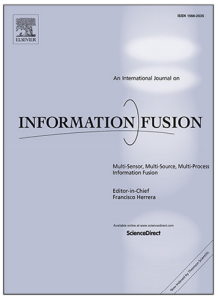
Journal Pre-proof



Multimodality information fusion for automated machine translation

Lin Li, Turghun Tayir, Yifeng Han, Xiaohui Tao, Juan D. Velásquez

PII: S1566-2535(22)00187-7
DOI: https://doi.org/10.1016/j.inffus.2022.10.018
Reference: INFFUS 1646

To appear in: *Information Fusion*

Received date : 5 April 2022
Revised date : 15 October 2022
Accepted date : 18 October 2022

Please cite this article as: L. Li, T. Tayir, Y. Han et al., Multimodality information fusion for automated machine translation, *Information Fusion* (2022), doi: https://doi.org/10.1016/j.inffus.2022.10.018.

Multimodality Information Fusion for Automated Machine Translation

Lin Li[a,*], Turghun Tayir[a], Yifeng Han[a], Xiaohui Tao[b], Juan D. Velásquez[c,d]

[a]*School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China*
[b]*School of Sciences, University of Southern Queensland, Toowoomb, Australia*
[c]*Instituto Sistemas Complejos de Ingenieria, ISCI*
[d]*Department of Industrial Engineering, Instituto Sistemas Complejos de Ingenieria, University of Chile, Santiago, Chile*

Abstract

Machine translation is a popular automation approach for translating texts between different languages. Although traditionally it has a strong focus on natural language, images can potentially provide an additional source of information in machine translation. However, there are presently two challenges: (i) the lack of an effective fusion method to handle the triangular-mapping function between image, text, and semantic knowledge; and (ii) the accessibility of large-scale parallel corpus to train a model for generating accurate machine translations. To address these challenges, this work proposes an effective multimodality information fusion method for automated machine translation based on semi-supervised learning. The method fuses multimodality information, texts and images to deliver automated machine translation. Specifically, our objective fuses multimodalities with alignment in a multimodal attention network, which advances the method through the power of mapping text and image features to their semantic information with accuracy. Moreover, a semi-supervised learning method is utilised for its capability in using a small number of parallel corpus for supervised training on the basis of unsupervised training. Conducted on the Multi30k dataset, the experimental results shows the promising performance of our proposed fusion method compared with state-of-the-art approaches.

*Corresponding author
Email addresses: cathylilin@whut.edu.cn (Lin Li), hotpes@whut.edu.cn (Turghun Tayir), tahanyifeng1110@163.com (Yifeng Han), xiaohui.tao@usq.edu.au (Xiaohui Tao), jvelasqu@dii.uchile.cl (Juan D. Velásquez)

Key words: Multimodal fusion, Machine translation, Multimodal alignment, Semi-supervised learning

1. Introduction

Multimodal machine learning is a vibrant field that aims to build models that process and correlate information from different modalities. In recent years, many research tasks based on multimodal information have been implemented, including multimodal machine translation (MMT) [1–3], cross-modal generation [4–6], visual dialogue [7–9], etc. Since the inputs and outputs of these models consist of different modal forms, these tasks deal with multimodal data. More recent multimodal tasks [1, 2, 4, 8] have produced state-of-the-art results based on Transformer structures [10], especially in MMT.

Image-based MMT is one of the mainstream multimodal machine learning approaches, aiming to translate one language into another language with semantic consistency via machine computation that uses images in the process. With image-based MMT, images and text are used as input for the translation model (as shown in Fig. 1) whereby images provide supplementary semantic information and disambiguation for the textual data, to improve the accuracy of machine translation [11–13]. The initial approach for adding images to a machine translation model is by concatenating the image and text features [2, 14]. However, since image and text belong to different modalities, direct concatenation overlooks the "one-to-many" and "many-to-one" correspondent relationships between text and image. Some studies [1, 15, 16] align image and text features via attention structure [17] to achieve remarkable results. In particular, some recent works [1, 18, 19] based on Transformer architecture have been outstanding. Our literature review reveals that models with different training parameters often lead to different translation results [13, 20, 21]. This situation is more pronounced with ensemble learning, as its quality depends on the accuracy and diversity of sub-models [22, 23]. Therefore, in addition to the accuracy of multimodal alignment, the diversity of translation models and their positive impact on machine translation needs to be considered.

Many studies also show that supervised machine translation (SMT) [24–26] has

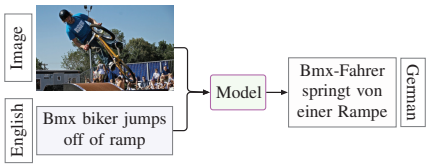


Figure 1: An example for image-based MMT.

become the mainstream direction for enhancing machine translation quality. Compared to statistical machine translation [27, 28], recent developments in deep learning has empowered machine translation approaches to become more popular and competitive [10, 17]. However, SMT model training requires large-scale, high-quality parallel corpora, and the acquisition of such corpora requires a major investment of human and material resources. In contrast, unsupervised machine translation (UMT) [29–31] can be trained without parallel corpus. Although UMT effectively solves the dependence of translation models on large-scale, high-quality parallel corpora, its translation quality is poor. The main solution to this problem is to leverage additional information, such as a large number of monolingual pre-training corpus [19, 30] or visual pivots [31, 32], where pre-training corpora is more effective. However, this approach not only increases material resources and training costs, but may also introduce low-quality pseudo-sentence pairs in model training, which hurts the performance of the machine translation model. Regardless of using supervised models or unsupervised models, both approaches are limited to a large-scale corpus. To this end, our study leverages a small amount of high-quality parallel corpus to establish a semi-supervised multimodal machine translation (Semi-MMT).

We redefine the training of the Semi-MMT model, which consists of supervised training and unsupervised training. In the unsupervised training component, a denoising auto-encoder [33] is applied to a given monolingual, multimodal corpus to reconstruct the source and target languages. Then, regularization pushes the latent encoding spaces aligned between the source and target language by constructing pseudo-pairs. For supervised training, a small parallel corpus is leveraged to further enhance this latent alignment to train the model and improve machine translation quality.

In terms of modal fusion, the text features from the encoder and the image features from the image feature extraction module are fused in a joint semantic space via weight learning. Through this processing, the multimodal feature weights and biases are continuously converged towards making the translation results more accurate during the training process. Finally, through integrating the prediction results of different parameter models, the diversity of models is more effectively utilized, which improves the translation performance.

The main contributions of this work are as follows:

- On the basis of unsupervised training, we leverage a small number of parallel corpora for further model training, thus forming Semi-MMT.
- Through a design of multi-perspective multimodal ensemble learning, we fuse the results of sub-models, effectively exploiting the benefits of our multimodal fusion method.
- The experimental results on the multimodal dataset Multi30k show that our model achieves a remarkable performance compared to the baseline model, using tens of millions of large-scale pre-training monolingual corpora.

The remainder of this paper is structured as follows. Section 2 introduces the recent related work of machine translation. Then, section 3 describes task definition before section 4 demonstrates the structure of the proposed model and its training paths. Section 5 presents the experiment and its results to verify the effectiveness of our approach. Section 6 summarizes our work and gives an outlook for future directions.

2. Related work

With the development of deep learning, many researchers apply various network structures to different natural language processing tasks, e.g., convolutional neural network (CNN) [34, 35], and recurrent neural network (RNN) for text based machine learning [36, 37], and deep recurrent belief networks for decision-making tasks [38]. In multimodal machine learning, since the data belongs to different modalities, bridging the gap between the different modalities is one of the major difficulties in this

task. Recently, several researchers proposed capsule networks as a potential solution to effectively address the problem of insufficient semantic interaction between modalities[39, 40]. Among them, work by [39] applied capsule networks on the Transformer and achieved state-of-the-art results in MMT.

2.1. Monomodal machine translation

SMT. Machine translation is a sequence-to-sequence learning task, which is typically implemented by an encoder-decoder structure [25]. In this structure, the encoder maps the source sentence into a distributed representation, which is then fed into the decoder to generate the target sentence word by word. With the advancement of deep learning, different neural networks have been used as encoder-decoder structure, such as RNN [17, 41], CNN [26] and Transformer [10]. Recently, Transformer has advanced the field of machine translation further than CNN and RNN in terms of translation quality and speed of convergence [10, 18], thus becoming the mainstream machine translation framework [1, 2, 18, 20]. Therefore, this study also uses Transformer as the main model to inherit this existing advanced technology.

UMT. Considering the machine translation of low-resource language pairs, some existing works [42, 43] have verified that UMT is a feasible solution. These works generally use modifications of the encoder-decoder schema and build a common latent space between two languages, learning translation by reconstruction in both domains. In terms of training data, they use a large amount of data to pre-train models or word embeddings. Although the above-mentioned studies do not leverage parallel corpus, they still rely on a large number of pre-trained embeddings. Thus, our approach explores the feasibility of reducing reliance on large-scale corpora and improving translation quality by using a small number of parallel corpora.

2.2. Semi-supervised machine translation

Since machine translation relies heavily on large-scale parallel bilingual corpora and only using monolingual corpora leads to the decline of translation quality, semi-supervised machine translation has attracted intensive attention. It is generally trained

in the following three ways: (1) source-to-target and target-to-source translation models are jointly trained by reconstructing the observed monolingual corpus using an auto-encoder [44]; (2) fusion between translation system and language model [45]; and (3) generating pseudo sentence pairs from monolingual corpora [46]. Compared with our model, the above works are performed on text-only data, with the parallel corpus accounting for at least 20% of the entire training data, much larger than our 7%. Our approach uses an image as a pivot to bridge the two unpaired languages, and with or without parallel corpus, image also plays the role of supplementary information to disambiguate. To the best of our knowledge this is a new Semi-MMT approach, but our task still draws on existing text-based ideas, such as reconstructing corpus and generating pseudo-sentences.

2.3. Multimodal machine translation

SMT. After several recent tasks [3, 11, 47] proposed image-based MMT, it has attracted extensive attention from researchers. Recent works such as [1, 2, 48, 49] have achieved remarkable translation results by feeding additional image information into text-only models. To introduce image information into the translation model, the image information is initially used as a part of the input sentence [47, 50] or used to initialize the encoder and decoder hidden states [51]. Subsequently, researchers use the attention mechanism to align and mine image information [11, 18, 21, 52]. The most recent state-of-the-art models [1, 2, 18, 48] are basically built on the Transformer, which provides the theoretical knowledge that support our decision to choose the Transformer as our main mode.

However, through the analysis of the MMT dataset and existing experimental results, we discover that text still plays a leading role in MMT, while the image is additional modal information. Treating text and image equally may encode too much irrelevant information from the image [1, 48]. This paper validates the importance of modal alignment methods by exploiting different multimodal alignment methods.

UMT. Recently, introducing image information into UMT has attracted widespread attention [19, 30, 31]. A problem in UMT is the lack of a target language that corresponds to the source language, therefore an additional image modality is introduced

Table 1: List of some abbreviations

Abbreviation	Expansion	Abbreviation	Expansion
$\mathbf{x}$	source sentence	$\widetilde{\mathbf{x}}$	translation of $\mathbf{x}$
$\mathbf{y}$	target sentence	$\widetilde{\mathbf{y}}$	translation of $\mathbf{y}$
$\mathbf{z}$	image	t	model text input
$\mathbf{z}_{\mathbf{x}}$	image corresponding to $\mathbf{x}$	T	text encoding
$\mathbf{z}_{\mathbf{y}}$	image corresponding to $\mathbf{y}$	I	image encoding
$\widehat{\mathbf{x}}$	reconstruction of $\mathbf{x}$	$\widetilde{y}$	model output
$\widehat{\mathbf{y}}$	reconstruction of $\mathbf{y}$	y	translation reference

as a pivot between unpaired bilinguals. Su et al. [19] investigated the possibility of using images for disambiguation and promoting the performance of UMT. Their hypothesis is intuitively based on the invariant property of the image, which means that the description of the same visual content in different languages should still be roughly similar. They achieve current state-of-the-art performance by exploring two training paths, such as auto-encoding loss and cycle-consistency loss. Thus, we use it as the basis for the unsupervised part of our model training. Existing unsupervised MMT models [19, 30] rely on large-scale corpora for model pre-training, which increases the cost of training. In addition, the alignment of multimodal features brings additional challenges to unsupervised MMT. Therefore, considering the above issues, this study uses a small amount of data to implement Semi-MMT with different alignments.

3. Task description

3.1. Task definition

As shown in Fig. 2, the training of the Semi-MMT model includes two parts: supervised and unsupervised. In unsupervised training, only monolingual corpora are used on both the source and the target language sides, and they come in the paired form of $(\mathbf{x}, \mathbf{z}_{\mathbf{x}}) \in \mathcal{X} \times \mathcal{Z}$ and $(\mathbf{y}, \mathbf{z}_{\mathbf{y}}) \in \mathcal{Y} \times \mathcal{Z}$. Therefore, the triple data of supervised learning is no longer available. Unsupervised training allows the model to generate a common latent space between the two languages and learns translation through re-

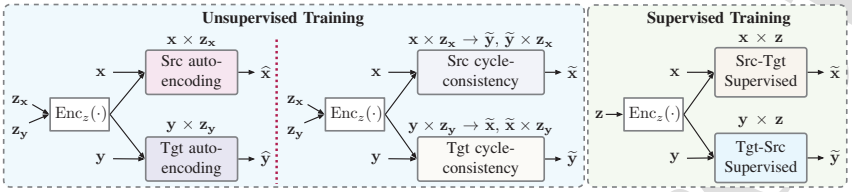


Figure 2: Illustration of the input and output of the proposed model. Src and Tgt are denoted as source and target. $\mathbf{x}$ and $\mathbf{y}$ denote source and target sentences, $\mathbf{z_x}$ and $\mathbf{z_y}$ represent images corresponding to $\mathbf{x}$ and $\mathbf{y}$, and $\mathrm{Enc}_z(\cdot)$ represents image encoding. In unsupervised training, $\mathbf{x}$ and $\mathbf{y}$ are non-parallel sentences. In supervised training, $\mathbf{x}$ and $\mathbf{y}$ are parallel sentences, while $\mathbf{z}$ represents their corresponding image. $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{y}}$ represent the reconstruction of source and target sentences, $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{y}}$ represent the translation of the target and the source sentences.

construction in the two languages. This part is composed of auto-encoding loss and cycle-consistency loss. In auto-encoding loss, the two mapping relations $\mathcal{X} \times \mathcal{Z} \rightarrow \mathcal{X}$ and $\mathcal{Y} \times \mathcal{Z} \rightarrow \mathcal{Y}$ are learned by reconstructing on the source and target languages, while in cycle-consistency loss, regularization pushes the latent encoding spaces aligned between the source and the target language and learns the mappings $\mathcal{Y} \times \mathcal{Z} \rightarrow \mathcal{X}$ and $\mathcal{X} \times \mathcal{Z} \rightarrow \mathcal{Y}$. In supervised training, an image and its descriptions in two different languages form a triplet $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in (\mathcal{X}, \mathcal{Y}, \mathcal{Z})$, and supervised-loss is utilized to learn the mapping relation $\mathcal{X} \times \mathcal{Z} \rightarrow \mathcal{Y}$ and $\mathcal{Y} \times \mathcal{Z} \rightarrow \mathcal{X}$. Some abbreviations and their extended meanings are shown in Table 1.

3.2. Unsupervised MMT

As shown in the unsupervised training part of Fig. 2, since UMT utilizes monolingual corpus, the model has no input for translation reference. On both the source and target languages, only non-overlapping monolingual multimodal data is provided for training, and the available data is $(\mathbf{x}, \mathbf{z_x}) \in \mathcal{X} \times \mathcal{Z}$ and $(\mathbf{y}, \mathbf{z_y}) \in \mathcal{Y} \times \mathcal{Z}$, where $\{\mathbf{x}\} \cap \{\mathbf{y}\} = \phi$. Since there is no clear pairing information across the source and the target languages, it is impossible to directly optimize the supervised likelihood. Although each language has different expressions, languages with the same meaning are similar in the latent space. In addition, visual content is used to disambiguate seman-

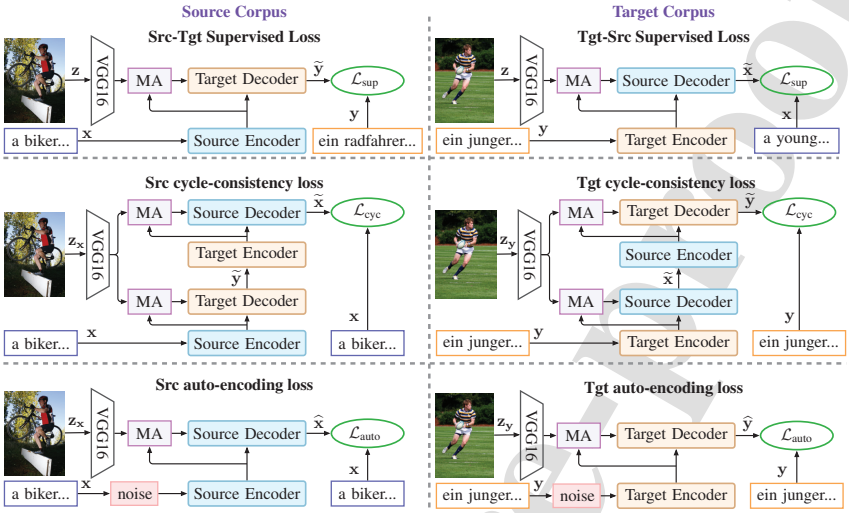


Figure 3: The overall training path of the Semi-MMT. MA represents a multimodal attention. Unsupervised training is implemented via src/tgt auto-encoding loss and src/tgt cycle-consistency loss, while supervised training is implemented via src-tgt/tgt-src supervised-loss. The encoder and decoder in the figure are represented as the encoder and decoder of the Transformer.

tics and promote latent space alignment, thus improving machine translation performance [19, 30]. The unsupervised part of our model is based on the UMNMT [19]. In this part, auto-encoding loss is used to reconstruct the source and target languages and then cycle-consistency loss is used to learn the mapping relations $\mathcal{X} \times \mathcal{Z} \rightarrow \mathcal{Y}$ and $\mathcal{Y} \times \mathcal{Z} \rightarrow \mathcal{X}$.

3.3. Supervised MMT

As shown in the supervised training part of Fig. 2, the MMT model utilizes image information, making up for the shortcomings of text-only translation and improving its performance. In this task, the image $\mathbf{z}$ and the two different language word sequences $\mathbf{x} = (x_1, ..., x_n)$ and $\mathbf{y} = (y_1, ..., y_m)$ describing $\mathbf{z}$ form a triple $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in (\mathcal{X}, \mathcal{Y}, \mathcal{Z})$. In the Transformer model, the encoder transforms the source language word sequence $\mathbf{x}$ into a hidden representation $\{\mathbf{h}_1^e, ..., \mathbf{h}_n^e\} = \mathrm{Enc}_x(\mathbf{x})$. Similarly, the image is encoded as $\{\mathbf{h}_1^i, ..., \mathbf{h}_k^i\} = \mathrm{Enc}_z(\mathbf{z})$, where k represents the number of image features. In

the decoder, at the time stamp t the encoder-decoder attention mechanism calculates a context vector $\mathbf{c}_i^t = \sum_{j=1}^{n} \alpha_i \mathbf{h}_j^e$ via a attention-based alignment $\{\alpha_1, ..., \alpha_n\} = \mathrm{Align}(\mathbf{h}_t^y, \{\mathbf{h}_1^i, ..., \mathbf{h}_n^i\})$, where $\sum_{j=1}^{n} \alpha_j = 1$ and $\mathbf{h}_t^y$ represents the decoder state. In this paper, *k* and *n* are equal, $\{\mathbf{h}_1^{ei}, ..., \mathbf{h}_n^{ei}\}$ represents the linear weighted sum of the corresponding items of $\{\mathbf{h}_1^i, ..., \mathbf{h}_k^i\}$ and $\{\mathbf{h}_1^e, ..., \mathbf{h}_n^e\}$. The context vector of our multimodal attention $\mathbf{c}_t$ is as follows:

$$\mathbf{c}_t = \mathrm{A}(\mathbf{h}_t^y, \{\mathbf{h}_1^{ei}, ..., \mathbf{h}_n^{ei}\}, \{\mathbf{h}_1^i, ..., \mathbf{h}_n^i\}) \tag{1}$$

The probability that the model predicts the next token in the decoder output can be written as:

$$p(y_t \mid \mathbf{y}_{<t}, \mathbf{x}, \mathbf{z}) = \mathrm{softmax}(g(\mathbf{c}_t, y_{t-1}, \mathbf{h}_{t-1}^y)) \tag{2}$$

where g(·) represents a non-linear function.

4. Our Semi-MMT Method

As shown in Fig. 3, the training of the proposed Semi-MMT method includes unsupervised and supervised parts. The unsupervised part includes four training paths, which are src/tgt auto-encoding loss (lower region of the Fig. 3) and src/tgt cycle-consistency loss (middle region of Fig. 3). The supervised part includes two training paths, which are src-tgt/tgt-src supervised-loss (top region of Fig. 3). These training paths are composed of a multimodal encoder-decoder structure, a multimodal attention structure and an image feature extraction module. A detailed structure is shown Fig. 4.

4.1. Multimodal encoder-decoder structure

As shown in the encoder-decoder part of Fig. 4, the multimodal encoder-decoder structure is built on the basis of Transformer [10]. For clarity, the layer normalization [53] of encoder and decoder is omitted. The input and output embeddings are trainable on the source and target sides, but are not pre-trained. Text and image features are received and fused by introducing additional multimodal attention between the encoder-decoder attention mechanism and the feed-forward sub-layer. The encoder converts the input source sentence into a vector with semantic information, and the

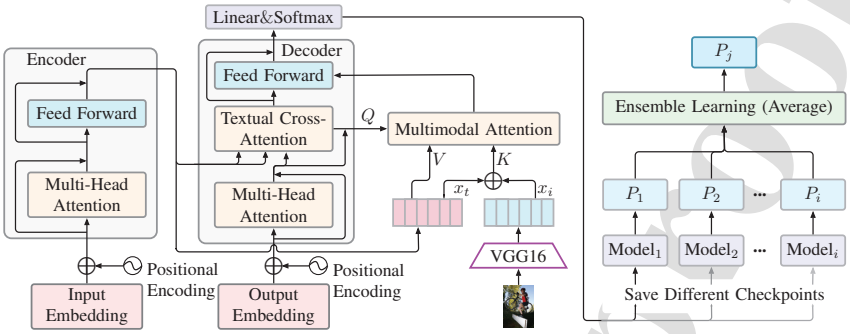


Figure 4: The detailed structure of the Semi-MMT. It includes a multimodal encoder-decoder structure, multimodal attention structure, image feature extraction module and multi-perspective fusion. This figure shows a part of the proposed model. The following takes the source language corpus as an example to introduce the figure: (1) If this figure is regarded as an auto-encoding loss (lower region of the Fig. 3) path, the encoder and decoder in the figure are in the source language; (2) If it is regarded as a cycle-consistency loss (middle region of Fig. 3) path, the source encoder to target decoder and target encoder to source decoder are applied jointly; (3) If it is regarded as a supervised-loss (top region of Fig. 3) path, the encoder and decoder are in the source and target languages. The target corpus training is also similar, but the encoder and decoder are on different corpora. The output of the decoder on the left of the figure represents multimodal ensemble learning for multi-perspective fusion. Layer normalization of Transformer is omitted for clarity.

decoder generates the target sentence according to the semantic information. Since the Transformer model does not use the display of sequence order, the model needs to encode the position of words to determine the positional relationship of different words in the sequence.

The encoder-decoder structure in the six training paths is designed as follows. In auto-encoding loss, as shown in the lower region of Fig. 3, the encoder and decoder of the same language and its corresponding images are applied once. In cycle-consistency loss, as shown in the middle region of Fig. 3, the encoder and decoder of the two languages and their corresponding images are cross-applied twice. Both the first two encoders and decoders are locked, and there is no parameter update. In supervised-loss, as shown in the top region of Fig. 3, the encoder and decoder of the source and target languages are cross-applied once. During the training process, the encoder and

decoder of the source and target languages share the parameters of the first three layers, respectively.

4.2. Image feature extraction module

As shown in Fig. 4, following prior work [47], a 4096 dimensional vector of the second fully connected layer of VGG16 [54] pre-trained on ImageNet [55] is used as the image feature. Then, it is averaged at every 8 dimensions and a 512 dimension image feature vector is obtained. This vector represents rough image information and the semantic value is greatly reduced compared to the original image. Finally, to correspond to the length of 50 sentences, it is self-replicated 50 times to obtain 50×512 dimensional image features. Through these operations, the semantic content of the image is greatly increased. Hereinafter, we call it one-dimensional (1D) global image features.

4.3. Multimodal alignment for modality fusion

In multimodal tasks, modality alignment is defined as finding relationships and correspondence between sub-components of instances from two or more modalities [56]. The attention mechanism allows the model to learn the alignment between different modalities, such as image and text [57]. For MMT, the alignment of the multimodal features is the most important factor affecting the results. Therefore, we introduce three different multimodal feature alignment methods: encoder and decoder gate (*Gate*) [58], double attentive (*Atten*) [18] and *IVTA* [59]. Furthermore, considering the impact of different multimodal alignment methods on ensemble learning, we only fuse different sub-models of these three models under the same model, but not all three models together.

4.3.1. Gate

The Gate structure is added at both ends of the encoder and decoder of the Transformer and is used to introduce image features into the encoder and decoder. On the decoder side, first its output s_j is mapped to an unnormalized distribution over the target vocabulary $y_j = W \cdot s_j + b$ and then a gating layer is added.

$$g_j = \sigma\left(U_{gate}^{dec} \cdot s_j + W_{gate}^{dec} \cdot I + b_{gate}^{dec}\right) \tag{3}$$

$$y'_j = y_j \odot g_j \tag{4}$$

where U^{dec}_{gate} and W^{dec}_{gate} are two weights, and b^{dec}_{gate} is the bias, and σ represents the sigmoid function. The operation of inserting image features *I* into the encoder side is similar to that of the decoder, but they have different roles in the translation process. The gate structure on the end of the decoder is used to filter out entity semantics that do not exist in the image, while the gate structure on the encoder side is used to disambiguate the encoded source sentence.

4.3.2. Atten

The image information is introduced into the model via an additional visual cross-attention mechanism. This resultant visual cross-attention layer is inserted between the encoder-decoder attention mechanism and the feed-forward sub-layer. Since the image information is set as the key *K* and *V* values for visual cross-attention, the key and value matrix are equal (*K*=*V*). They directly introduce image information into the attention mechanism without any processing. Through the use of image information for disambiguation, they can thus model the denotation of words.

$$A(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{5}$$

4.3.3. IVTA

As shown in the Fig. 4, we use coordinated representation learning to perform a linear weighted transformation on text and image features and then add them to get *IVTA*, such as in Eq. (6).

$$IVTA = W_t \cdot T + b_t + W_i \cdot I + b_i \tag{6}$$

where W_t, b_t, W_i and b_i represent the weights and biases of the linear transformation of text and image features, respectively, and *T* and *I* represent text and image features themselves. The weights and biases are trainable and they continuously converge towards the accuracy of the translation results during training with the help of a loss function. After the text and image vectors are mapped onto a similar common semantic space, they satisfy the addition and subtraction operations. This means that

different modal features can be encoded in a similar semantic space via the multimodal alignment of Eq. (6).

Text and image information is received via a multimodal attention structure, which is inserted between text cross-attention and feed-forward layers, as shown in Fig. 4. Unlike *Atten*, the multimodal attention structure in *IVTA* receives the output of text cross-attention as Q, *IVTA* as K, and the encoder output as V, as shown in Eq. (7)

$$A = \text{softmax}\left(\frac{Q(IVTA)^T}{\sqrt{d}}\right)V \tag{7}$$

The modal fusion part on Fig. 3 and Fig. 4 refers to the IVTA alignment. Model training and multimodal multi-perspective fusion are performed separately for the three modality alignments.

4.4. Loss function

In this paper, $mean\_loss$ is selected as the loss function of the training. $Mean\_loss$ is constructed on the cross-entropy loss function. Let $\widetilde{y}$ denote the distribution of the machine translation model output and y denote the translation reference. The cross-entropy loss can then be defined as:

$$H(\widetilde{y}_j, y_j) = -\sum_{j=1}^{|J|} \widetilde{y}_j log(y_j) \tag{8}$$

where $\widetilde{y}_j$ and y_j represent the j-*th* dimension of the vectors $\widetilde{y}$, and y and $|J|$ represent the dimension of the output vector $\widetilde{y}$. The $mean\_loss$ can be written as follows:

$$mean\_loss = \frac{\sum_{j=1}^{L} H(\widetilde{y}_j, y_j)}{L} \tag{9}$$

where *L* represents the maximum length of the source sentence, which is set to 50.

4.5. Training path

The model training is conducted according to the steps shown in Algorithm 1, which includes unsupervised and supervised parts. Line 1 to line 4 corresponds to the input data for the two training parts. Lines 8 and 9 represent the encoding of image and text. Line 10 to 12 refers to decoding image and text fusion to obtain another

Algorithm 1 Training algorithm of the proposed model

Input: $t = \{t_1, ..., t_n\}$, $z = \{z_1, ..., z_n\}$ ▷ t: text, z: image and $(\mathbf{x}, \mathbf{y}) \in t$

Output: $\widetilde{y} = \{\widetilde{y_1}, ..., \widetilde{y_n}\}$ ▷ $\widetilde{y}$ is predicted sentence sequence

1: **if** $\mathcal{L}_{\text{auto}}$ **or** $\mathcal{L}_{\text{cyc}}$ **then**
2: Use monolingual data $(\mathbf{x}, \mathbf{z_x})$ and $(\mathbf{y}, \mathbf{z_y})$ ▷ $\{\mathbf{z_x}\} \cap \{\mathbf{z_y}\} = \phi$
3: **else** ▷ $\mathcal{L}_{\text{sup}}$
4: Use parallel data $(\mathbf{x}, \mathbf{y}, \mathbf{z})$
5: **repeat**
6: $i = 1$
7: **for** $i<N$ **do**
8: $T_i \leftarrow Enc(t_i)$ ▷ Use Section 4.1 for text encoding
9: $I_i \leftarrow VGG16(z_i)$ ▷ Use Section 4.2 for image encoding
10: Use Section 4.3.3 to fuse T_i and I_i
11: $\widetilde{y_i} \leftarrow Dec(T_i, I_i)$ ▷ Use Section 4.5 for decoding
12: $\mathcal{L} \leftarrow mean\_loss(\widetilde{y_i}, y_i)$ ▷ y_i is reference and use Section 4.4
13: Timely adjust $\mathcal{L}_{\text{auto}}$, $\mathcal{L}_{\text{cyc}}$ and $\mathcal{L}_{\text{sup}}$ in Section 4.5.
14: **end for**
15: **until** min($\mathcal{L}$)
16: **end if**

language sentence, and calculate the loss function between this language sentence and the reference sentence. Line 13 corresponds to the training path in this section, and these multiple training paths are adjusted during the training process. Line 15 indicates that the training is carried out until the minimum loss is obtained. The unsupervised part is based on the UMNMT model [19], which is trained via auto-encoding loss and cycle-consistency loss. Supervised training is conducted via supervised-loss.

Auto-encoding loss. In unsupervised training, we leverage the denoising auto-encoder to reconstruct the source and target languages. As shown in the lower region of Fig. 3, two denoising auto-encoding losses are constructed with noisy monolingual data $\mathbf{x}$, $\mathbf{y}$ and their corresponding images.

$$\mathrm{Dec}_x\left(\mathrm{Enc}_x(\mathbf{x}), \mathrm{Enc}_z(\mathbf{z_x})\right) = \widehat{\mathbf{x}} \tag{10}$$

$$\mathcal{L}_{\text{auto}}\left(\mathbf{x}, \mathbf{z}_{\mathbf{x}}\right) = mean\_loss\left(\widehat{\mathbf{x}}, \mathbf{x}\right) \tag{11}$$

where $\widehat{\mathbf{x}}$ represents the output of the decoder in auto-encoding loss, $\mathrm{Enc}_x(\cdot)$ and $\mathrm{Dec}_x(\cdot)$ represent the encoder and decoder of the source language, and $\mathrm{Enc}_z(\cdot)$ represents the image encoder VGG16. Similarly, we derive the auto-encoding loss of the target language part:

$$\mathrm{Dec}_y\left(\mathrm{Enc}_y(\mathbf{y}), \mathrm{Enc}_z(\mathbf{z}_{\mathbf{y}})\right) = \widehat{\mathbf{y}} \tag{12}$$

$$\mathcal{L}_{\text{auto}}\left(\mathbf{y}, \mathbf{z}_{\mathbf{y}}\right) = mean\_loss\left(\widehat{\mathbf{y}}, \mathbf{y}\right) \tag{13}$$

Cycle-consistency loss. The central idea of auto-encoding loss is to use auto-encoders to reconstruct monolingual corpora, in which source-to-source and target-to-target relationships are learned. However, these two relationships are not the desired result of this task, so we use cycle-consistency loss to regularize and push the latent encoding spaces between the source and the target language. From this, the expected mapping $\mathcal{X} \times \mathcal{Z} \rightarrow \mathcal{Y}$ is learned. As shown in the middle region of Fig. 3, in the cycle-consistency loss of the source language, we input the encoded source sentence and image into the decoder of the target language and infer the target sentence $\widetilde{\mathbf{y}}$.

$$\mathrm{Dec}_y\left(\mathrm{Enc}_x(\mathbf{x}), \mathrm{Enc}_z(\mathbf{z}_{\mathbf{x}})\right) = \widetilde{\mathbf{y}} \tag{14}$$

Since the source sentence $\mathbf{x}$ does not have its corresponding target sentence $\mathbf{y}$, the loss function cannot be used directly. The decoder output $\widetilde{\mathbf{y}}$ is a low-quality pseudo target sentence, in which the source-to-target translation models serve as the encoder and decoder. If the pseudo-sentence $\widetilde{\mathbf{y}}$ and the golden reference $\mathbf{x}$ continue to be the training data of the target-to-source model, then construct a pseudo-supervised triple ($\mathbf{x}$, $\widetilde{\mathbf{y}}$, $\mathbf{z}_{\mathbf{x}}$), which satisfies the parallel corpus for model training.

$$\mathrm{Dec}_x\left(\mathrm{Enc}_y(\widetilde{\mathbf{y}}), \mathrm{Enc}_z(\mathbf{z}_{\mathbf{x}})\right) = \widetilde{\mathbf{x}} \tag{15}$$

Finally, the outputs $\widetilde{\mathbf{x}}$ corresponding to the pseudo input $\widetilde{\mathbf{y}}$, and learns the source-to-target translation. In the source-to-target-to-source joint training process of the entire

cycle-consistency loss, the input $\mathbf{x}$ and the output $\widetilde{\mathbf{x}}$ are in the same language. The relationship in this whole process can be written as:

$$\mathrm{Dec}_x\left(\mathrm{Enc}_y\left(\mathrm{Dec}_y\left[\mathrm{Enc}_x(\mathbf{x}), \mathrm{Enc}_z(\mathbf{z})\right]\right), \mathrm{Enc}_z(\mathbf{z})\right) = \widetilde{\mathbf{x}} \tag{16}$$

With the help of the loss function, the intermediate "pseudo-sentence $\widetilde{\mathbf{y}}$" is also updated with a more correct trend during the training. From cycle-consistency loss function on the source language, the mapping $\mathcal{Y} \times \mathcal{Z} \rightarrow \mathcal{X}$ can be successfully refined.

$$\mathcal{L}_{\text{cyc}}(\mathbf{x}, \mathbf{z}_{\mathbf{x}}) = mean\_loss\left(\widetilde{\mathbf{x}}, \mathbf{x}\right) \tag{17}$$

Similarly, in the cycle-consistency loss of the target language, the mapping relationship $\mathcal{X} \times \mathcal{Z} \rightarrow \mathcal{Y}$ can be learned.

Supervised loss. In the unsupervised training part, the use of low-quality pseudo-sentences as model input and the large number of model layers make it difficult for the model to update to the optimal level, which seriously hurts the performance of the translation. Therefore, we continue to train the model with a small amount of parallel corpus, through which the model can be further updated and achieve better translation results. At this stage, an image $\mathbf{z}$ and the description of it in two languages form a triplet $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in (\mathcal{X}, \mathcal{Y}, \mathcal{Z})$, to directly train the model. As shown in the top region of Fig. 3, the supervised-loss of the source language is:

$$\mathrm{Dec}_y\left(\mathrm{Enc}_x(\mathbf{x}), \mathrm{Enc}_z(\mathbf{z})\right) = \widetilde{\mathbf{y}} \tag{18}$$

$$\mathcal{L}_{\text{sup}}(\mathbf{x}, \mathbf{z}, \mathbf{y}) = mean\_loss\left(\widetilde{\mathbf{y}}, \mathbf{y}\right) \tag{19}$$

Similarly, the mapping relationship $\mathcal{Y} \times \mathcal{Z} \rightarrow \mathcal{X}$ is learned in the supervised-loss of the target language.

4.6. Multimodal multi-perspective fusion

With the improvement of computer processing capabilities, machine translation adopts the end-to-end translation structure of a neural network [10, 24, 25] to implement the mapping from the source sequence to target sequence. However, since neural network training tends to fall into the local optimal solution, the final model training

Algorithm 2 Multimodal multi-perspective fusion algorithm

Input: sub-model probability $p1$, $p2$, ..., pj, maximum length of sentence l

Output: probability distribution p of the ensemble model

1: l=50; i = 1
2: **for** i<l **do**
3: p_s = 0; n = 1
4: **for** n<j **do**
5: $p[n][i]$ = sub_model[n](p[1], p[2], ..., p[i-1])
6: p_s += $p[n][i]$ ▷ j is the number of sub-models
7: n++
8: $p[i]$ = arg max(p_s / j) ▷ p_s is the probability vector
9: **end for**
10: i++
11: **end for**
12: **return** p

result may not be the global optimal solution [13, 60, 61]. To avoid this situation, ensemble learning is used to fuse several models with different parameters. Traditional direct ensemble learning [62] has achieved significant improvements by fusing multiple locally optimal solutions. However, there is a word order constraint in the text sequence of machine translation, and directly fusing translation results from different training setting is unsatisfactory.

To address the above challenges, in the later stage of training we save one checkpoint model at each epoch with better performance, as shown in Fig. 4, and regard the model as a sub-model. In the test process, multiple sub-models with different optimal solutions are fused using a multimodal multi-perspective fusion Algorithm 2.

For example, as shown in Fig. 5 and line 5 in Algorithm 2, the sentence “ein junger mann wirft einen football” is input into three different sub-models, and the current word “ein” outputs different translation results across these sub-models. These probability values are averaged according to line 8 in Algorithm 2, and the word corresponding to the maximum value of the averaged vector is regarded as the translation result of “ein”.

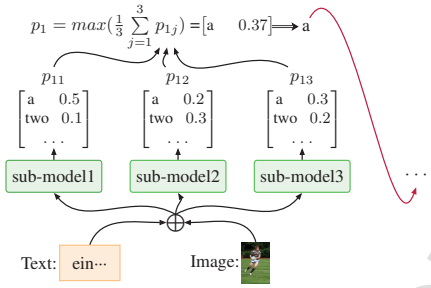


Figure 5: An example for multimodal and multi-perspective fusion based on the prediction results of the three models.

In machine translation, each current sequence depends on the result of the previous sequence. As shown in line 5 in Algorithm 2, the next word is predicted by inputting the prediction result of the current word into the decoder of a different sub-model, then repeating the above process until the end of the sentence. We added the symbol "+" after the model name to indicate that the model has used the multi-perspective fusion method.

5. Experiments

In this section, we firstly introduce the experimental dataset, pre-processing methods, experimental settings and baseline models. Secondly, the baseline model and the proposed model are tested, and the effectiveness of the method is verified by the experimental results. Thirdly, the multimodal multi-perspective fusion method is applied to the models with different parameters. Finally, experiments are conducted on models with parallel data of different sizes. Our translation experiments are all conducted on English $\Rightarrow$ German.

5.1. Dataset

This study conducts experiments on the Multik30k [63] dataset, which is an extended version of Flickr30k [64]. As shown in Fig. 6, the dataset contains 29,000 training and 1,014 validation images. For testing, we used the 2016 and 2017 test sets, each containing 1,000 images. Each image is paired with its English descriptions, as well

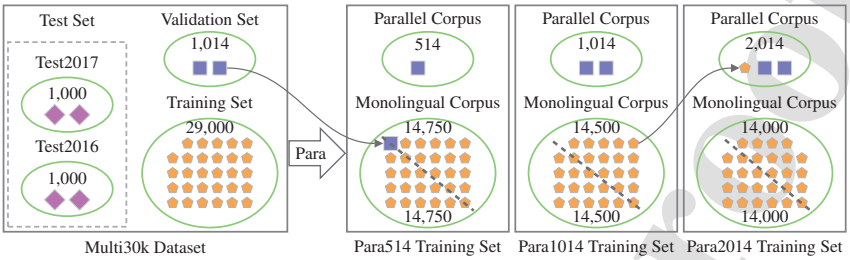


Figure 6: Multi30k multimodal data distribution and the process of reorganizing training set of Para data. To aid comprehension of data adjustment, a dotted line is used to indicate the random split of the paired of the two languages to form a monolingual corpus of the source and target languages.

as human translations of German. The training dataset Para is composed of monolingual corpus and parallel corpus, and the size of Para is adjusted in three ways, namely Para514, Para1014 and Para2014. For supervised training, **Para514** uses the 514 data in the Multi30k validation set as training parallel corpus, **Para1014** uses the entire validation set as training parallel corpus and **Para2014** takes the entire validation set and 1,000 Multi30k training sets as training parallel corpus.

In unsupervised training, to ensure that the model does not see any pairwise sentence, we follow the approaches in works such as [19, 30] that randomly split half of the training set into one language and the complementary half for the other language. After reorganizing a parallel corpus of each Para data, the remaining training data (500 validation instances not included in the Para 514) is randomly split for unsupervised training. The distribution of supervised and unsupervised training data is shown in Fig. 6.

5.2. Experimental setup

We implemented our experiments on a machine with a single 12GB TITAN Xp GPU. For text pre-processing, this paper follows the work of [3] and uses the hyper-parameters listed in Table 2 to train model. And then, label smoothing [65] is set to 0.1 and Adam optimizer [66] is applied for parameter optimization. In unsupervised training, injecting noise into the input data is a common trick, which generally includes

Table 2: List of some hyperparameters

Hyperparameter	Value
Word embedding dimension	512
Position encoding dimension	512
Feed forward layer dimension	2,048
Sentence length	50
Multi-head attention heads	8
Encoder/decoder layers	4
Dropout rate	0.1

disrupting the order of sentences, and dropping and/or shielding some words [42]. Similar to the work [42], we use $prob\_drop$ = 0.1 to indicate that each word has a 10% probability of being deleted and $word\_move\_dis$ = 2 to indicate that the displacement of each word does not exceed 2.

5.3. Evaluation indicator

For evaluating translation quality of our model, we use three different evaluation indicators: BLEU [67], METEOR [68] and ROUGE [69]. They focus on different parts of the sentence during the evaluation process. We define each in detail as follows.

BLEU. It is a document-level automatic indicator that calculates the geometric mean of the *n-gram* matching accuracy between the translation reference and the translation output. Generally, *n-gram* is defaulted to 4-*gram*. The calculation formula of BLEU score is as follows:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \cdot \log \frac{c_h}{c_o}\right) \tag{20}$$

where *N* represents the size of the largest *n-gram* considered, *BP* is the Brevity Penalty. w_n generally takes a constant value for all *n*, namely 1/*N*. c_h represents the number of hits of the *n-gram* in the translation reference and c_o represents the total number of

n-grams. The Brevity Penalty calculation formula can be written as follows:

$$BP = \begin{cases} 1 & c > r \\ \exp\left(1 - \frac{r}{c}\right) & c \leq r \end{cases} \tag{21}$$

where *c* represents the sentence length of the translation and *r* represents the sentence length of the translation reference.

METEOR. It combines unigram accuracy and recalls with internal alignment mechanisms between words in the translation reference and the translation output. It aims to solve some deficiencies in BLEU, such as the matching of synonyms. The calculation formula of METEOR is:

$$METEOR = (1 - Pen)F_{\text{mean}} \tag{22}$$

where the punishment mechanism is: $Pen = \gamma\left(c_c/c_h\right)^{\theta}$ and the score of machine translation is: $F_{mean} = PR/[\alpha P + (1-\alpha)R]$. The accuracy rate P is the ratio of the number of hits in the translation to the total number of words. The recall rate R is the ratio of the number of hit words in the translation to the total number of translation reference words. c_c represents the number of matching chunks, α, γ and θ are the default parameters for evaluation.

ROUGE. It is a commonly used evaluation metric for machine translation and article summaries. There are four ROUGE methods proposed in the original paper [69]. We utilize ROUGE-L, which takes into account the longest common sub-sequence between the reference and the translation result when calculating. The procedure for calculating it is as follows:

$$F_{lcs} = \frac{\left(\beta^2 + 1\right) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{23}$$

where $R_{lcs} = LCS(\widetilde{y}, y)/len(y)$, $R_{lcs} = LCS(\widetilde{y}, y)/len(\widetilde{y})$ and $\beta = P_{lcs}/R_{lcs}$. $\widetilde{y}$ is the model-generated answer and y is the reference answer. $LCS(\widetilde{y}, y)$ is to get the length of the longest common subsequence.

5.4. Baseline models

To verify the feasibility of the proposed method, we compare it with the existing works, such as UMNMT [19], 3Iter [42] and B+img [31].

- **UMNMT.** It is the basic model of the unsupervised part of our model, which is built through training paths such as auto-encoding loss and cycle-consistency loss. We take its four experiments as the baseline model, namely S-txt, S-txt-img, P-txt and P-txt-img. **S-txt:** Only half of the text training set of Multi30k is used for training. **S-txt-img:** Half of the text and image training set of Multi30k is used for training. **P-txt:** More than 10 million text-only monolingual corpus are used for pre-training and then half of the text training set of Multi30k are used for fine-tuning. **P-txt-img:** More than 10 million text-only monolingual corpora are used for pre-training and then half of the text and image training set of Multi30k are used for fine-tuning.

- **3Iter.** The principle is to iteratively improve the model based on reconstruction loss, then use a discriminator to align the latent distribution of the source and target language. Before training the model with monolingual text-only corpus, a large amount of synthetic pairing data is used for pre-training.

- **Base+img.** To the best of our knowledge, it is the most recent work on multi-modal UMT. It uses additional visual modalities to recover sentences that has previously masked some words and it is trained using Multi30k monolingual data.

5.5. Experimental results

5.5.1. Comparison with the baseline models

As shown in Table 3, we compare our experimental results with those of existing works. It is noted that Gate, Atten, and IVTA models are trained on 2,014 parallel and 14,000 non-parallel instances. Gate, Atten and IVTA belong to our Semi-MMT systems, and training on 2,014 small-scale parallel and 1,4000 non-parallel instances sees their results surpass the P-txt-img pre-trained on tens of millions of large-scale monolingual corpora. Compared with the S-txt-img model, the BLEU value of the our (IVTA^{+}) model increased from 8.85 to 25.54, and even for the P-txt-img, the BLEU of our model increased by 8.59%. Compared with 3Iter, the BLEU value increased from

Table 3: Comparison of experimental results

Model	Test2016			Test2017		
	BLEU	METEOR	ROUGE	BLEU	METEOR	ROUGE
Existing UMT systems						
S-txt [19]	6.27	11.60	30.60	–	–	–
S-txt-img [19]	8.85	13.80	34.20	–	–	–
P-txt [19]	20.97	25.40	53.90	–	–	–
P-txt-img [19]	23.52	26.10	55.10	–	–	–
3Iter [42]	22.74	–	–	–	–	–
Base+img [31]	16.58	–	–	–	–	–
Our Semi-MMT Systems (***Trained on Para2014 data***)						
Gate	23.39	46.84	63.13	16.33	31.83	41.17
Atten	23.41	47.75	62.80	16.95	35.83	42.14
IVTA	22.95	47.14	56.32	15.62	32.36	41.61
Gate^{+}	23.79	46.96	62.92	16.52	32.28	42.65
Atten^{+}	25.35	48.22	63.26	17.91	36.24	44.77
IVTA^{+}	**25.54**	**50.18**	**65.21**	**18.04**	**36.94**	**46.85**

22.74 to 25.54. Finally, compared with the Base-img, our model obtained 9 BLEU scores improvement, thus verifying the effectiveness of the proposed method.

It can be seen in Table 3 that there is a clear gap between the results of S-txt and S-txt-img. The only difference between these two models is that S-txt-img uses additional image features but S-txt does not. Therefore, it can be concluded that the performance improvement of S-txt-img is from the image. It also supports our view that images play a pivotal role in unsupervised training.

5.5.2. Modality fusion

Gate model utilizes the Gate structure to introduce image information into the encoder and decoder of the Transformer. Atten introduces an additional attention mechanism in the Transformer decoder to introduce image information. IVTA is the fusion of text and image features in a joint semantic space through weight learning. The mul-

timodal feature weights and biases are continuously converged, making the translation results more accurate during training. The experimental results of these three multimodal fusion methods without ensemble learning are basically similar, and the results of IVTA are not as good as the other two.

5.5.3. Multimodal multi-perspective fusion

The symbol "+" represents the model using the multi-perspective fusion method. To avoid the model result being a local optimal solution and improve model performance, we use an ensemble learning method for each sub-model saved in the later stage of model training to fuse their prediction results. Meanwhile, to observe the impact of different multimodal alignment methods on multimodal multi-perspective fusion, ensemble learning is used for these three models. Note that the fused models are sub-models of the same model, not sub-models of the three models fused together.

From our Semi-MMT systems in Table 3, it can be seen that the results of the three ensemble models demonstrate improved performance compared to the results of their sub-models. The ensemble learning strategy fuses some sub-models with different parameters, which gives the model better performance. There is no significant improvement between Gate and Gate$^+$, while there is an improvement between Atten and Atten$^+$. Compared with these three models, the improvement between IVTA and IVTA$^+$ is the most obvious. Although the performance of the IVTA sub-model is not the best, the ensemble model IVTA$^+$ results are better than the other two models. The only difference between these three models is multimodal alignment. From this, it can be seen that IVTA multimodal alignment method has the best performance in multimodal multi-perspective fusion.

To further verify our conclusions, we calculated pared-samples T-test on the results of the three indicators between the sub-model and the fusion model. The p-value is 0.146 between Gate and Gate$^+$, 0.029 between Atten and Atten$^+$, and 0.007 between IVTA and IVTA$^+$. The smaller the p-value, the greater the deviation of the corresponding two model results. Therefore, these results further support our conclusion.

Table 4: Experiment results of parallel data of different sizes

Model	Training data	Test2016			Test2017		
		BLEU	METEOR	ROUGE	BLEU	METEOR	ROUGE
Gate	Para514	17.02	34.83	45.70	10.11	21.95	29.56
	Para1014	21.08	41.05	52.62	15.47	29.71	40.24
	Para2014	23.39	46.84	63.13	16.33	31.83	41.17
Atten	Para514	17.60	34.47	45.89	11.12	22.04	29.73
	Para1014	22.71	45.71	54.24	15.93	37.58	44.15
	Para2014	23.41	47.75	62.80	16.95	35.83	42.14
IVTA	Para514	17.08	34.23	44.81	9.26	19.04	24.46
	Para1014	21.70	42.17	52.79	15.49	30.92	39.35
	Para2014	22.95	47.14	56.32	15.62	32.36	41.61

5.6. Experiments on parallel datasets of different sizes

It can be seen in Table 4 that for the Gate, Atten and IVTA models, the expansion of the parallel corpus significantly improves translation quality. The parallel training corpus Para is increased twice, with 500 for the first time and 1,000 for the second time, respectively. In this process, BLEU indicator showed that the Gate model increased by 23.85% and 10.96%, the Atten model increased by 29.03% and 3.08%, and the IVTA model increased by 27.05% and 5.76% in the Test2017. It can be observed that although the increase of the parallel corpus the second time is larger than that of the first time, the improvement of translation quality in the second time is not as high as that of the first time. In Test2016 and other indicators, with the increase of parallel corpora, the improvement of translation performance of the three modes is still gradually declining. In addition, in order to further verify the above conclusions, as shown in Table 5 we used pared-samples T-test to calculate the deviation of three indexes between Para514 and Para1014 (p-value1) and between Para1014 and Para2014 (p-value2). As can be seen from Table 5, for these three models, p-value1 is smaller than the corresponding p-value2, which means that the improvement between Para514 and Para1014 is greater than that between Para1014 and Para2014.

Table 5: pared-samples T-test of experimental results with parallel data of different sizes

Model	p-value1	p-value2
Gate	0.001	0.058
Atten	0.032	0.404
IVTA	0.022	0.024

Table 6: Experimental results on Multi30k fully parallel data

Model	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
TFMT [10]	37.65	64.81	30.28	57.11	25.94	54.02
Under 1D Image Features						
Gate [58]	38.45	65.63	30.36	57.78	27.42	54.15
Atten [18]	38.84	65.65	30.54	58.45	27.68	55.04
IVTA [59]	38.82	65.78	31.21	58.17	28.12	55.17
Under 2D Image Features						
Gate [58]	38.23	64.47	30.64	57.82	27.34	54.54
Atten [18]	38.71	65.75	30.65	58.45	27.83	54.70
IVTA [59]	38.80	65.70	31.32	58.26	27.97	54.29

In Table 6, to further observe the effect of parallel corpora of different sizes, the experiments of these three models were conducted on fully parallel data of Multi30k. As can be seen from Tables 4 and 6, semi-supervised machine translation techniques still lag behind machine translation techniques with fully parallel corpora.

These models apply six training paths, among which src-tgt/tgt-src supervised-loss uses the parallel data of the Para set for training. Although the Para training data is adjusted, the total data volume remains the same. Therefore, the model training time cost on different Para datasets is not much different. The experiments on a commodity machine equipped with TITAN Xp and 12G memory take about 5 hours.

Table 7: Experimental results on different image granularities (*in Para2014 data*)

Model	Test2016			Test2017		
	BLEU	METEOR	ROUGE	BLEU	METEOR	ROUGE
Under 1D Image Features						
Gate	23.39	46.84	63.13	16.33	31.83	41.17
Atten	23.41	47.75	62.80	16.95	35.83	42.14
IVTA	22.95	47.14	56.32	15.62	32.36	41.61
Under 2D Image Features						
Gate	18.90	40.55	51.31	12.4	22.84	32.02
IVTA	21.87	46.24	52.32	15.31	33.43	41.25
Multi-perspective fusion Under 1D Image Features						
Gate$^+$	23.79	46.96	62.92	16.52	32.28	42.65
Atten$^+$	25.35	48.22	63.26	17.91	36.24	44.77
IVTA$^+$	**25.54**	**50.18**	**65.21**	**18.04**	**36.94**	**46.85**

5.7. Experiments on different image features

To study the influence of image features with different granularity on the translation results, we also conducted experiments using two-dimensional (2D) local image features. To obtain 2D local image features, a $7\times7\times512$ matrix is derived from pool5 layer of pre-trained VGG16, and then this matrix is linearly transformed into a 49×512 matrix. Finally, pad its last row with zeros to make it 50×512. 2D indicates that the model utilizes 2D local image features and 1D indicates that the model utilizes 1D image features (in Section 4.2). As can be seen from Table 7, the experimental results of the three multimodal fusion models in 1D image features are better than those in 2D image features. The Atten model performs best on 1D image features, and the experimental results on 2D image features are unstable or even negligible. This situation is not obvious in Table 6, which may be because global image features are more beneficial to unsupervised translation. In Table 6, in addition to the test sets introduced previously, we also use the MSCOCO test data [3], and the IVTA model has more obvious advantages on it.

By comparing the improvement of the S-txt-img model over the S-txt model in Table 3 and the improvement of the MMT models over the text-only Transformer machine translation (TFMT) in Table 6, it can be seen that the image in the unsupervised model is more effective than that in the supervised model. Even the model Atten (1D), which performs best on the BLEU of the Test2016 in Table 6, has only a 3.6% improvement over TFMT model, which is far less than the 41% improvement between the S-text and S-text-img trained with half Multi30k data. This is consistent with recent research conclusions on MMT, which found that image information has a greater impact on the MMT model in a limited textual context than in a sufficient textual context [70–72].

5.8. Case study

In Table 8, SRC and REF represent source and translation reference sentences, while 2D and $1D^+$ represent the translated sentences of IVTA(2D) and $IVTA^+$ trained in Para2014. IVTA(2D) represents the IVTA trained under 2D image features.

5.8.1. The impact of image feature representation

In case 1 (the first image), the model $IVTA^+$ translates "player" and "guitar" into "artist" and "jacket" respectively. It is likely that the ensemble model is affected by the form of the player and jacket on the image. In addition, comparing the translation results of the model IVTA(2D), $IVTA^+$ correctly translated "on the street", "woman", etc. Although the compression operation of the 1D image features lose the accuracy of the image representation, the text description is often the main body of the image. This is further demonstrated by the experimental results in Table7, where the model IVTA with 1D image features outperforms the model with 2D image features. Therefore, compressing the image may play a role in removing roughness and refinement, making the translation of key parts of the text more accurate. Moreover, 1D image features contain more semantics, making it easier to align data of different modalities for Semi-MMT.

5.8.2. Long sentence translation quality

In case 2 (second image), the experimental result of model $IVTA^+$ is not semantically smooth. However, compared with the translation results of the model IVTA

Table 8: Case study

	**SRC**:	a female performer with a violin plays on a street while a woman with a blue guitar looks on
	REF:	eine musikantin mit einer violine spielt auf der straße während eine frau mit einer blauen gitarre zusieht
	2D:	eine frau spielt mit einem geige auf einer decke während ein mann mit blauen gitarre zusieht
	1D$^+$:	eine künstlerin spielt geige auf einer straße während eine frau mit blauem oberteil zusieht
	**SRC**:	two males seem to be conversing while standing in front of a truck aposs back and behind a metal item while four people stand around them
	REF:	zwei männer stehen vor dem heck eines lasters und hinter einem metallgegenstand und unterhalten sich anscheinend während vier weitere personen um sie herum stehen
	2D:	zwei menschen warten auf einen lastwagen während in der nähe eines autos und einem grill unterhalten
	1D$^+$:	zwei männliche personen die sich zu unterhalten während sie vor einem lastwagen stehen und hinter ihnen steht

(2D), the IVTA$^+$ model noticed that the two speakers were "men", thus the gender information was correctly translated. In addition, it correctly translated the information of "standing in front of the truck". The reason is that the introduction of 1D images can make each word correspond to the complete compressed image information, so the relationship between different entities is more accurate. The IVTA$^+$ model also failed to obtain the information of the other four people, which may be a defect of our model in complex instances that lack obvious features.

5.8.3. Translation quality in various scenarios

On the other hand, the two cases correspond to the "outdoor" and "indoor" scenarios. In case 1, the IVTA$^+$ model translates the "blanket" into "street", i.e., the model re-recognized that the text has the attribute of "outdoor" through the supplement of 1D image features. In case 2, there is no mention of "indoor" in the source sentence, but the 1D image features applied by the IVTA$^+$ model also supplements the text with the semantics of "what are these two people standing behind". It can be seen that the complementary effect of 1D image features on semantics is reflected in the translation of different scenes. After introducing 1D image features, the performance of the model is improved in the scenario of entity information and relationship between entities.

6. Conclusion and future work

This paper proposes a Semi-MMT method, which includes unsupervised and supervised training parts. The unsupervised part is based on the training of existing unsupervised MMT models, while the supervised part uses a small number of parallel corpora to further train the model, thus improving the model performance. Moreover, by comparing the results on ensemble learning of models with different modality alignments, we demonstrate the importance of selecting a reasonable modality alignment. Compared with the baseline model, our model shows better performance, which verifies the feasibility of our proposed idea.

In future research, we plan to further investigate the impact of the proportion of parallel and non-parallel corpora in the total training data on machine translation. More parallel corpora will be used in experiments to further reveal the relationship between the increase of parallel data and translation quality. In addition, because visual inputs generate better translations in limited text contexts, enhancing the impact of images in Semi-MMT would also be a promising direction for future research.

CRediT authorship contribution statement

Lin Li: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Turghun Tayir: Conceptualization, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Yifeng Han:** Conceptualization, Methodology, Investigation, Resources. **Xiaohui Tao:** Investigation, Writing – review & editing, Supervision, Funding acquisition. **Juan D. Velásquez:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors gratefully acknowledge financial support from Grant 62276196 from National Natural Science Foundation of China, Grant 2021BAA030 from Department of Science and Technology of Hubei Province, China, ANID PIA/APOYO AFB180003, and Grant DP220101360 from the Australian Research Council.

References

[1] S. Yao, X. Wan, Multimodal transformer for multimodal machine translation, in: Association for Computational Linguistics, 2020, pp. 4346–4350.

[2] O. Caglayan, M. Kuyu, M. S. Amac, P. Madhyastha, E. Erdem, A. Erdem, L. Specia, Cross-lingual visual pre-training for multimodal machine translation, in: European Chapter of the Association for Computational Linguistics, 2021, pp. 1317–1324.

[3] D. Elliott, S. Frank, L. Barrault, F. Bougares, L. Specia, Findings on second shared task on multimodal machine translation and multilingual image description, in: The Second Conference on Machine Translation, 2017, pp. 215–233.

[4] Y. Huang, H. Xue, B. Liu, Y. Lu, Unifying multimodal transformer for bi-directional image and text generation, in: International Conference on Multimedia, 2021, pp. 1138–1147.

[5] Z. Zhang, L. Schomaker, Divergan: An efficient and effective single-stage framework for diverse text-to-image generation, Neurocomputing 473 (2022) 182–198.

[6] R. Zhao, Z. Shi, Text-to-remote-sensing-image generation with structured generative adversarial networks, IEEE Geosci. Remote. Sens. Lett. 19 (2022) 1–5.

[7] J. Yu, X. Jiang, Z. Qin, W. Zhang, Y. Hu, Q. Wu, Learning dual encoding model for adaptive visual understanding in visual dialogue, IEEE Trans. Image Process. 30 (2021) 220–233.

[8] F. Chen, F. Meng, X. Chen, P. Li, J. Zhou, Multimodal incremental transformer with visual grounding for visual dialogue generation, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP, 2021, pp. 436–446.

[9] Y. Shi, Y. Tan, F. Feng, C. Zheng, X. Wang, Category-based strategy-driven question generator for visual dialogue, in: Chinese Computational Linguistics, 2021, pp. 177–192.

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

[11] O. Caglayan, W. Aransa, Y. Wang, M. Masana, M. García-Martínez, F. Bougares, L. Barrault, J. van de Weijer, Does multimodality help human and machine for translation and image captioning?, in: The First Conference on Machine Translation, 2016, pp. 627–633.

[12] L. Specia, S. Frank, K. Sima'an, D. Elliott, A shared task on multimodal machine translation and crosslingual image description, in: The First Conference on Machine Translation, 2016, pp. 543–553.

[13] L. Tan, L. Li, Y. Han, D. Li, K. Hu, D. Zhou, P. Wang, An empirical study on ensemble learning of multimodal machine translation, in: International Conference on Multimedia Big Data, 2020, pp. 63–69.

[14] X. Qian, Z. Zhong, J. Zhou, Multimodal machine translation with reinforcement learning, 2018, pp. 1–9, arXiv preprint arXiv:1805.02356.

[15] M. Zhou, R. Cheng, Y. J. Lee, Z. Yu, A visual attention grounding neural model for multimodal machine translation, in: Proceedings of Empirical Methods on Natural Language Processing, 2018, pp. 3643–3653.

[16] W. Zhang, J. Yu, H. Hu, H. Hu, Z. Qin, Multimodal feature fusion by relational reasoning and attention for visual question answering, Inf. Fusion 55 (2020) 116–126.

[17] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: International Conference on Learning Representations, 2015, pp. 1–14.

[18] J. Helcl, J. Libovický, D. Varis, CUNI system for the WMT18 multimodal translation task, in: The Third Conference on Machine Translation, 2018, pp. 616–623.

[19] Y. Su, K. Fan, N. Bach, C. J. Kuo, F. Huang, Unsupervised multi-modal neural machine translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10482–10491.

[20] L. Li, T. Tayir, K. Hu, D. Zhou, Multi-modal and multi-perspective machine translation by collecting diverse alignments, in: Pacific Rim International Conference on Artificial Intelligence, 2021, pp. 311–322.

[21] L. Li, T. Tayir, Multimodal machine translation enhancement by fusing multimodal-attention and fine-grained image features, in: International Conference on Multimedia Information Processing and Retrieval, 2021, pp. 267–272.

[22] K. Imamura, E. Sumita, Ensemble and reranking: Using multiple models in the NICT-2 neural machine translation system at WAT2017, in: Proceedings of the 4th Workshop on Asian Translation, 2017, pp. 127–134.

[23] Y. Wang, L. Wu, Y. Xia, T. Qin, C. Zhai, T. Liu, Transductive ensemble learning for neural machine translation, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020, pp. 6291–6298.

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

[25] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, in: Advances in Neural Information Processing Systems, 2014, pp. 3104–3112.

[26] J. Gehring, M. Auli, D. Grangier, D. Yarats, Y. N. Dauphin, Convolutional sequence to sequence learning, in: International Conference on Machine Learning, 2017, pp. 1243–1252.

[27] P. F. Brown, J. Cocke, S. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, P. S. Roossin, A statistical approach to machine translation, Comput. Linguistics 16 (2) (1990) 79–85.

[28] P. F. Brown, S. D. Pietra, V. J. D. Pietra, R. L. Mercer, The mathematics of statistical machine translation: Parameter estimation, Comput. Linguistics 19 (2) (1993) 263–311.

[29] G. Lample, M. Ott, A. Conneau, L. Denoyer, M. Ranzato, Phrase-based & neural unsupervised machine translation, in: Proceedings of Empirical Methods on Natural Language Processing, 2018, pp. 5039–5049.

[30] P. Huang, J. Hu, X. Chang, A. G. Hauptmann, Unsupervised multimodal neural machine translation with pseudo visual pivoting, in: Association for Computational Linguistics, 2020, pp. 8226–8237.

[31] P. Huang, S. Sun, H. Yang, Image-assisted transformer in zero-resource multimodal translation, in: International Conference on Acoustics, Speech and Signal Processing, 2021, pp. 7548–7552.

[32] S. Chen, Q. Jin, J. Fu, From words to sentences: A progressive learning approach for zero-resource machine translation with visual pivots, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 2019, pp. 4932–4938.

[33] P. Vincent, H. Larochelle, Y. Bengio, P. Manzagol, Extracting and composing robust features with denoising autoencoders, in: International Conference on Machine Learning, 2008, pp. 1096–1103.

[34] G. Yuan, J. Li, H. Li, Y. Du, Y. Li, A. Yu, Label-embedding-based multi-core convolution for text categorization, in: International Conference on Advanced Computational Intelligence, 2020, pp. 269–276.

[35] Y. Zhang, M. Lease, B. C. Wallace, Exploiting domain knowledge via grouped weight sharing with application to text categorization, in: Association for Computational Linguistics, 2017, pp. 155–160.

[36] W. Wang, Z. Gan, W. Wang, D. Shen, J. Huang, W. Ping, S. Satheesh, L. Carin, Topic compositional neural language model, in: Artificial Intelligence and Statistics, PMLR, 2018, pp. 356–365.

[37] R. Johnson, T. Zhang, Effective use of word order for text categorization with convolutional neural networks, in: North American Chapter of the Association for Computational Linguistics, 2015, pp. 103–112.

[38] Y. Wang, X. Tan, Deep recurrent belief propagation network for pomdps, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 10236–10244.

[39] H. Lin, F. Meng, J. Su, Y. Yin, Z. Yang, Y. Ge, J. Zhou, J. Luo, Dynamic context-guided capsule network for multimodal machine translation, in: International Conference on Multimedia, 2020, pp. 1320–1329.

[40] Y. Li, H. Sun, S. Feng, Q. Zhang, S. Han, W. Du, Capsule-lpi: a lncrna-protein interaction predicting tool based on a capsule network, BMC Bioinform. 22 (2021) 1–19.

[41] Y. Cheng, S. Shen, Z. He, W. He, H. Wu, M. Sun, Y. Liu, Agreement-based joint training for bidirectional attention-based neural machine translation, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016, pp. 2761–2767.

[42] G. Lample, A. Conneau, L. Denoyer, M. Ranzato, Unsupervised machine translation using monolingual corpora only, in: International Conference on Learning Representations, 2018, pp. 1–14.

[43] T. Mohiuddin, S. R. Joty, Unsupervised word translation with adversarial autoencoder, Comput. Linguistics 46 (2) (2020) 257–288.

[44] Y. Cheng, W. Xu, Z. He, W. He, H. Wu, M. Sun, Y. Liu, Semi-supervised learning for neural machine translation, in: Association for Computational Linguistics, 2016, pp. 1965–1974.

[45] I. Skorokhodov, A. Rykachevskiy, D. Emelyanenko, S. Slotin, A. Ponkratov, Semi-supervised neural machine translation with language models, in: Proceedings of the Workshop on Technologies for MT of Low Resource Languages, 2018, pp. 37–44.

[46] W. Xu, X. Niu, M. Carpuat, Dual reconstruction: a unifying objective for semi-supervised neural machine translation, in: Findings of the Association for Computational Linguistics, 2020, pp. 2006–2020.

[47] P. Huang, F. Liu, S. Shiang, J. Oh, C. Dyer, Attention-based multimodal neural machine translation, in: The First Conference on Machine Translation, 2016, pp. 639–645.

[48] P. Liu, H. Cao, T. Zhao, Gumbel-attention for multi-modal machine translation, 2021, pp. 1–8, arXiv preprint arXiv:2103.08862.

[49] X. Liu, J. Zhao, S. Sun, H. Liu, H. Yang, Variational multimodal machine translation with underlying semantic alignment, Inf. Fusion 69 (2021) 73–80.

[50] I. Calixto, Q. Liu, Incorporating global visual features into attention-based neural machine translation, in: Proceedings of Empirical Methods on Natural Language Processing, 2017, pp. 992–1003.

[51] P. S. Madhyastha, J. Wang, L. Specia, Sheffield multimt: Using object posterior predictions for multimodal machine translation, in: The Second Conference on Machine Translation, 2017, pp. 470–476.

[52] I. Calixto, Q. Liu, N. Campbell, Doubly-attentive decoder for multi-modal neural machine translation, in: Association for Computational Linguistics, 2017, pp. 1913–1924.

[53] L. J. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, 2016, pp. 1–14, arXiv preprint arXiv:1607.06450.

[54] J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, Imagenet: A large-scale hierarchical image database, in: Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[55] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015, pp. 1–14.

[56] T. Baltrusaitis, C. Ahuja, L. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2019) 423–443.

[57] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning, 2015, pp. 2048–2057.

[58] S. Grönroos, B. Huet, M. Kurimo, J. Laaksonen, B. Mérialdo, P. Pham, M. Sjöberg, U. Sulubacak, J. Tiedemann, et al., The memad submission to the WMT18 multimodal translation task, in: The Third Conference on Machine Translation, 2018, pp. 603–611.

[59] Y. Han, L. Li, J. Zhang, A coordinated representation learning enhanced multimodal machine translation approach with multi-attention, in: International Conference on Multimedia Retrieval, 2020, pp. 571–577.

[60] R. Sennrich, B. Haddow, A. Birch, Edinburgh neural machine translation systems for WMT 16, in: The First Conference on Machine Translation, 2016, pp. 371–376.

[61] C. Hokamp, Ensembling factored neural machine translation models for automatic post-editing and quality estimation, in: The Second Conference on Machine Translation, 2017, pp. 647–654.

[62] L. K. Hansen, P. Salamon, Neural network ensembles, IEEE Trans. Pattern Anal. Mach. Intell. 12 (10) (1990) 993–1001.

[63] D. Elliott, S. Frank, K. Sima'an, L. Specia, Multi30k: Multilingual english-german image descriptions, in: Proceedings of the 5th Workshop on Vision and Language, 2016, pp. 70–74.

[64] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Trans. Assoc. Comput. Linguistics 2 (2014) 67–78.

[65] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[66] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, 2015, pp. 1–15.

[67] K. Papineni, S. Roukos, T. Ward, W. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Association for Computational Linguistics, 2002, pp. 311–318.

[68] A. Lavie, A. Agarwal, METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments, in: The Second Workshop on Statistical Machine Translation, 2007, pp. 228–231.

[69] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81.

[70] B. Li, C. Lv, Z. Zhou, T. Zhou, T. Xiao, A. Ma, J. Zhu, On vision features in multimodal machine translation, in: Association for Computational Linguistics, 2022, pp. 6327–6337.

[71] Z. Wu, L. Kong, W. Bi, X. Li, B. Kao, Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation, in: Association for Computational Linguistics and International Joint Conference on Natural Language Processing, 2021, pp. 6153–6166.

[72] O. Caglayan, P. Madhyastha, L. Specia, L. Barrault, Probing the need for visual context in multimodal machine translation, in: North American Chapter of the Association for Computational Linguistics, 2019, pp. 4159–4170.

Highlights

- An effective fusion method improves the mapping between different modalities
- Using a small number of parallel corpora can improve UMT results
- Different modal alignments affect ensemble learning results

CRediT authorship contribution statement

Lin Li: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Turghun Tayir:** Conceptualization, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Yifeng Han:** Conceptualization, Methodology, Investigation, Resources. **Xiaohui Tao:** Investigation, Writing – review & editing, Supervision, Funding acquisition. **Juan D. Velásquez:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: