{TIR}} \in \mathbb{R}^{l_I \cdot d_D}$ denotes the decoder feature. θ_D^{TIR} denotes the parameters of recognition decoder. $P(\hat{x}_t|\mathbf{I}, \hat{\mathbf{X}}_{<t})$ denotes the decoding probability at step t and $\hat{\mathbf{X}}_{<t}$ denotes the recognition history before step t. exp$(\cdot)$ represents the exponential function and $W_o^{\text{TIR}} \in \mathbb{R}^{|\mathcal{V}_{\mathbf{x}}| \cdot d_D}$ represents a weight ma-

[1] Our code will be released to the public.

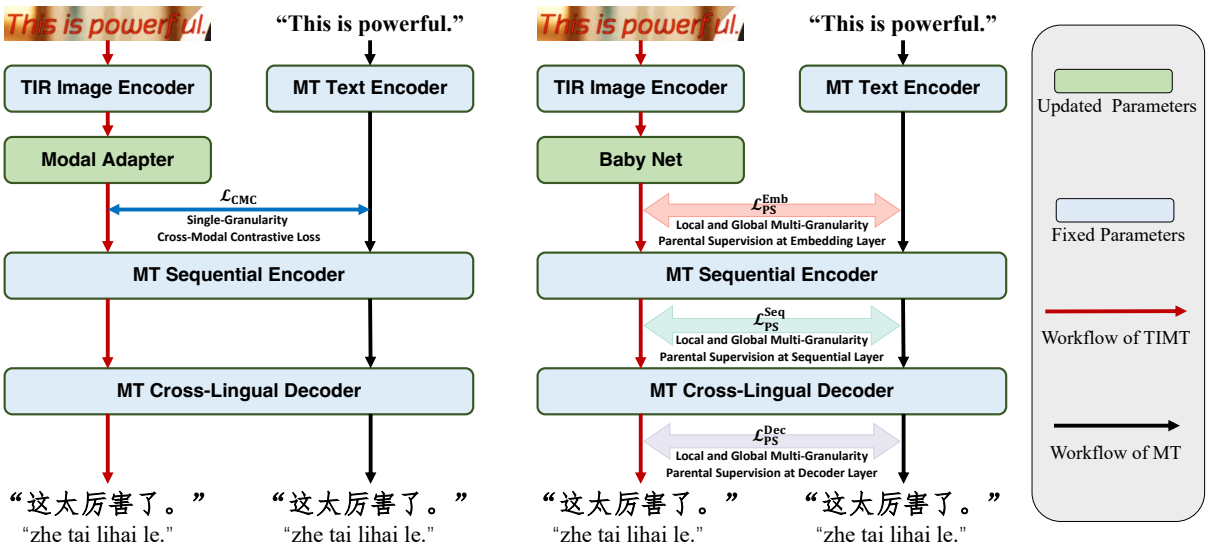


Figure 2: Diagram of our proposed BabyNet with Hierarchical Parental Supervision and comparison with related E2TIMT work (Ma et al., 2023b). Our work extends the knowledge transfer from the MT model by incorporating multi-granularity and hierarchical parental supervision. Furthermore, both cross-modal contrastive and knowledge distillation losses are utilized to fully inherit knowledge from the pre-trained MT model.

trix, which maps the decoder feature into source language vocabulary $\mathcal{V}_{\mathbf{X}}$.

MT Model Machine translation model is used to translate source language text sentences into the target language. Similar to the TIR model, the MT model also has three sub-modules: embedding encoder, sequential encoder, and cross-lingual decoder. The transformer based MT model is formulated as:

$$\begin{aligned} F_T^{\text{MT}} &= \text{Text\_Embedding}(\mathbf{X}; \theta_T^{\text{MT}}) \\ F_S^{\text{MT}} &= \text{TransformerEncoder}(F_T^{\text{MT}}; \theta_S^{\text{MT}}) \\ F_D^{\text{MT}} &= \text{Tgt\_TransformerDecoder}(F_S^{\text{MT}}; \theta_D^{\text{MT}}) \\ P(\hat{y}_t|\mathbf{X}, \hat{\mathbf{Y}}_{<t}) &\propto \exp(W_o^{\text{MT}} F_D^{\text{MT}}) \end{aligned} \tag{5}$$

where $\text{Text\_Embedding}(\cdot)$ denotes an embedding matrix that maps the source language words into dense word embeddings, $\text{Tgt\_TransformerDecoder}(\cdot)$ represents the target language transformer decoder as in Vaswani et al., (2017), $W_o^{\text{MT}} \in \mathbb{R}^{|\mathcal{V}_{\mathbf{Y}}| \cdot d_D}$ represents a mapping matrix for translation, and $\mathcal{V}_{\mathbf{Y}}$ denotes the target language vocabulary. $\theta_T^{\text{MT}}, \theta_S^{\text{MT}}, \theta_D^{\text{MT}}$ represent the parameters of text encoder, MT sequential encoder, and cross-lingual translation decoder, respectively.

2.2.2. BabyNet: The Child Model for TIMT

Text image machine translation has similar encoding functions to the TIR image encoder, while the sequential semantic encoding and cross-lingual decoding functions are similar between TIMT and MT decoders. Thus, the TIR encoder, MT sequential encoder, and MT decoder are re-combined for the TIMT task. Meanwhile, to keep the capacity of pre-trained modules, the parameters of inherited modules are not updated. Additionally, a trainable BabyNet is inserted between the TIR encoder and MT sequential encoder for variation, which aims at bridging the task gap between the TIR and MT modules. Specifically, the BabyNet receives the image feature F_I^{TIMT} and transforms it into MT feature space with a multi-layer transformer encoder architecture. By inheriting from parent models and expanding a variation module, the TIMT processing progress is then formulated as:

$$\begin{aligned} F_I^{\text{TIMT}} &= \text{Patch\_Embedding}(\mathbf{I}; \theta_I^{\text{TIR}}) \\ F_{\text{Baby}}^{\text{TIMT}} &= \text{BabyNet}(F_I^{\text{TIMT}}; \theta_{\text{BabyNet}}) \\ F_S^{\text{TIMT}} &= \text{TransformerEncoder}(F_{\text{Baby}}^{\text{TIMT}}; \theta_S^{\text{MT}}) \\ F_D^{\text{TIMT}} &= \text{Tgt\_TransformerDecoder}(F_S^{\text{TIMT}}; \theta_D^{\text{MT}}) \\ P(\hat{y}_t|\mathbf{I}, \hat{\mathbf{Y}}_{<t}) &\propto \exp(W_o^{\text{MT}} F_D^{\text{TIMT}}) \end{aligned} \tag{6}$$

where $\theta_T^{\text{TIR}}, \theta_S^{\text{MT}}, \theta_D^{\text{MT}}$ and W_o^{MT} represent inherited parameters which are not updated during training. While θ_{BabyNet} denotes the trainable variation parameters in the BabyNet. By introducing BabyNet between the TIR image encoder and the MT sequential encoder, the pre-trained TIR and MT modules are combined together, which utilizes the cascade parameters and end-to-end architecture to

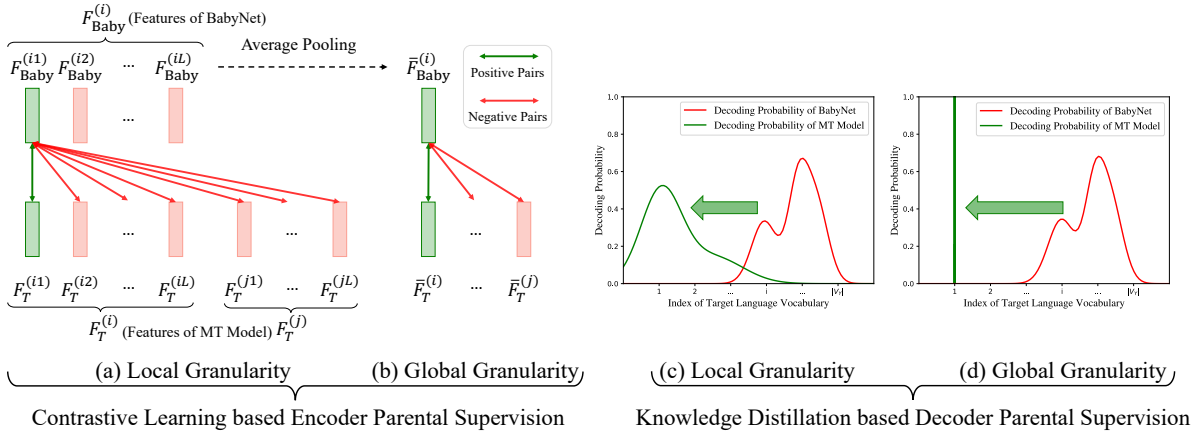


Figure 3: Diagram of contrastive and knowledge distillation based parental supervision with various granularities. With parental supervision, BabyNet is guided to have a similar feature and decoding distribution as the MT model.

translate text images with an efficient structure.

2.3. Hierarchical Parental Supervision

Features encoded by BabyNet are further fed into MT sequential encoder as shown in Eq. (6). Therefore, optimizing the BabyNet feature $F_{\text{Baby}}^{\text{TIMT}}$ to match the distribution of MT text feature F_T^{MT} can effectively improve the consistency between the BabyNet and MT sequential encoder. Figure 2 shows the hierarchical knowledge transfer method from parental supervision. Specifically, embedding encoder, sequential encoder, and decoder layer knowledge guidance from the pre-trained MT model is introduced to improve the capacity of BabyNet. Meanwhile, global- and local-granularity parental supervision is incorporated together to provide coarse- and fine-grained parental supervision.

2.3.1. Parental Supervision at Embedding Encoder Layer

To better map the BabyNet feature to the space of the MT text encoder, the contrastive learning based supervision is first calculated with the guidance from the MT text feature. Local and global granularity are utilized to provide more comprehensive supervision signals.

For local-granularity contrastive learning, as shown in Figure 3 (a), the BabyNet features and MT text features at the same position of the same sample are treated as positive pairs, while features from different positions and different samples are negative pairs. The InfoNCE (Gutmann and Hyvärinen, 2010) based contrastive loss (Chen et al., 2020a) is utilized to bring the feature representations of positive example pairs together and push away the feature representations of negative pairs. The local-granularity contrastive loss function at the embedding layer is:

$$\mathcal{L}_{\text{Local}}^{\text{Emb}} = -\sum_{k=1}^{L} \log \frac{\exp(d(F_{\text{Baby}}^{(ik)}, F_T^{(ik)})/\tau)}{\sum_{j=1}^{B}\sum_{l=1}^{L}\exp(d(F_{\text{Baby}}^{(ik)}, F_T^{(jl)})/\tau))} \tag{7}$$

where the numerator term indicates the distance between positive sample features, while the denominator term represents the distances between positive and negative sample features. i and j represent i-th and j-th training samples in a mini-batch, respectively. k and l represent the k-th and l-th position of features within a training example. B represents the batch size, and L represents the sequence length of the samples. $d(\cdot)$ denotes a vector similarity metric and cosine distance is used in our work. τ represents the temperature coefficient, which is set to 0.2 as in Chen et al., (2020b). Besides local parental supervision, global guidance is also introduced as shown in Figure 3 (b):

$$\mathcal{L}_{\text{Global}}^{\text{Emb}} = -\log \frac{\exp(d(\bar{F}_{\text{Baby}}^{(i)}, \bar{F}_T^{(i)})/\tau)}{\sum_{j=1}^{B}\exp(d(\bar{F}_{\text{Baby}}^{(i)}, \bar{F}_T^{(j)})/\tau))}$$
$$\bar{F}_{\text{Baby}}^{(i)} = \frac{1}{L}\sum_{k=1}^{L}(F_{\text{Baby}}^{(ik)}); \quad \bar{F}_T^{(i)} = \frac{1}{L}\sum_{k=1}^{L}(F_T^{(ik)}) \tag{8}$$

where $\bar{F}_{\text{Baby}}$ and $\bar{F}_T$ respectively represent the global features of the BabyNet and the text encoder. The BabyNet feature and the text feature with the same sentence content are considered positive pairs, while features with different sentence content are considered negative examples.

2.3.2. Parental Supervision at Sequential Encoder Layer

The BabyNet features are further encoded by the MT sequential encoder to incorporate contextual semantic information. To align the TIMT and MT sequential features, the MT sequential encoder provides both local and global supervision signals, which guides the TIMT sequential feature to have a consistent distribution as the MT sequential feature. Similar to embedding layer parental supervision, the contrastive loss is introduced to distinguish the TIMT and MT features and the local sequential encoder supervision loss is:

$$\mathcal{L}^{\text{Seq}}_{\text{Local}} = -\sum_{k=1}^{L} \log \frac{\exp(d(F^{(ik)}_{\text{S-TIMT}}, F^{(ik)}_{\text{S-MT}})/\tau)}{\sum_{j=1}^{B}\sum_{l=1}^{L}\exp(d(F^{(ik)}_{\text{S-TIMT}}, F^{(jl)}_{\text{S-MT}})/\tau))} \tag{9}$$

where $F_{\text{S-TIMT}}$ and $F_{\text{S-MT}}$ represent the TIMT and MT sequential features, respectively. The positive and negative pairs are constructed as that in embedding layer supervision. Besides, the global guidance at the sequential encoder layer is formulated as:

$$\begin{aligned} \mathcal{L}^{\text{Seq}}_{\text{Global}} &= -\log \frac{\exp(d(\bar{F}^{(i)}_{\text{S-TIMT}}, \bar{F}^{(i)}_{\text{S-MT}})/\tau)}{\sum_{j=1}^{B}\exp(d(\bar{F}^{(i)}_{\text{S-TIMT}}, \bar{F}^{(j)}_{\text{S-MT}})/\tau))} \\ \bar{F}^{(i)}_{\text{S-TIMT}} &= \frac{1}{L}\sum_{k=1}^{L}(F^{(ik)}_{\text{S-TIMT}}); \quad \bar{F}^{(i)}_{\text{S-MT}} = \frac{1}{L}\sum_{k=1}^{L}(F^{(ik)}_{\text{S-MT}}) \end{aligned} \tag{10}$$

where $\bar{F}^{(i)}_{\text{S-TIMT}}$ and $\bar{F}^{(i)}_{\text{S-MT}}$ denote the global TIMT and MT sequential features through average pooling for the i-th sample, respectively.

2.3.3. Parental Supervision at Decoder Layer

To transfer the cross-lingual generation capacity from the parent MT model to the child TIMT model, the decoding distribution of the MT decoder is utilized to distillate the decoding knowledge into the TIMT decoder. As shown in Figure 3 (c), the local-granularity knowledge transfer replaces the ground-truth one-hot distribution into the decoding probability distribution generated by the MT decoder generated when calculating cross-entropy loss function:

$$\mathcal{L}^{\text{Dec}}_{\text{Local}} = -\sum_{t=1}^{L}\sum_{y_t \in \mathcal{V}_{\mathbf{Y}}} Q(\hat{y}_t|\hat{\mathbf{Y}}_{<t}, \mathbf{X}^{(i)}) \log P(\hat{y}|\hat{\mathbf{Y}}_{<t}, \mathbf{I}^{(i)}) \tag{11}$$

where $Q(\hat{y}_t|\hat{\mathbf{Y}}_{<t}, \mathbf{X}^{(i)})$ represents the MT decoding distribution given i-th source language sentence and decoding history before t-th step. While $P(\hat{y}|\hat{\mathbf{Y}}_{<t}, \mathbf{I}^{(i)})$ denotes the TIMT decoding probability. $\mathcal{V}_{\mathbf{Y}}$ represents the target language vocabulary. Global-granularity parental supervision at the decoder layer is introduced to replace ground-truth one-hot distribution with the MT model decoded one-hot distribution as shown in Figure 3 (d):

$$\mathcal{L}^{\text{Dec}}_{\text{Global}} = -\sum_{t=1}^{L}\sum_{y_t \in \mathcal{V}_{\mathbf{Y}}} \mathbb{I}_{\text{MT}}(\hat{y}_t) \log P(\hat{y}_t|\hat{\mathbf{Y}}_{<t}, \mathbf{I}^{(i)}) \tag{12}$$

where $\mathbb{I}_{\text{MT}}(\cdot)$ denotes the indicator function that takes the value of 1 when the decoded token $\hat{y}_t$ by the TIMT decoder is the same as MT generated token, and 0 otherwise.

2.4. Fused Loss Functions

By integrating the multi-granularity and hierarchical parental supervision (PS), the final optimization objective during training is:

$$\begin{aligned} \mathcal{L}_{\text{ALL}} &= \lambda_{\text{TIMT}}\mathcal{L}_{\text{TIMT}} + \lambda_{\text{PS}}(\mathcal{L}^{\text{Emb}}_{\text{PS}} + \mathcal{L}^{\text{Seq}}_{\text{PS}} + \mathcal{L}^{\text{Dec}}_{\text{PS}}) \\ \mathcal{L}^{\text{Emb}}_{\text{PS}} &= \mathcal{L}^{\text{Emb}}_{\text{Local}} + \mathcal{L}^{\text{Emb}}_{\text{Global}}, \mathcal{L}^{\text{Seq}}_{\text{PS}} = \mathcal{L}^{\text{Seq}}_{\text{Local}} + \mathcal{L}^{\text{Seq}}_{\text{Global}} \\ \mathcal{L}^{\text{Dec}}_{\text{PS}} &= \mathcal{L}^{\text{Dec}}_{\text{Local}} + \mathcal{L}^{\text{Dec}}_{\text{Global}}, \ \lambda_{\text{TIMT}} + \lambda_{\text{PS}} = 1 \end{aligned} \tag{13}$$

where λ_{TIMT} and λ_{HPS} are hyper-parameters to control the weight of various loss functions. $\mathcal{L}^{\text{Emb}}_{\text{HPS}}, \mathcal{L}^{\text{Seq}}_{\text{HPS}}, \mathcal{L}^{\text{Dec}}_{\text{HPS}}$ represent parental supervision at embedding, sequential, and decoder layers. During training, both source language text images and corresponding texts contribute to parent supervision through hierarchical knowledge transfer and contrastive learning. For inference, only the source language text image is used for encoding, and source language texts are not utilized. In essence, inference transforms the source language text image into target language translation results.

3. Experiment

3.1. Datasets

TIR and MT Datasets for Parent Models. To provide a fair comparison, the dataset utilized to train the parent models is kept the same as multi-task learning based end-to-end TIMT work (Chen et al., 2020c; Ma et al., 2022). For TIR task, MJSynth (MJ) (Jaderberg et al., 2014)[2], SynthText (ST) (Gupta et al., 2016)[3], and Synthetic Text Line Dataset (Ma et al., 2022) are utilized to optimize the parameters of TIR model. Parallel sentences from the Workshop of Machine Translation [4] are utilized to train the MT parent model.

TIMT Dataset for BabyNet. The public end-to-end TIMT dataset released by Ma et al., (2022) is utilized to train all end-to-end TIMT models and our proposed BabyNet. This dataset contains 1

[2]https://www.robots.ox.ac.uk/vgg/data/text/
[3]https://www.robots.ox.ac.uk/vgg/data/scenetext/
[4]http://www.statmt.org/wmt18/

Architecture	Synthetic			Subtitle		Street
	En⇒Zh	En⇒De	Zh⇒En	En⇒Zh	Zh⇒En	Zh⇒En
Existing End-to-End TIMT Methods						
ItNet (Jain et al., 2021)	18.43	15.71	11.38	16.91	10.07	0.94
CLTIR (Chen et al., 2020c)	19.44	16.31	13.52	17.96	11.25	1.74
RTNet (Su et al., 2021)	19.63	16.78	14.01	18.82	11.50	1.93
MTETIMT(Ma et al., 2022)	21.96	18.84	15.62	19.17	12.11	5.84
MHCMM(Chen et al., 2022)	22.08	18.97	15.66	19.24	12.12	5.87
E2E MC-TIT (Lan et al., 2023)	22.17	19.21	15.74	19.28	12.14	5.95
PEIT (Zhu et al., 2023)	22.23	19.29	15.81	19.30	12.15	6.04
MTKD (Ma et al., 2023c)	22.26	19.38	15.84	19.31	12.17	6.08
E2TIMT (Ma et al., 2023b)	22.53	19.67	16.25	19.46	12.39	6.24
Our Proposed BabyNet with Various Parental Supervision (PS.) Granularities						
BabyNet w/ Local PS.	23.58	19.91	16.74	19.50	12.44	6.32
BabyNet w/ Global PS.	23.40	19.86	16.67	19.49	12.41	6.26
BabyNet w/ Fused PS.	**23.65**	**20.13**	**16.82**	**19.53**	**12.47**	**6.37**

Table 1: Comparison of end-to-end text image machine translation models.

million synthetic text line images and corresponding translation pairs for training. The evaluation sets have three translation directions: English-to-Chinese (EnZh), English-to-German (EnDe), and Chinese-to-English (ZhEn). Besides the synthetic evaluation domain, two real-world (subtitle and street-view) test sets are also utilized to evaluate the generalization of the models. For more details of this dataset please refer to Ma et al., (2022).

3.2. Experimental Settings

TIR parent models have four variants in our paper to evaluate the influence of different parent models: CRNN (Shi et al., 2017), TRBA (Baek et al.), TRT (Ma et al., 2023b), and TrOCR (Li et al., 2021). While the base and big MT models utilize the same architecture proposed in Vaswani et al., (2017). Parent models are firstly trained with TIR or MT datasets, respectively. Parameters of parent models are then frozen during the optimization on the TIMT dataset. For the child model, the architecture of BabyNet is a 6-layer transformer encoder with 8 attention heads, and the hidden dimension is set to 512. The batch size is 64, and the training step is set to 300,000. The maximum lengths for English, German, and Chinese sentences are set to 80, 80, and 40 respectively. Parameters of the BabyNet 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. The dropout rate is 0.1, and the initial learning rate is set to 2e-3. Sacre-BLEU (Papineni et al., 2002) [5] is utilized as the metric to evaluate the translation performance.

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

3.3. Compared End-to-End TIMT Models

We compare our method with existing end-to-end TIMT models: **ItNet** utilizes CNN based image encoder and transformer decoder for target language generation (Jain et al., 2021). **CLTIR** is multi-task trained with TIR task (Chen et al., 2020c). **RTNet** utilizes a feature transformer to bridge the TIR and MT modules (Su et al., 2021). **MTETIMT** is trained with MT auxiliary task (Ma et al., 2022). **MHCMM** is trained with hierarchical mimic learning (Chen et al., 2022). **E2E MC-TIT** (Lan et al., 2023) incorporates multimodal codebook to quantize image features into discrete code to further improve the cascade TIMT. To provide a comparison of the end-to-end methods, we reproduce the multimodal codebook based method with an end-to-end architecture. **PEIT** (Zhu et al., 2023) employs a two-stage pre-training strategy and alignment with an auxiliary MT task. **MTKD** is a multi-teacher knowledge distillation method (Ma et al., 2023c). **E2TIMT** is an parameter-efficient model with modal adapter (Ma et al., 2023b). To provide a fair comparison, all models are trained and evaluated with the same TIMT data set released by Ma et al., (2022).

3.4. Comparison with Existing End-to-End TIMT Methods

Table 1 shows the results of our method and existing end-to-end TIMT methods on the synthetic and two real-world (subtitle and street-view) evaluation sets. For the comparison with existing methods like E2TIMT (Ma et al., 2023b), our proposed BabyNet utilizes a trainable external bridge module to link the pre-trained OCR modules and

Architecture	Params.(↓) (Million)	Time(↓) (Second)	BLEU(↑) (%)
Cascade	195.1	0.33	20.46
End-to-End	121.9	0.19	18.02
Multi-Task	147.6	0.19	19.44
E2TIMT	13.2	0.19	22.53
BabyNet	13.2	0.19	23.65

Table 2: Comparison of model parameters and decoding time among various models on English-to-Chinese translation direction.

MT modules, while E2TIMT focuses on parameter-efficient tuning architecture like injecting adapter modules inside the pre-trained models. Furthermore, both local and global features of BabyNet are utilized during the calculation of loss functions, while E2TIMT only considers global features and lacks fine-grained knowledge supervision tailored for TIMT task, which limits the performance.

In all translation directions (En⇒Zh, En⇒De, and Zh⇒En), BabyNet with hierarchical parental supervision consistently outperforms the existing end-to-end TIMT methods with an average improvement of 0.72 BLEU scores on the synthetic domain across three translation directions. As for different granularities of parental supervision, local granularity outperforms global granularity, which we attribute to the fine-grained parental supervision signal providing more accurate guidance for the optimization of BabyNet. Furthermore, the fused parental supervision, which combines local and granularity guidance together, achieves the best results indicating that these two granularities are complementary for end-to-end TIMT task.

3.5. Comparison on Model Size and Decoding Time of TIMT Models

Table 2 shows the comparison of model parameters and decoding time. The cascade methods utilize separated TIR and MT models, which have parameter redundancy and a long decoding delay. The end-to-end method has fewer parameters and faster decoding. Multi-task learning based method incorporated external parameters for auxiliary tasks, while the decoding speed is fast due to the single-task inference during evaluation. E2TIMT is a parameter-efficient method, which just fine-tunes the modal adapter parameters. Similar to E2TIMT, our proposed method optimizes the inserted BabyNet between the TIR encoder and MT decoder, which just has 10.8% parameters of the end-to-end model to train. Meanwhile, the decoding speed of BabyNet is much faster than the cascade model with 41.1% less time. Furthermore, BabyNet has better translation performance than existing methods with an improvement of 1.12 BLEU score on the English-to-Chinese synthetic test set, which takes full advantage of cascade and end-to-end methods.

3.6. Analysis on Different Parent Model Combinations

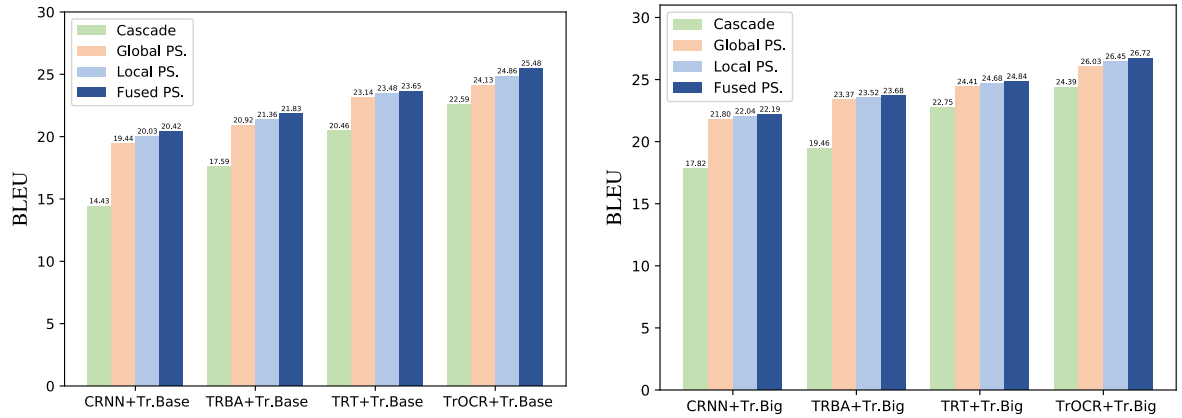
The BabyNet method can bridge various parent models, which is flexible in practical applications. To validate the generalization of BabyNet, Figure 4 shows the experimental results of different parent model recombinations. Four TIR models are evaluated in this paper: CRNN (Shi et al., 2017), TRBA (Baek et al.), TRT (Ma et al., 2023b) and TrOCR (Li et al., 2021). While two MT models are utilized: transformer-base and transformer-big as in Vaswani et al., (2017).

Figure 4 (a) shows the results of various TIR models and transformer-base combinations, while (b) shows the results of corresponding TIR models with transformer-big MT model. The cascade model in each group denotes the corresponding TIR and MT pipeline structure. BabyNet inherits the corresponding TIR encoder and MT decoder in each comparison group. Various granularity settings are also compared to illustrate the effectiveness of parental supervision.

From this comparison, BabyNet with different TIR and MT combinations consistently outperforms the corresponding cascade model, indicating the good generalization of BabyNet. Meanwhile, with a better TIR image encoder (like the TrOCR encoder), BabyNet achieves better performance through the inheritance of stronger image encoding capacity. Furthermore, models in Figure 4 (b) all outperform corresponding models in Figure 4 (a), indicating stronger transformer-big MT decoder further enhances the BabyNet translation performance. Additionally, the local granularity is better than the global granularity and fused granularity achieves the best results, revealing that multi-granularity parental supervision is vital for BabyNet optimization.

3.7. Ablation Study

To verify the performance gain of parental supervision at different layers, ablation studies were conducted on different combinations of loss functions. As shown in Table 3, Row 1 represents the BabyNet trained without any parental supervision. Rows 2, 3, and 5 show the single-layer parental supervision and parental supervision at the image encoder layer achieves the best performance with an improvement of 3.06 BLEU compared with Row 1. This is because aligning the features of text image and machine translation at lower layers facilitates



(a) Combinations of Various TIR Models and Transformer Base (b) Combinations of Various TIR Models and Transformer Big

Figure 4: Results of Different Parent Model Combinations. 'PS.' is the abbreviation of Parental Supervision. Tr.Base and Tr.Big represent Transformer Base and Transformer Big as in (Vaswani et al., 2017), respectively

	$\mathcal{L}_{\text{PS}}^{\text{Emb}}$	$\mathcal{L}_{\text{PS}}^{\text{Seq}}$	$\mathcal{L}_{\text{PS}}^{\text{Dec}}$	Local	Global	Fused
1	0	0	0	-	-	22.63
2	0	0	1	24.81	23.96	24.98
3	0	1	0	25.04	24.14	25.16
4	0	1	1	25.45	24.52	25.71
5	1	0	0	25.22	24.39	25.69
6	1	0	1	25.63	24.71	25.82
7	1	1	0	25.70	25.02	25.98
8	1	1	1	**25.74**	**25.28**	**26.06**

Table 3: Results of Ablation Study on English-to-Chinese Synthetic Validation Set.

better consistency with the pre-trained MT model in the subsequent processes.

For bi-layer combinations (Row 4, 6, and 7), it can be observed that combining bi-layer parental supervision outperforms single-layer guidance with 0.56 BLEU. Meanwhile, the fusion of parental supervision at the embedding and sequential encoding layer achieves the best performance for bi-layer settings. The eighth row shows the results of incorporating parental supervision at all three layers and it achieves the best performance compared with other layer combinations. Furthermore, the comparison between column local, global, and fused granularities shows that local granularity outperforms global guidance and various granularities are complementary.

3.8. Analysis on Hyper-Parameter

Hyper-parameter λ_{PS} in Eq. (13) is utilized to control the optimization weight of end-to-end TIMT and parental supervision loss functions. The constraint between λ_{TIMT} and λ_{PS} is: $\lambda_{\text{TIMT}} + \lambda_{\text{PS}} = 1$. As shown in Figure 5, when $\lambda_{\text{PS}} = 0$, the BabyNet is only optimized with $\mathcal{L}_{\text{TIMT}}$ and the performance is limited due to the lack of guidance from parent models. As the weight of the parental supervision loss increases, more constraints from parental features are incorporated, leading to improved translation performance of the model. When $\lambda_{\text{PS}} = 0.8$, the model achieves the best translation performance on the English-to-Chinese validation set with an improvement of 3.43 BLEU. When $\lambda_{\text{PS}} = 1.0$ and $\lambda_{\text{TIMT}} = 0.0$, the performance slightly decreases, indicating end-to-end TIMT loss is also useful to provide direct translation knowledge guidance. The hyper-parameter analysis demonstrates that the TIMT loss and parental supervision loss can provide complementary information and the joint optimization achieves the best results.

4. Related Work

To translate source language text images, in-image machine translation aims at generating images containing target translation results (Mansimov et al., 2020; Hinami et al., 2021), while text image machine translation (TIMT) is designed to decode target language texts (Chen et al., 2020c). The research of TIMT is highly related to our work and it can be mainly categorized into two types: cascaded models and end-to-end models.

Cascaded models combine pre-trained text image recognition models (Baek et al.; Li et al., 2021; Zhang et al., 2021, 2019, 2018) with machine translation models (Vaswani et al., 2017; Zhao et al., 2023, 2020) to translate text images in the source language (Afli and Way, 2016; Chen et al., 2015; Du et al., 2011; Wong et al., 2011; Chang et al., 2009; Yang et al., 2002). MC-TIT (Lan

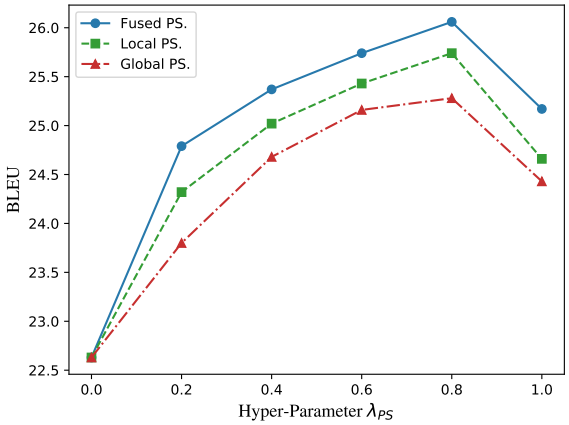


Figure 5: Comparison Experiment Results of Hyper-parameter λ_{PS}.

et al., 2023) incorporates a multimodal codebook to quantize image features into discrete code to further improve the cascade TIMT.

End-to-end models directly translate the input source language text images into target language sentences. To incorporate external TIR or MT datasets, multi-task learning based methods are proposed to improve TIMT performance (Chen et al., 2020c; Ma et al., 2022). While multi-teacher knowledge distillation (Ma et al., 2023c) and cross-modal mimic learning (Chen et al., 2022) are studied to transfer knowledge into end-to-end TIMT models. PEIT (Zhu et al., 2023) employs a two-stage pre-training strategy and alignment with an auxiliary MT task. Modal contrastive learning (Ma et al., 2023a) is proposed to align text and image representation while quantized feature shows better performance of text image representation learning (Ma et al., 2024). Another solution is inspired by parameter-efficient tuning (Zaken et al., 2022; Sun et al., 2021; Rothe et al., 2020; Le et al., 2021; Rebuffi et al., 2017), which takes advantage of pre-trained TIR or MT models with an external trainable feature transformation (Su et al., 2021) or modal adapter (Ma et al., 2023b). Different from existing research, our work aims at incorporating multi-granularity hierarchical parental supervision to guide the training of parameter-efficient BabyNet, which achieves new state-of-the-art and has a good generalization for various parent combinations.

5. Conclusion

In this paper, we propose a novel BabyNet optimized with hierarchical parental supervision for end-to-end TIMT. BabyNet bridges the pre-trained TIR and MT parent modules and effectively improves the translation performance. The Analyses on supervision granularity show local parental supervision outperforms global granularity due to the more accurate fine-grained knowledge guidance. Meanwhile, lower layer alignment with parental supervision has better consistency in the decoding process, which achieves better translation performance. Additionally, BabyNet achieves significant improvements with various combinations of parent model structures, indicating the good generalization of our method. In the future, we will explore more parental supervision functions to further improve translation performance.

6. Acknowledgement

This work has been supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62106265.

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.

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*.

Yi Chang, Datong Chen, Ying Zhang, and Jie Yang. 2009. An image-based automatic arabic translation system. *Pattern Recognit.*, 42(9):2127–2134.

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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020a. A simple frame-

work for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 1597–1607.

Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. 2020b. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297.

Zhuo Chen, Fei Yin, Qing Yang, and Cheng-Lin Liu. 2022. Cross-lingual text image recognition via multi-hierarchy cross-modal mimic. *IEEE Transactions on Multimedia (TMM)*, pages 1–13.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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.

Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. 2016. Synthetic data for text localisation in natural images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2315–2324. IEEE Computer Society.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, pages 297–304.

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)*.

Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Synthetic data and artificial neural networks for natural scene text recognition. *CoRR abs/1406.2227*.

Puneet Jain, Orhan Firat, Qi Ge, and Sihang Liang. 2021. Image translation network.

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)*.

Zhibin Lan, Jiawei Yu, Xiang Li, Wen Zhang, Jian Luan, Bin Wang, Degen Huang, and Jinsong Su. 2023. Exploring better text image translation with multimodal codebook. In *ACL*, pages 3479–3491.

Hang Le, Juan Miguel Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers),*

Virtual Event, August 1-6, 2021, pages 817–824. Association for Computational Linguistics.

Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. Trocr: Transformer-based optical character recognition with pre-trained models. *CoRR*, abs/2109.10282.

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 *26th International Conference on Pattern Recognition (ICPR)*.

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 *17th International Conference on Document Analysis and Recognition (ICDAR)*.

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 *17th International Conference on Document Analysis and Recognition (ICDAR)*.

Cong Ma, Yaping Zhang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2024. Vector quantization knowledge transfer for end-to-end text image machine translation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12471–12475.

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.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 506–516.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Trans. Assoc. Comput. Linguistics*, 8:264–280.

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.

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.

Zewei Sun, Mingxuan Wang, and Lei Li. 2021. Multilingual translation via grafting pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2735–2747. Association for Computational Linguistics.

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.

Fai Wong, Sam Chao, and Wai Kit Chan. 2011. Cyclops - snapshot translation system based on mobile device. *J. Softw.*, 6(9):1664–1671.

Jie Yang, Xilin Chen, Jing Zhang, Ying Zhang, and Alex Waibel. 2002. Automatic detection and translation of text from natural scenes. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2002, May 13-17 2002, Orlando, Florida, USA*, pages 2101–2104.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1–9. Association for Computational Linguistics.

Yaping Zhang, Shan Liang, Shuai Nie, Wenju Liu, and Shouye Peng. 2018. Robust offline handwritten character recognition through exploring writer-independent features under the guidance of printed data. *Pattern Recognit. Lett.*, 106:20–26.

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, Lu Xiang, Junnan Zhu, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2020. Knowledge graph enhanced neural machine translation via multi-task learning on sub-entity granularity. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4495–4505. International Committee on Computational Linguistics.

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

Shaolin Zhu, Shangjie Li, Yikun Lei, and Deyi Xiong. 2023. PEIT: bridging the modality gap with pre-trained models for end-to-end image translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13433–13447. Association for Computational Linguistics.