When Pairs Meet Triplets: Improving Low-Resource Captioning via Multi-Objective Optimization

YIKE WU, College of Computer Science, Nankai University
SHIWAN ZHAO, IBM Research - China
YING ZHANG and XIAOJIE YUAN, College of Computer Science, Nankai University
ZHONG SU, IBM Research - China

Image captioning for low-resource languages has attracted much attention recently. Researchers propose to augment the low-resource caption dataset into (image, rich-resource language, and low-resource language) triplets and develop the dual attention mechanism to exploit the existence of triplets in training to improve the performance. However, datasets in triplet form are usually small due to their high collecting cost. On the other hand, there are already many large-scale datasets, which contain one pair from the triplet, such as caption datasets in the rich-resource language and translation datasets from the rich-resource language to the low-resource language. In this article, we revisit the caption-translation pipeline of the translation-based approach to utilize not only the triplet dataset but also large-scale paired datasets in training. The caption-translation pipeline is composed of two models, one caption model of the rich-resource language and one translation model from the rich-resource language to the low-resource language. Unfortunately, it is not trivial to fully benefit from incorporating both the triplet dataset and paired datasets into the pipeline, due to the gap between the training and testing phases and the instability in the training process. We propose to jointly optimize the two models of the pipeline in an end-to-end manner to bridge the training and testing gap, and introduce two auxiliary training objectives to stabilize the training process. Experimental results show that the proposed method improves significantly over the state-of-the-art methods.

CCS Concepts: • **Computing methodologies → Natural language generation;**

Additional Key Words and Phrases: Image captioning, low-resource, paired dataset, triplet dataset, bridge the gap, auxiliary training objective

1 INTRODUCTION

Much research has been conducted on image captioning in English and achieved quite good performance in recent years. However, for image captioning in some other languages, the performance is far from satisfactory. One issue behind is that caption datasets in these languages are much smaller than those in English, which causes that they cannot provide sufficient supervisory signals for fully training a caption model. We regard such languages as low-resource languages in terms of image captioning.

As directly training on the low-resource caption dataset does not yield good performance, researchers have proposed to augment the low-resource caption dataset (image, low-resource language) pairs into (image, rich-resource language, and low-resource language) triplets [11], which we call *triplet dataset* in this work. By involving an additional sentence in the rich-resource language, they expand the single kind of paired labels in the low-resource caption dataset into three kinds of paired labels,[1] which provides more supervisory signals for model training. Dual attention [18] has been proposed to exploit the triplet dataset to achieve better performance, whose architecture is shown on the right side of Figure 2. In the training process, it first encodes the image and its corresponding English caption, and then feeds their encoded features into the decoder simultaneously to generate captions in the low-resource language. Wu et al. [37] employ a similar architecture with dual attention, and enhance it by adding the cycle consistency constraint on the attention maps in the cycle of image regions, words in the rich-resource language and words in the low-resource language, which makes further use of the supervisory signals in the triplet dataset. Although the triplet dataset preliminarily mitigates the deficiency of supervisory signals in low-resource captioning, the number of triplets in it is still too small due to the high collecting cost, which restricts further improvement on the captioning performance.

We notice that the triplet dataset is not the only dataset that could be leveraged for the low-resource captioning task. In the English image captioning and machine translation tasks, there are monolingual English caption datasets and large parallel corpora which are large in scale and have been already collected. The English image caption dataset provides paired labels for the (image, rich-resource language) pair in the triplet, and the machine translation dataset provides paired labels for the (rich-resource language, low-resource language) pair in the triplet. We refer to such datasets as *paired dataset* considering that they only contain one pair from the triplet. In this article, we propose an effective approach, which exploits not only the triplet dataset but also the large-scale paired datasets to improve the low-resource captioning performance. To clarify the datasets involved in our approach, we illustrate the difference among several distinct training settings in Figure 1: directly training on the low-resource caption dataset, training on the triplet dataset, and the proposed method which performs training on both the triplet dataset and paired datasets (i.e., monolingual English caption dataset and English–German parallel corpus). We consider German as an example of the low-resource languages in this work and the reason is three-fold. First, there is no sufficient caption data in German as in English. Second, the benchmark dataset Multi30K [11] is a triplet dataset of German and English, and we follow this for the convenience of comparison in the experiment. Third, we just verify the effectiveness of our idea via this example.

Unfortunately, it is not trivial to improve the low-resource captioning performance via training on the triplet dataset and paired datasets simultaneously. There are mainly two challenges related to the model architecture and training strategy, respectively. First, as the triplet dataset contains triplet labels and paired datasets contain labels of pairs from the triplet, we need a flexible model architecture so that it could benefit from both triplet labels and paired labels. As shown on the right side of Figure 2, the dual attention model [18] is not flexible because it requires simultaneous

---

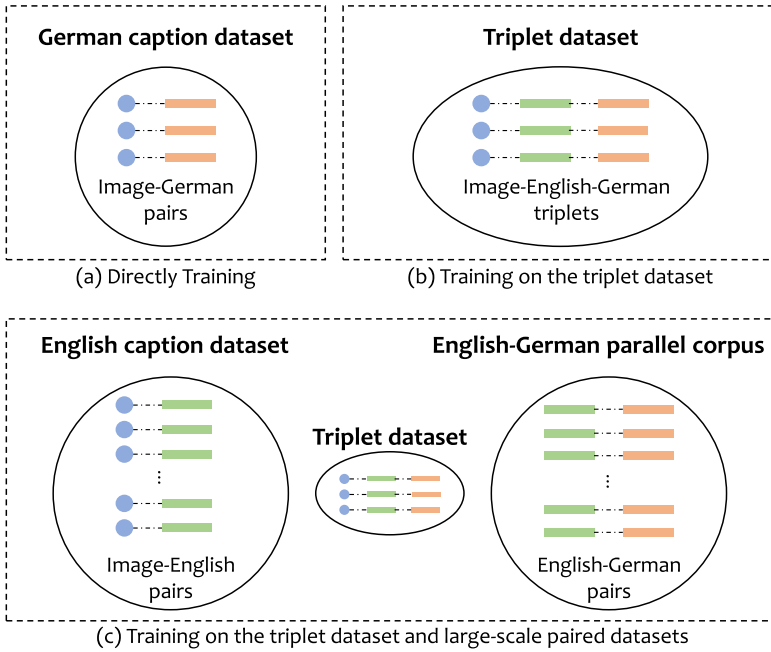[1]Any two elements in the triplet can form one kind of paired labels.

Fig. 1. Datasets in different training settings of the captioning task for low-resource languages. We take English and German as examples of the rich-resource language and low-resource language, respectively.

appearance of image, English, German in triplet during training. To further exploit the paired datasets, we need an English caption model and an English-German translation model to form a caption-translation pipeline, instead of one highly coupled model like the dual attention. The architecture of the caption-translation pipeline is shown on the left side of Figure 2. Since both the caption model and translation model require only paired labels for training, they could benefit from the additional paired datasets. This exactly shares the same spirit with the translation-based methods [13, 22, 23]. However, the previous translation-based methods train the two models only with the paired datasets, which suffers from the inconsistency between different types of paired labels. For example, the English captions from image-English pairs in MSCOCO are very different from English sentences from English-German pairs in a parallel corpus in terms of the language style. This inconsistency makes the translation-based methods even gain no improvement over the model directly trained on the low-resource caption dataset. In this work, we revisit the caption-translation pipeline and mitigate the inconsistency by further leveraging the triplet dataset besides the paired datasets.

Second, incorporating both the triplet dataset and paired datasets into the caption-translation pipeline to achieve better performance requires an elaborate training strategy. On the one hand, training the two models separately does not work well due to the gap between the training and testing phases. More concretely, the translation model is fed with English ground truths in the training phase, but in the testing phase, it is fed with English captions generated from the caption model. To bridge the gap, we propose to optimize the whole pipeline in an end-to-end manner, and use Gumbel-Softmax reparameterization [19, 27] to make the discrete sampling process between the caption model and translation model differentiable. On the other hand, we find that the end-to-end optimization is unstable due to the violent fluctuation of model parameters, and introduce two auxiliary training objectives to stabilize the training process.
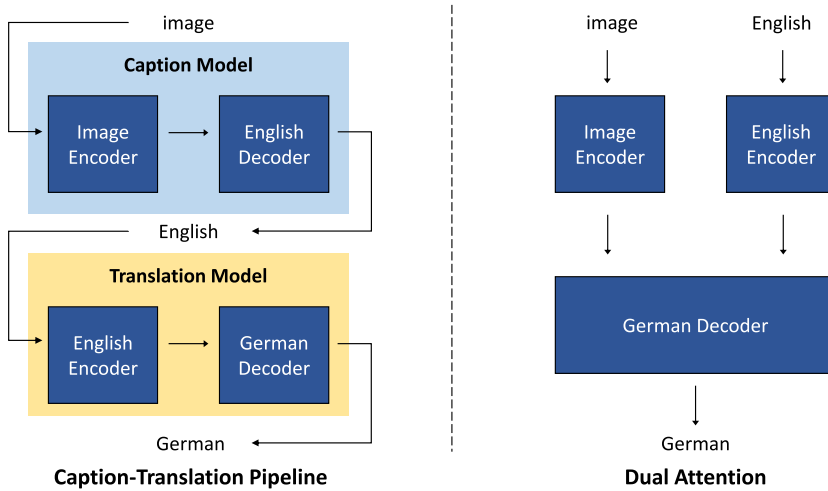
Fig. 2. Architecture comparison between the caption-translation pipeline and the dual attention model [18]. The dual attention model can only leverage the triplet labels since it requires the simultaneous appearance of image, English, German in triplet during training. In contrast, the caption-translation pipeline is flexible to benefit from both triplet labels and paired labels.

In summary, our contributions are three-fold:

— **Data Usage.** To the best of our knowledge, we are the first to exploit not only the triplet dataset but also the large-scale paired datasets to improve captioning for low-resource languages. To benefit from both triplet labels and paired labels simultaneously, we revisit the caption-translation pipeline, which is a flexible architecture composed of an English caption model and an English-German translation model.

— **Optimization Method.** To give full play to the advantage of our data usage over the caption-translation pipeline, we propose an elaborate multi-objective optimization method, which jointly optimizes the whole pipeline in an end-to-end way with two auxiliary training objectives stabilizing the training process. We use Gumbel-Softmax reparameterization to make the end-to-end training differentiable.

— **Experimental Effectiveness.** The proposed data usage and optimization method are indispensable to each other and they together provide a novel paradigm for low-resource captioning. The experimental results verify that this paradigm really makes full use of both the triplet dataset and large-scale paired datasets in an effective way, which achieves state-of-the-art performance on the low-resource captioning.

2 RELATED WORK

2.1 Recent Progress on Image Captioning

Recent years have witnessed much progress on the image captioning methods based on deep learning. Vinyals et al. [36] propose a CNN-RNN architecture to automatically generate image captions in an end-to-end way. They take CNN as the encoder to extract the image feature and feed it into the RNN decoder to trigger the decoding process of captioning. Xu et al. [38] introduce an attention-based model to focus on different spatial grids of an image during the caption generation. Moreover, Lu et al. [25] propose an adaptive attention mechanism, which automatically decides when to attend to the image and when to just rely on the language model, considering that some

words do not have corresponding visual signals in the image. Rennie et al. [33] introduce **reinforcement learning** (**RL**) to the image captioning, which presents a self-critical sequence training approach to optimize the evaluation metric directly. Guo et al. [15] further improve RL-based image captioning by constraining the action space with an n-gram language prior. Researchers also make improvement on captioning performance using more informative image representation for downstream processing. You et al. [42] combine the visual features and visual concepts of an image, and Anderson et al. [2] and Lu et al. [26] exploit an object detector to represent an image as a group of object regions. Yao et al. [40] further boost the region-level features of [2] in a relation-aware manner by considering the semantic and spatial relationships between object regions. Yao et al. [41] encode an image as a hierarchical tree composed of the whole image, region-level features and instance-level features to provide a better image representation for caption generation. Some works [5, 39] also leverage the scene graph representation of an image for image captioning. Instead of improving the image representation, Ke et al. [20] focus on the target decoding side and enhance its long sequential modeling ability. More recently, some researchers also explore the application of new technical advances in the field of image captioning. Cornia et al. [7] present a meshed transformer with memory, and Pan et al. [30] propose a unified X-linear attention block to model the second order interactions with attention mechanism. Besides, many other closely-related works, such as [8, 9, 35] also contribute to the development of image captioning.

2.2 Low-Resource Image Captioning

Although researchers have proposed various image captioning methods and made great progress in the rich-resource languages, there are few studies for low-resource language image captioning. Existing works can be roughly divided into two categories.

Alignment-Based Methods. The first kind of approach usually leverages the triplet dataset composed of data triplets (image, rich-resource language, and low-resource language), and takes the rich-resource language as the only additional input to learn better alignment between visual and linguistic modalities in a common latent space, which leads to improvement on the low-resource captioning performance. Elliott et al. [10] propose several multi-modal architectures that fuse the features of both images and English captions in various ways to generate German captions. However, there are two issues in this work: first, besides the image, it also needs the parallel English captions as input in the testing phase, which is difficult to achieve in practical applications; second, it actually does coarse-level alignment between images and sentences, which could be further refined. Jaffe [18] solves the above two issues by proposing a dual attention model. This model takes only images as input in the testing phase and generates pseudo parallel English captions from a pretrained English caption model. Note that most of research works on the low-resource captioning task follow this setting, i.e., without the parallel English captions. On the other hand, it does fine-grained alignment between image regions and words in a sentence instead of coarsely aligning the images and sentences. Wu et al. [37] further enhance the dual attention model [18] by adding the cycle consistency constraint on the attention maps in the cycle of image regions, English words and German words. In addition, the previous work [29] also falls into this kind of approach and does coarse-level alignment, although it does not leverage the triplet dataset. It first pretrains an English caption model with MSCOCO [24], and then replaces the decoder by a randomly initialized one and trains this new caption model on a Japanese caption dataset. In this way, it transfers knowledge from a rich-resource language caption dataset to the low-resource image captioning.

Our Method vs. Alignment-Based Methods. The two models dual attention [18] and cycle attention [37] do fine-grained alignment between visual and linguistic modalities, which perform the best among the alignment-based methods. Thus, we focus on comparing with them. The main

limitation of these models lies in that they can only leverage the triplet dataset that is too small for training a model to achieve satisfactory results, because their architecture needs the simultaneous appearance of image, English and German in triplet during training. The high requirement for the format of training data limits the usage of other types of datasets in these models, which greatly restricts further improvement on low-resource captioning performance. For example, the large-scale paired datasets contain a number of paired labels that could be beneficial to model training, but these models cannot leverage them since such datasets cannot meet the requirement. In comparison, the caption-translation pipeline in our method is flexible to leverage both triplet dataset and large-scale paired datasets, which is superior to the two models [18, 37] in terms of data usage.

Translation-Based Methods. The second kind of approach is usually based on a caption-translation pipeline composed of a caption model and a translation model, and enhances low-resource captioning only with the large-scale paired datasets, such as monolingual English caption datasets and parallel corpora for machine translation. Li et al. [23] first translate English ground truths into Chinese via an online machine translation service that is trained on a large-scale parallel corpus, and then train a caption model on the pseudo Chinese caption dataset. Lan et al. [22] argue that the quality of dataset constructed as in the work [23] can be improved, and propose evaluation methods to filter sentences in the pseudo dataset. Gu et al. [13] point out that a caption-translation pipeline suffers from the different distributions between caption data and translation data, and involve regularizers to mitigate the problem. However, the translation-based methods even gain no improvement on automatic metrics comparing with the monolingual caption model, which is directly trained on the low-resource caption dataset. We attribute their unsatisfactory performance to two issues. First, they only leverage the paired datasets that contain labels in pair (e.g., image-English or English-German) instead of in triplet (e.g., image-English-German). Although the paired datasets are large in scale, the different types of paired labels are usually inconsistent with each other. This inconsistency leads to a mismatch between the caption model trained on one type of paired labels (e.g., image-English) and the translation model trained on the other type (e.g., English–German), which is detrimental to the performance of the caption-translation pipeline. Second, there exists a gap between training and testing phases if we only train the two models separately in the pipeline, which hinders the full utilization of large-scale paired datasets.

Our Method vs. Translation-Based Methods. Although our method also follows the architecture of the caption-translation pipeline, it solves the two issues in the translation-based methods and significantly ($p < 0.05$) outperforms the monolingual caption model over which the translation-based methods usually gain no improvement (as shown in Section 4.3). First, besides the paired datasets, our method also leverages the triplet dataset to train the caption-translation pipeline. The triplet labels are helpful to "sewing up" the inconsistency between different types of paired labels. Second, it designs an elaborate training strategy that jointly optimizes the whole pipeline in an end-to-end way with two auxiliary training objectives stabilizing the training process. In this way, our method successfully bridges the gap between training and test phases and makes full use of both triplet dataset and large-scale paired datasets.

Positioning Our Method. To the best of our knowledge, this work is the first to leverage both the triplet dataset and large-scale paired datasets, which integrates alignment-based methods and translation-based methods in the view of data usage. In order to give full play to the advantage of such data usage, we propose an elaborate multi-objective optimization method for the caption-translation pipeline composed of a joint training objective and two auxiliary training objectives, which is helpful to making full use of both two kinds of datasets in the pipeline. The proposed data usage and optimization method provide a novel paradigm for low-resource captioning, which significantly improves the captioning performance over the state-of-the-art methods.
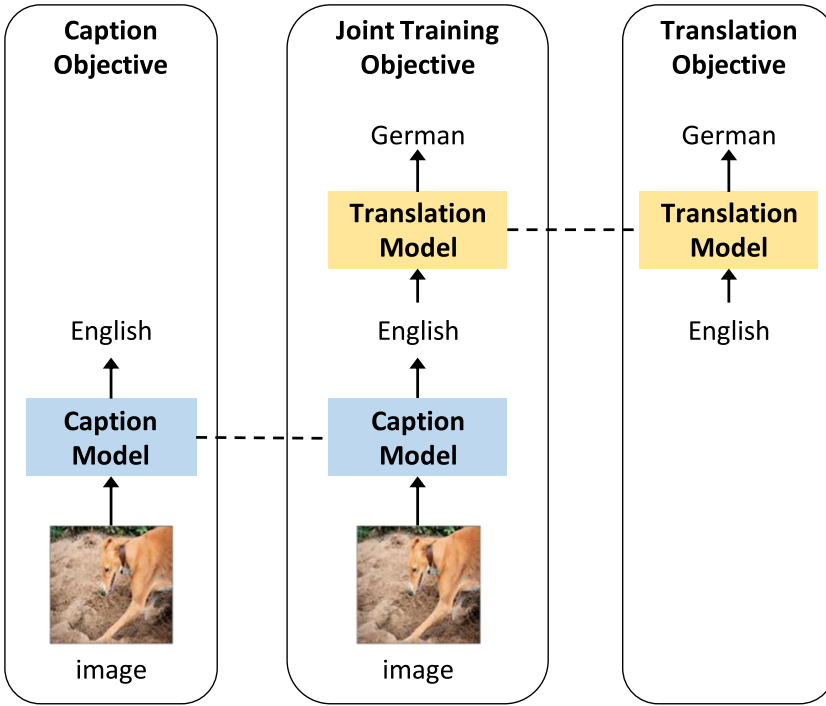
Fig. 3. Framework overview of our method. The caption-translation pipeline is composed of a caption model and a translation model. We train the pipeline with three objectives. The joint training objective jointly optimizes the whole pipeline in an end-to-end manner to bridge the gap between the training and testing phases. The caption objective and translation objective optimize the two models in the pipeline separately to stabilize the training process. Note that the models connected with dotted lines are exactly the same one, which is displayed twice for clarity.

3 METHODOLOGY

3.1 Background

We revisit the caption-translation pipeline in this work, which is composed of an English caption model [2] and an English-German translation model [3]. This pipeline is flexible enough to enable the utilization of the triplet dataset and paired datasets simultaneously. In the inference, we first generate the English caption for a given image by the caption model, and then translate the English caption into German by the translation model. Finally, we can obtain the German caption of the given image.

3.2 Framework Overview

In this work, we leverage both the triplet dataset (*img-En-De* triplets, which can be decomposed into *img-En* pairs and *En-De* pairs) and paired datasets (*img-En* pairs and *En-De* pairs) to train the caption-translation pipeline, where the notations *img*, *En*, and *De* denote image, English, and German, respectively.

As shown in Figure 3, the framework of our method consists of three training objectives, which are built on different types of data: (1) **Joint Training Objective.** We optimize the caption-translation pipeline in an end-to-end manner on the *img-De* pairs of the triplet dataset. (2) **Caption Objective.** We train the caption model separately on the *img-En* pairs of both the triplet dataset

(a) Training the two
models separately

(b) Testing phase
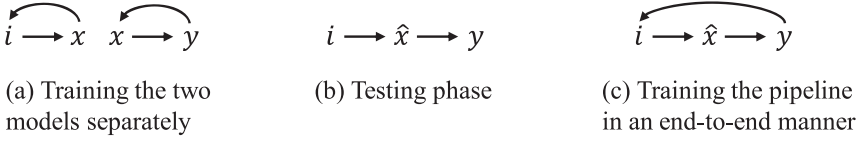
(c) Training the pipeline
in an end-to-end manner

Fig. 4. As shown in (a) and (b), the training phase is inconsistent with the testing phase if we just train the two models in the pipeline separately. To bridge the gap, we optimize the pipeline in an end-to-end manner as shown in (c). The symbols i, x, y, and $\hat{x}$ denote an image, an English ground truth, a German ground truth, and a generated English caption, respectively, and the arrows pointing right and left represent the forward pass and backward pass, respectively.

and paired datasets. (3) **Translation Objective.** We train the translation model separately on the *En-De* pairs of both the triplet dataset and paired datasets.

The three objectives are all indispensable in our approach. On the one hand, if we only train the caption model and translation model separately without the joint training objective, the gap between training and testing phases will be detrimental to the model performance. On the other hand, if we only optimize the whole pipeline in an end-to-end manner without the auxiliary training objectives (i.e., caption objective and translation objective), the training process will be instable and cannot converge well.

Therefore, the total loss $\mathcal{L}$ of our method is a combination of the joint training objective $\mathcal{L}_{joint}$, caption objective $\mathcal{L}_{cap}$ and translation objective $\mathcal{L}_{trans}$, which can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{joint} + \lambda(\mathcal{L}_{cap} + \mathcal{L}_{trans}), \tag{1}$$

where the coefficient λ is introduced to balance the impact of the joint training objective and auxiliary training objectives. In the following sections, we will elaborate how we obtain the losses $\mathcal{L}_{joint}$, $\mathcal{L}_{cap}$ and $\mathcal{L}_{trans}$, respectively.

3.3 Bridge the Gap between Training and Testing Phases

If we only train the two models in the caption-translation pipeline separately, there still exists a gap between the training and testing phases of the pipeline, which is detrimental to German captioning performance. Next, we will explain the gap and illustrate how we bridge it.

3.3.1 What Is the Gap. Formally, we denote an image, an English ground truth, and a German ground truth as i, x, and y, respectively. Given a triplet dataset $\mathcal{D}_t = \{(i, x, y)\}$ or a combination of two paired datasets $\mathcal{D}_p = \{(i, x)\} \cup \{(x, y)\}$, we can optimize two models in the pipeline separately as shown in Figure 4(a), where the translation model is fed with an English ground truth x. However, in the testing phase in Figure 4(b), the translation model is fed with an English caption $\hat{x}$ generated from the caption model. Therefore, the training phase based on x is inconsistent with the testing phase based on $\hat{x}$, which is detrimental to German captioning performance.

3.3.2 Jointly Training the Pipeline Based on Gumbel-Softmax. To bridge the gap between the training and testing phases, our idea is to jointly optimize the two models of the pipeline in an end-to-end manner as shown in Figure 4(c). The joint training objective is built on the image-German pairs $\{(i, y)\}$ of the triplet dataset $\mathcal{D}_t$, which can be formulated as:

$$\mathcal{L}_{joint} = -\mathbb{E}_{(i,y)\sim\mathcal{D}_t}\left[\mathbb{E}_{\hat{x}\sim p(X|i;\Theta_1)}\log P(y|\hat{x};\Theta_2)\right], \tag{2}$$

where Θ_1 and Θ_2 are the parameters of the caption model and translation model, respectively. In this way, we maximize the probability of generating a German ground truth y directly conditioned

on the image i without involving the English ground-truth, which is consistent with the testing phase.

In the forward pass, we first sample an English caption from the caption model. Each word in the caption is represented as a C-dimensional one-hot vector $\boldsymbol{u}$, which is sampled by the Gumbel-Max Trick [14, 28] from a categorical distribution $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_k, \ldots, \pi_C)$ over the English vocabulary of the caption model as follows:

$$\boldsymbol{u} = \text{one\_hot}(\text{argmax}(\boldsymbol{g} + \log \boldsymbol{\pi})), \tag{3}$$

where $\boldsymbol{g} = (g_1, g_2, \ldots, g_k, \ldots, g_C)$ is a vector in which each dimension $g_k \sim \text{Gumbel}(0, 1)$. Then, we feed the English sample composed of one-hot vectors into the translation model to compute $\mathcal{L}_{joint}$.

Unfortunately, in the backward pass, the sampling process between the two models in the pipeline is not differentiable. To solve the problem, we exploit the Gumbel-Softmax reparameterization [19, 27], which circumvents the high variance of reinforcement learning. It provides a contiguous approximation for the one-hot vector $\boldsymbol{u}$:

$$\tilde{\boldsymbol{u}} = \text{softmax}((\boldsymbol{g} + \log \boldsymbol{\pi})/\tau), \tag{4}$$

where the temperature $\tau \in (0, \infty)$ controls the proximity between $\tilde{\boldsymbol{u}}$ and $\boldsymbol{u}$, and thus we can derive the gradient $\frac{\partial \tilde{\boldsymbol{u}}}{\partial \boldsymbol{\pi}}$ as follows:

$$\frac{\partial \tilde{u}_j}{\partial \pi_k} = \tilde{u}_j(\delta_{jk} - \tilde{u}_k)/\tau, \quad j, k \in \{1, 2, ..., C\}, \tag{5}$$

where δ_{jk} is $\mathbb{1}[j = k]$. However, if we feed $\tilde{\boldsymbol{u}}$ into the translation model, a mixture of word embeddings will lead to accumulated mix error during the propagation through RNN as mentioned in [12]. Therefore, we further exploit the Straight-Through version of Gumbel-Softmax [19] to avoid this. Concretely, we feed $\boldsymbol{u}$ into the translation model in the forward pass, and replace $\tilde{\boldsymbol{u}}$ with $\boldsymbol{u}$ in Equation (5) in the backward pass to make the sampling process differentiable.

3.4 Auxiliary Training Objectives

In practice, we find that the end-to-end optimization in Equation (2) is unstable. This is mainly because the parameters of the two models in the pipeline are sharply changed during the optimization, which leads to the instability in the sampling process. Therefore, during the end-to-end optimization, we also introduce two auxiliary training objectives to provide supervision for the two models in the pipeline, respectively, which avoids the violent fluctuation of the model parameters, and thus is beneficial to stabilizing the training process.

3.4.1 Caption Objective. The training objective of the caption model parameterized by Θ_1 is built on the image-English pairs $\{(i, x)\}$ of both the triplet dataset $\mathcal{D}_t$ and paired datasets $\mathcal{D}_p$, which can be formulated as:

$$\mathcal{L}_{cap} = -\mathbb{E}_{(i,x) \sim \mathcal{D}_t \cup \mathcal{D}_p} \log P(x|i; \Theta_1). \tag{6}$$

Next, we elaborate how we obtain $P(x|i; \Theta_1)$, as shown in the example of Figure 5. Specifically, given an image i and its English ground truth $x = \{x_1, x_2, ..., x_L\}$, the caption model first extracts a group of feature vectors $V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_M\}$ from i by an object detector, and then feeds $\overline{\boldsymbol{v}} = \frac{1}{M} \sum \boldsymbol{v}$ into an RNN decoder as the initial hidden state $\boldsymbol{h}_0$. Next, at each time step t of the RNN decoder, we calculate the context vector $\hat{\boldsymbol{v}}_t$ via an attention mechanism f_{att} as follows:

$$\boldsymbol{\alpha}_t = \text{softmax}(f_{att}(V, \boldsymbol{h}_{t-1})), \tag{7}$$

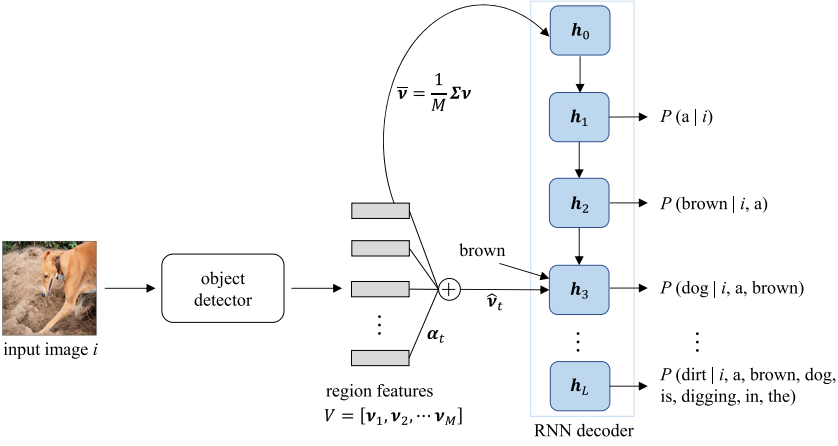$$\hat{\boldsymbol{v}}_t = \boldsymbol{\alpha}_t \cdot V. \tag{8}$$

Fig. 5. The network architecture of the caption model with the data flow of caption objective in it.

For each feature vector $\boldsymbol{v} \in V$, the attention mechanism f_{att} is implemented as Equation (9), where $FC_{1,2,3}$ denote full-connected layers with different parameters and ReLU is the activation function:

$$f_{att} := FC_3(ReLU(FC_1(\boldsymbol{v}) + FC_2(\boldsymbol{h}_{t-1}))), \quad \boldsymbol{v} \in V. \tag{9}$$

Finally, we compute the probability of generating the word x_t and further obtain $P(x|i; \Theta_1)$:

$$P(x_t|i, x_1, ..., x_{t-1}; \Theta_1) = f_{dec}(\hat{\boldsymbol{v}}_t, \boldsymbol{h}_{t-1}, x_{t-1}), \tag{10}$$

$$P(x|i; \Theta_1) = \prod_{t=1}^{L} P(x_t|i, x_1, ..., x_{t-1}; \Theta_1), \tag{11}$$

where f_{dec} denotes the decoding process of the caption model at each time step t.

3.4.2 Translation Objective. The training objective of the translation model parameterized by Θ_2 is built on the English-German pairs $\{(x, y)\}$ of both the triplet dataset $\mathcal{D}_t$ and paired datasets $\mathcal{D}_p$, which can be formulated as:

$$\mathcal{L}_{trans} = -\mathbb{E}_{(x,y) \sim \mathcal{D}_t \cup \mathcal{D}_p} \log P(y|x; \Theta_2). \tag{12}$$

Next, we elaborate how we obtain $P(y|x; \Theta_2)$, as shown in the example of Figure 6. Specifically, given an English sentence $x = \{x_1, x_2, ..., x_L\}$ and its parallel German sentence $y = \{y_1, y_2, ..., y_N\}$, the translation model first encodes x into a sequence of hidden states $S = \{\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_L\}$, and then feeds $\boldsymbol{s}_L$ into an RNN decoder as the initial hidden state $\boldsymbol{g}_0$. Next, we calculate the context vector $\hat{s}_t$ at each time step t of the RNN decoder as follows:

$$\boldsymbol{\beta}_t = softmax(f'_{att}(S, \boldsymbol{g}_{t-1})), \tag{13}$$

$$\hat{s}_t = \boldsymbol{\beta}_t \cdot S, \tag{14}$$

where f'_{att} is also an attention mechanism implemented as follows:

$$f'_{att} := tanh(FC([\boldsymbol{s}; \boldsymbol{g}_{t-1}])), \quad \boldsymbol{s} \in S, \tag{15}$$

where tanh is the activation function, FC denotes a full-connected layer, and [;] denotes the concatenation of two vectors. Finally, we compute the generation probability of the word y_t and
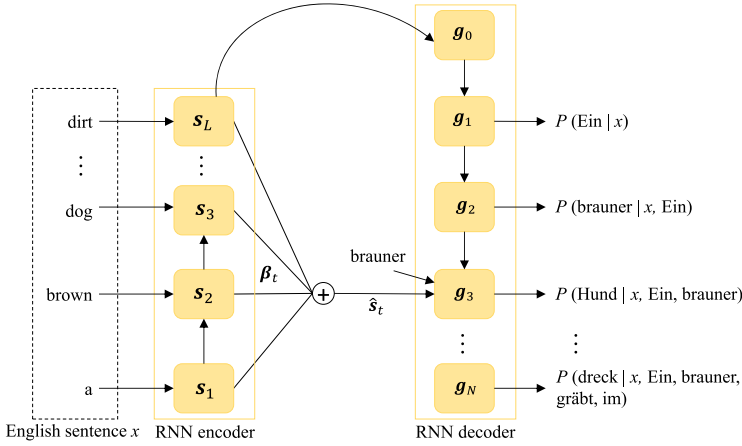
Fig. 6. The network architecture of the translation model with the data flow of translation objective in it.

further obtain $P(y|x; \Theta_2)$ as follows:

$$P(y_t|x, y_1, ..., y_{t-1}; \Theta_2) = f'_{dec}(\hat{s}_t, \boldsymbol{g}_{t-1}, y_{t-1}), \tag{16}$$

$$P(y|x; \Theta_2) = \prod_{t=1}^{N} P(y_t|x, y_1, ..., y_{t-1}; \Theta_2), \tag{17}$$

where f'_{dec} denotes the decoding process of the translation model at each time step t.

4 EXPERIMENTS

4.1 Datasets

Two types of datasets are involved in the experiments: a triplet dataset and two large-scale paired datasets. The large-scale paired datasets include a monolingual English caption dataset and an English-German parallel corpus. The numbers of triplets and pairs in each dataset are listed in Table 1.

Triplet Dataset. We use the Multi30K dataset [11], which is extended from Flickr30K [43]. The Flickr30K dataset consists of 29,000, 1,014, and 1,000 images for training, validation, and testing, and each image is annotated by five English captions. The Multi30K dataset extends it into two versions. Here, we use the translation version, denoted as Multi30K-trans. Multi30K-trans introduces an additional human-translated German caption for one of the five English captions per image, which forms image-English-German triplets.

Large-Scale Paired Datasets. We use MSCOCO [24] as the monolingual English caption dataset in which each image is annotated by five English captions. We use the WMT 2015[2] English-German data as the English-German parallel corpus, denoted as WMT-2015$_{En-De}$, which consists of about 4.3 M sentence pairs from Europarl v7, News Commentary and Common Crawl corpora.

Data Preprocessing. We resize the images of Multi30K-trans and MSCOCO into 450x450, and use Faster R-CNN [32] with a backbone of ResNet-101 [16] to extract the feature vectors from an image. The feature vectors correspond to 36 salient object regions in the image. We filter the English-German pairs of WMT-2015$_{En-De}$ by the sequence length between 3 and 50. In addition, we convert all the sentences in each dataset into lower case, then remove punctuation characters

---

[2]http://www.statmt.org/wmt15/translation-task.html.

Table 1. Statistics of Datasets

Dataset	img-De	img-En	En-De	img-En-De
Multi30K-trans	31,014	31,014	31,014	31,014
MSCOCO	-	616,435	-	-
WMT-2015$_{En-De}$	-	-	4,535,522	-

The notation "img-En-De" denotes the image-English-German triplets.
Accordingly, the notations "img-De", "img-En"", and "En-De" denote
different pairs, respectively. The symbol "-" means the dataset does not
contain the pairs or triplets.

from them, and split them into sequences of tokens and add *<start>* and *<end>* tags at the beginning and end of them, respectively. By collecting the tokens, we construct English vocabulary of 10,536 words from both Multi30K-trans and MSCOCO, and German vocabulary of 3,536 words from Multi30K-trans.

4.2 Experimental Settings

Model Architecture. We employ a bidirectional GRU [6] as the encoder of the translation model, and use single-layer LSTM [17] for the other RNN models in our method. We set all the hidden size, embedding size, and the dimension of attention layer as 512. The dropout rate is set as 0.5.

Training and Inference. The maximum training epoch is set as 50 and we do early stop in training if the performance on CIDEr does not improve for 20 epochs on the validation set. We use Adam optimizer [21] with a learning rate $4x10^{-4}$ and the batch size is set as 32. In the end-to-end optimization, we apply a smaller learning rate 4×10^{-5} for stability, and the temperature τ of Gumbel-Softmax and the coefficient λ are set as 0.5 and 0.8, respectively. In the inference, we adopt the beam search strategy with the beam size of 3, and the maximum decoding step is 50.

Evaluation. We evaluate the German captioning performance on the test split of Multi30K dataset. For each image, we use all six German captions from both the translation version and caption version of Multi30K as the references for stable evaluation. The evaluation metrics include BLEU [31], METEOR [4], CIDEr [34], and SPICE [1]. The BLEU measures the n-gram precision between generated captions and several references. The METEOR focuses on the unigrams, and it judges the word matching based on not only surface forms but also stemmed forms and meaning, which is more consistent with the human judgement. The CIDEr and SPICE are designed for image captioning task. The CIDEr performs a TF-IDF weighting for each n-gram. It gives high weights to the n-grams, which frequently appear in the captions of the specific image and are rarely shared by other image captions in the dataset. The SPICE maps the generated captions and references into scene graphs for evaluation.

4.3 Quantitative Analysis

4.3.1 Compared Methods. We compare our method with two kinds of state-of-the-art methods on common metrics to validate its effectiveness: (1) Translation-Based Methods. Since this kind of methods usually gain no improvement comparing with the monolingual caption model, we only choose the vanilla caption-translation pipeline [23] as the representative of translation-based methods for comparison. (2) Alignment-Based Methods. This kind of methods usually achieve better results than translation-based ones, and thus we focus on comparing with them. We compare our method with the two most competitive alignment-based methods dual attention [18] and cycle attention [37]. All the caption models or translation models follow the same implementation for fairness. In addition, we also list the training data leveraged by compared methods in Table 2 for clear comparison.

Table 2. The Training Data Leveraged by the Different Methods

#	Model	Multi30K-trans	MSCOCO	WMT-2015$_{En-De}$
		img-En-De	img-En	En-De
1	Monolingual Caption Model [2]	✓ (only using img-De)		
2	Cap-Trans Pipeline (paired) [23]		✓	✓
3	Cap-Trans Pipeline (triplet)	✓		
4	Cap-Trans Pipeline (both)	✓	✓	✓
5	Dual Attention [18]	✓		
6	Cycle Attention [37]	✓		
7	Ours	✓	✓	✓

"✓" denotes that the method uses the dataset in the training process. Note that the monolingual caption model
in Row 1 only uses image-German pairs from the image-English-German triplets in Multi30K-trans.

— **Monolingual Caption Model** [2]**.** We train a German caption model on the image-German
pairs of Multi30K-trans. It generates the German caption for a given image directly.
— **Cap-Trans Pipeline (paired, triplet, both).** The previous work [23] trains two models
in the caption-translation pipeline only on the paired datasets, respectively (pair). For com-
pleteness, we also explore to train the caption-translation pipeline only on the triplet dataset
(triplet) and on both triplet and paired datasets (both).
— **Dual Attention** [18]**.** It generates a German caption conditioned on both an image and an
English caption, which requires simultaneous appearance of image, English, and German in
triplet for training. Given an image in the inference, it first generates an English caption by
a pretrained English caption model, and then takes both the given image and the generated
English caption as input to infer the German caption. We pretrain the English caption model
on MSCOCO and the image-English pairs of Multi30K-trans.
— **Cycle Attention** [37]**.** It shares a similar architecture with dual attention and proposes to
add the cycle consistency constraint on the attention maps in the cycle of image regions,
English words, and German words.
— **Ours.** Our method leverages both the triplet dataset and paired datasets to train the caption-
translation pipeline. Before the training process, we pretrain the caption model and transla-
tion model on all the image-English pairs and English-German pairs from different datasets,
respectively.

4.3.2 Results Analysis. Table 3 exhibits the German captioning results of compared methods on
common metrics. Overall, we can see that our approach outperforms all the other methods on most
metrics. First, comparing Row 1 with Row 7, the boost on the performance over the monolingual
caption model validates the effectiveness of our approach for the low-resource image captioning.
Second, comparing with the methods, which only leverage the paired datasets or triplet dataset,
our approach effectively leverages both two kinds of datasets to achieve significant improvement.
On the one hand, comparing with only using the paired datasets (Row 2) to train the two models
in the caption-translation pipeline, respectively, the proposed method achieves much better per-
formance although they follow the same model architecture. On the other hand, compared to the
dual attention model (Row 5), which can only exploit the triplet dataset in training, our method
improves 3.02 on CIDEr, 0.45 on BLEU-4, and 0.28 on SPICE by further benefiting from the paired
datasets. A similar observation holds for cycle attention (Row 6). Third, although training the two
models in the caption-translation pipeline, respectively with both two kinds of datasets (Row 4)
outperforms that with only single kind (Rows 2 and 3), our method achieves further improvement
over Row 4. This indicates that the elaborate training strategy in our method contributes much to

Table 3. Experimental Results of German Captioning on Common Metrics

#	Model	CIDEr	B@1	B@2	B@3	B@4	METEOR	SPICE
1	Monolingual Caption Model [2]	30.77	53.69	35.46	23.51	15.16	**18.01**	4.05
2	Cap-Trans Pipeline (paired) [23]	17.39	47.05	28.20	16.35	9.56	14.23	1.89
3	Cap-Trans Pipeline (triplet)	29.32	53.58	35.58	23.16	14.79	17.32	4.33
4	Cap-Trans Pipeline (both)	31.01	54.59	36.43	23.98	15.83	17.64	4.11
5	Dual Attention [18]	30.82	54.74	36.69	24.30	16.02	17.78	4.08
6	Cycle Attention [37]	31.02	55.07	36.81	24.27	15.90	17.86	4.08
7	Ours	**33.84**†	**55.92**†	**37.69**†	**25.13**†	**16.47**†	17.85	**4.36**†

The notations "B@1, 2, 3, 4" are short for BLEU-1, 2, 3, and 4, respectively. On the metrics with superscript "$\dagger$", our method **significantly (p<0.05)** performs better than both the monolingual caption model and dual attention model. Details on the t-test are in the appendix.

Table 4. Ablation Study to Demonstrate the Contributions from Different Training Objectives

#	JTO	CapO	TransO	CIDEr	B@1	METEOR	SPICE
1		✓	✓	31.01	54.59	17.64	4.11
2	✓			30.67	55.22	17.61	4.27
3	✓		✓	31.38	54.92	17.73	4.09
4	✓	✓		32.50	54.96	17.85	4.28
5	✓	✓	✓	**33.84**	**55.92**	17.85	**4.36**

(JTO: joint training objective; CapO: caption objective; and TransO: translation objective).

the final improvement, which successfully bridges the gap between training and testing phases of the pipeline and thus makes better use of both two kinds of datasets.

4.4 Ablation Study

We provide an ablation study in Table 4 to demonstrate the contributions from different training objectives, which shows that the effective utilization of both the triplet dataset and paired datasets is not trivial and requires an elaborate training strategy. In Table 4, Row 5 represents the proposed method, and Rows 1–4 represent its several variants, respectively. First, we observe that our approach performs better than the variant without the joint training objective in Row 1. This indicates that only training the caption model and translation model separately does not work well, and our approach successfully bridges the gap between the training and testing phases by jointly optimizing the caption-translation pipeline in an end-to-end manner. Second, comparing Row 2 with Row 5, we find that the captioning performance decreases if we only perform the joint training without the auxiliary training objectives (Row 2). This verifies the necessity of the auxiliary training objectives to stabilize the training process. And we also make a deeper discussion on the auxiliary training objectives in Section 4.7. Third, we explore to perform the joint training with only single auxiliary training objective, i.e., translation objective in Row 3 and caption objective in Row 4, respectively. Comparing Row 3/4 with Row 2, we can see that single auxiliary training objective could also lead to mild improvement. Furthermore, the caption objective contributes more than the translation one (The results in Row 4 are better than that in Row 3). This is reasonable since the the sampling process closely related to the training instability is performed based on the output of the caption model. Finally, comparing Row 3/4 with Row 5, we observe that

Table 5. The Training Time Per Epoch of Methods with the Same Model Architecture (Caption-translation Pipeline)

#	Model		Training Time	
			caption model	translation model
1	Cap-Trans Pipeline	pair [23]	21 min	8 h
2		triplet	2 min	3 min
3		both	23 min	8h + 3 min
4	Ours		+ 7 min	

We list the training time of the caption model and the translation model in the pipeline, respectively. "h"Ï denotes an hour, "min"Ï denotes a minute.

using both auxiliary training objectives achieves better results than only using single one, which demonstrates that both of them are indispensable in the training process.

4.5 Empirical Analysis on Training Time

We analyze the additional time cost caused by our method in Table 5. Specifically, we compare the training time of different methods with the caption-translation baseline as the same model architecture, which only differ from each other in terms of data usage and optimization method. First, comparing with the translation-based method [23] that only leverages the paired datasets (Row 1), the additional time cost of also training on the triplet dataset (Row 3) is negligible (2 min vs. 21 min for caption model, 3 min vs. 8 h for translation model). Second, with the utilization of both paired datasets and triplet dataset, the multiple-objective optimization in our method (Row 4) only involves additional 7 min comparing with training the two models in the pipeline separately (Row 3). In summary, either the data usage (using both two kinds of datasets) or the optimization method proposed in this work only slightly increases the training time cost while significantly improving the low-resource captioning performance.

4.6 Qualitative Analysis

We display some generated examples of different methods in Figure 7. In Example (a), our approach generates a more detailed caption than the other methods by capturing the texture "rocky" of the "beach". In Example (b), it describes the image most accurately by identifying the object "skateboard". In Example (c), considering the correctness, our method also performs better than the other methods that misidentify the "drum" as "guitar". The image content of Example (d) is more complicated and difficult for captioning, and only our method successfully describes the " dogs" while the other methods not. To summarize, by effectively leveraging both the triplet dataset and paired datasets, the proposed method can identify the objects in the image more accurately and understand the image content more in detail.

4.7 Discussion

We analyze the trend of captioning performance of our method with/without the auxiliary training objectives (abbreviated as ATO), respectively in Figure 8(a). Without ATO, the CIDEr on both training and validation sets drops during the training process, while it is more steady and shows a rising trend in the setting with ATO. This shows that the auxiliary training objectives are really important to stabilizing the training process. We hypothesize this is mainly because the violent fluctuation of model parameters leads to the instability in the sampling process, and the auxiliary training objectives can prevent the model parameters from changing dramatically and thus make the training process more stable.

(a)	**Monolingual Caption Model:** ein mann steht am strand (a man is standing on the beach) **Dual Attention:** ein mann geht am strand (a man is walking on the beach) **Ours:** ein mann steht an einem **felsigen** strand (a man is standing on a **rocky** beach)
(b)	**Monolingual Caption Model:** ein mann fährt auf einem geländer auf einer \<unk> (a man is driving on a railing on a \<unk>) **Dual Attention:** ein mann in einem blauen hemd springt über ein geländer (a man in a blue shirt jumps over a railing) **Ours:** ein mann in einem blauen hemd macht ein kunststück auf einem **skateboard** (a man in a blue shirt is doing a trick on a **skateboard**)
(c)	**Monolingual Caption Model:** ein mann mit einem hut spielt gitarre (a man with a hat is playing guitar) **Dual Attention :** ein mann mit hut und einem weißen hemd spielt gitarre (a man with a hat and a white shirt plays the guitar) **Ours:** ein mann mit einem **strohhut** spielt **trommel** (a man with a **straw hat** is playing **drum**)
(d)	**Monolingual Caption Model:** eine gruppe von menschen steht auf einem feld (a group of people stands in a field) **Dual Attention :** eine gruppe von menschen steht im gras (a group of people is standing in the grass) **Ours:** eine gruppe von menschen steht mit **zwei hunden** auf einem feld (A group of people stand**s** with **two dogs** in a field)

Fig. 7. Examples of German captions generated by different methods. *Green* and *red* indicate the highlights and mistakes in the captions generated by our method, respectively. We also display the English translations of German captions in brackets for readability.

We also analyze the robustness of our approach when the coefficient λ varies in Figure 8(b). We observe that the proposed method always outperforms the baselines with different values of λ, which indicates that our approach is robust to λ.

For completeness, we further investigate how the model performance varies with different coefficients of three training objectives, as shown in Table 6. Different from the analysis in Figure 8(b), we assign different values to the coefficient of the joint training objective instead of keeping it always 1.0, and tune the coefficients of two auxiliary objectives separately instead of taking them as a whole (i.e., the coefficient λ). We start from the coefficient setting of our final model (Row 13), and vary the three coefficients, respectively, in $[0.2, 1.0]$ with a step of 0.2. We observe that with various coefficient settings, our method can achieve state-of-the-art experimental results in most cases, which verifies its effectiveness on low-resource captioning.

5 CONCLUSION

In this article, we propose to simultaneously exploit the triplet dataset and large-scale paired datasets, which only contain paired labels of the triplet to improve captioning for low-resource languages. We revisit the translation-based approach as it is flexible to incorporate both triplet labels and paired labels. We show that it is not trivial to achieve better performance by this incorporation due to the gap between training and testing and the instability in the training process. To
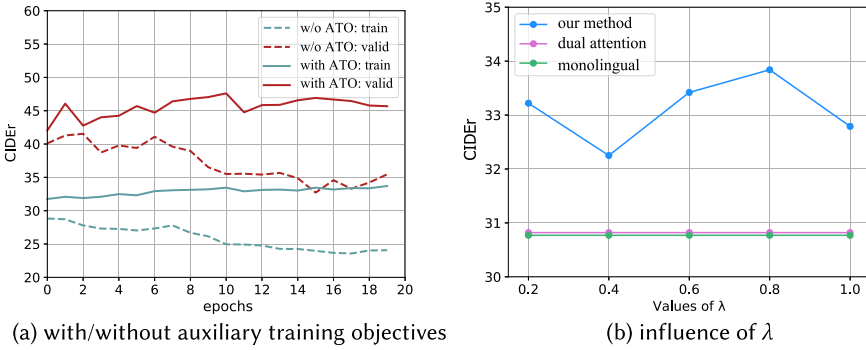
(a) with/without auxiliary training objectives   (b) influence of λ

Fig. 8. Analysis of the auxiliary training objectives and coefficient λ. (a) plots CIDEr of our method from 0 to 20 epoch. The notation "with ATO" or "w/o ATO" denotes the setting with or without the auxiliary training objectives, "train" or "valid" denotes the performance on the training set or validation set. (b) plots CIDEr of our method with different coefficient λ, and the CIDEr of the monolingual caption model and dual attention model are also shown for comparison.

Table 6. Experimental Results of German Captioning with Different Coefficients of the Joint Training Objective (JTO), the Caption Objective (CapO), and the Translation Objective (TransO)

#	JTO	CapO	TransO	CIDEr	B@1	B@2	B@3	B@4	METEOR	SPICE
1	1.0	0.2	0.8	31.95	53.47	35.25	23.03	14.57	17.37	4.11
2	1.0	0.4	0.8	32.89	55.96	37.82	25.30	16.61	17.65	4.30
3	1.0	0.6	0.8	32.94	55.67	37.04	24.67	16.13	18.03	4.27
4	1.0	1.0	0.8	32.27	55.80	37.91	25.61	17.03	18.18	4.38
5	1.0	0.8	0.2	33.57	55.49	37.41	25.12	16.75	18.02	4.33
6	1.0	0.8	0.4	32.23	55.99	37.73	25.25	16.67	17.90	4.34
7	1.0	0.8	0.6	32.78	56.22	37.97	25.58	17.03	17.99	4.31
8	1.0	0.8	1.0	32.18	55.68	37.57	25.16	16.59	17.86	4.28
9	0.2	0.8	0.8	31.30	54.73	36.53	24.26	15.87	17.55	3.90
10	0.4	0.8	0.8	31.44	55.75	37.74	25.27	16.77	18.02	4.00
11	0.6	0.8	0.8	32.84	55.97	37.70	25.48	17.00	18.01	4.32
12	0.8	0.8	0.8	32.39	56.16	38.08	25.57	16.92	18.07	4.12
13	1.0	0.8	0.8	33.84	55.92	37.69	25.13	16.47	17.85	4.36

bridge the gap, we jointly optimize the whole pipeline by making the sampling process from the caption model differentiable using Gumbel-Softmax reparameterization. Furthermore, we introduce two auxiliary training objectives to stabilize the training process. Experimental results show that our approach effectively leverages both the triplet dataset and large-scale paired datasets to significantly improve over the state-of-the-art methods.

APPENDIX

A SIGNIFICANCE TEST

We additionally run our method with the two baselines (monolingual caption model and dual attention model) with 8 different random seeds. A total of 9 groups of experimental results (including one group presented in the main body of the paper and additional 8 groups) are listed in Table 7.

Table 7. Experimental Results of German Captioning on Common Metrics (Nine Groups with Different Random Seeds in Total)

Group Number	Model	CIDEr	B@1	B@2	B@3	B@4	METEOR	SPICE
1	Monolingual	29.77	53.01	34.52	22.23	14.19	17.70	3.98
	Dual Attention	31.19	54.99	36.77	24.38	16.09	17.82	4.11
	Our Method	**33.41**	**55.72**	**37.92**	**25.40**	**16.82**	**17.84**	**4.27**
2	Monolingual	29.46	53.19	35.46	23.59	15.45	17.67	4.03
	Dual Attention	30.90	55.22	37.03	24.49	16.17	17.77	4.05
	Our Method	**32.55**	**55.57**	**37.46**	**24.78**	**16.25**	**17.89**	**4.11**
3	Monolingual	28.94	53.81	35.72	23.33	14.82	17.47	3.92
	Dual Attention	30.74	54.73	36.55	24.16	15.93	17.79	4.09
	Our Method	**32.44**	**55.51**	**37.14**	**24.71**	**16.15**	**17.88**	**4.30**
4	Monolingual	29.36	52.47	34.34	22.47	14.38	17.69	4.02
	Dual Attention	30.87	54.67	36.44	23.91	15.74	17.62	4.04
	Our Method	**32.33**	**55.71**	**37.07**	**24.42**	**15.93**	**17.82**	**4.28**
5	Monolingual	29.88	53.35	35.24	23.13	14.74	17.71	3.84
	Dual Attention	30.60	54.48	36.29	23.79	15.51	17.56	3.90
	Our Method	**32.55**	**55.03**	**36.81**	**24.51**	**15.98**	**18.05**	**4.26**
6	Monolingual	29.47	53.66	35.34	23.31	14.85	**17.80**	**4.10**
	Dual Attention	30.59	54.47	36.10	23.47	15.25	17.58	4.03
	Our Method	**32.63**	**55.44**	**37.24**	**24.60**	**16.02**	17.78	4.08
7	Monolingual	30.38	51.80	33.81	22.01	14.16	17.80	**4.27**
	Dual Attention	30.52	55.02	36.78	24.41	**16.11**	17.65	4.02
	Our Method	**33.26**	**55.65**	**37.30**	**24.71**	16.08	**18.00**	4.26
8	Monolingual	32.10	54.60	36.41	24.11	15.69	18.17	**4.40**
	Dual Attention	30.79	54.47	36.48	23.98	15.71	17.66	4.08
	Our Method	**33.12**	**55.78**	**37.44**	**25.10**	**16.53**	**18.18**	4.17
9	Monolingual	30.77	53.69	35.46	23.51	15.16	**18.01**	4.05
	Dual Attention	30.82	54.74	36.69	24.30	16.02	17.78	4.08
	Our Method	**33.84**	**55.92**	**37.69**	**25.13**	**16.47**	17.85	**4.36**
Average	Monolingual	30.01 ±0.90	53.29 ± 0.77	35.14± 0.75	23.08± 0.66	14.83±0.51	17.78±0.19	4.07± 0.16
	Dual Attention	30.78 ±0.19	54.75 ± 0.26	36.57± 0.27	24.10 ± 0.32	15.84 ± 0.29	17.69± 0.09	4.04 ± 0.06
	Our Method	**32.90 ± 0.49**	**55.59 ±0.24**	**37.34±0.31**	**24.82 ±0.31**	**16.25±0.28**	**17.92 ±0.12**	**4.23 ± 0.09**

Table 8. The P-values of the T-test Performed on the Nine Groups of Results, Comparing Our Method with Two Baselines, Respectively

Model Comparison	p-value						
	CIDEr	B@1	B@2	B@3	B@4	METEOR	SPICE
Ours vs. Monolingual	$5.962 \times e^{-7}$	$4.497 \times e^{-7}$	$9.334 \times e^{-7}$	$4.179 \times e^{-6}$	$3.343 \times e^{-6}$	0.099	0.023
Ours vs. Dual Attention	$4.412 \times e^{-9}$	$4.701 \times e^{-6}$	$7.168 \times e^{-5}$	$2.929 \times e^{-4}$	0.011	$6.538 \times e^{-4}$	$1.137 \times e^{-4}$

In Table 8, we perform t-test on the results in Table 7 and totally obtain 14 p-values by comparing our method with the two baselines, respectively, on all 7 metrics. Most of them (13 of 14) are smaller than 0.05 (significant), and even 11 p-values are far smaller than 0.01.

REFERENCES

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision.* Springer, 382–398.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 6077–6086.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations. 2015.*

[4] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization.* 65–72.

[5] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. 2020. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 9962–9971.

[6] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, 8th Workshop on Syntax, Semantics and Structure in Statistical Translation*, 103.

[7] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10578–10587.

[8] Songtao Ding, Shiru Qu, Yuling Xi, and Shaohua Wan. 2019. A long video caption generation algorithm for big video data retrieval. *Future Generation Computer Systems* 93 (2019), 583–595.

[9] Songtao Ding, Shiru Qu, Yuling Xi, and Shaohua Wan. 2020. Stimulus-driven and concept-driven analysis for image caption generation. *Neurocomputing* 398 (2020), 520–530.

[10] Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multilingual image description with neural sequence models. arXiv:1510.04709 . Retrieved from https://arxiv.org/abs/1510.04709.

[11] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*. 70–74.

[12] Jiatao Gu, D. Im, and V. Li. 2018. Neural machine translation with gumbel-greedy decoding. In *Proceedings of the Association for the Advancement of Artificial Intelligence*.

[13] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. 2018. Unpaired image captioning by language pivoting. In *Proceedings of the European Conference on Computer Vision*. 503–519.

[14] Emil Julius Gumbel. 1954. Statistical theory of extreme values and some practical applications. *NBS Applied Mathematics Series* 33 (1954), 51.

[15] Tszhang Guo, Shiyu Chang, Mo Yu, and Kun Bai. 2018. Improving reinforcement learning based image captioning with natural language prior. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 751–756.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[17] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[18] Alan Jaffe. 2017. Generating image descriptions using multilingual data. In *Proceedings of the Second Conference on Machine Translation*. 458–464.

[19] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. arXiv:1611.01144. Retrieved from https://arxiv.org/abs/1611.01144.

[20] Lei Ke, Wenjie Pei, Ruiyu Li, Xiaoyong Shen, and Yu-Wing Tai. 2019. Reflective decoding network for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8888–8897.

[21] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980. Retrieved from https://arxiv.org/abs/1412.6980.

[22] Weiyu Lan, Xirong Li, and Jianfeng Dong. 2017. Fluency-guided cross-lingual image captioning. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 1549–1557.

[23] Xirong Li, Weiyu Lan, Jianfeng Dong, and Hailong Liu. 2016. Adding chinese captions to images. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 271–275.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*. Springer, 740–755.

[25] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 375–383.

[26] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7219–7228.

[27] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2016. The concrete distribution: A continuous relaxation of discrete random variables. In *Proceedings of the International Conference on Learning Representations*. International Conference on Learning Representations.

[28] Chris J Maddison, Daniel Tarlow, and Tom Minka. 2014. A* sampling. In *Proceedings of the Advances in Neural Information Processing Systems*. 3086–3094.

[29] Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the ACL*.

[30] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10971–10980.

[31] 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 on Association for Computational Linguistics*. Association for Computational Linguistics, 311–318.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in Neural Information Processing Systems*. 91–99.

[33] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7008–7024.

[34] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4566–4575.

[35] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2017. Captioning images with diverse objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5753–5761.

[36] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

[37] Yike Wu, Shiwan Zhao, Jia Chen, Ying Zhang, Xiaojie Yuan, and Zhong Su. 2019. Improving captioning for low-resource languages by cycle consistency. In *Proceedings of the 2019 IEEE International Conference on Multimedia and Expo*. IEEE, 362–367.

[38] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*. 2048–2057.

[39] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10685–10694.

[40] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision*. 684–699.

[41] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2019. Hierarchy parsing for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2621–2629.

[42] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4651–4659.

[43] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.