UNISON: Unpaired Cross-Lingual Image Captioning

Jiahui Gao[1], Yi Zhou[2], Philip L.H. Yu[3*], Shafiq Joty[4], Jiuxiang Gu[5]

[1]The University of Hong Kong, Hong Kong [2]Johns Hopkins University, USA
[3]The Education University of Hong Kong, Hong Kong
[4]Nanyang Technological University, Singapore [5]Adobe Research, USA
sumiler@hku.hk, yzhou188@jhu.edu, plhyu@eduhk.hk, srjoty@ntu.edu.sg, jigu@adobe.com

Abstract

Image captioning has emerged as an interesting research field in recent years due to its broad application scenarios. The traditional paradigm of image captioning relies on paired image-caption datasets to train the model in a supervised manner. However, creating such paired datasets for every target language is prohibitively expensive, which hinders the extensibility of captioning technology and deprives a large part of the world population of its benefit. In this work, we present a novel unpaired cross-lingual method to generate image captions without relying on any caption corpus in the source or the target language. Specifically, our method consists of two phases: (*i*) a cross-lingual auto-encoding process, which utilizing a sentence parallel (bitext) corpus to learn the mapping from the source to the target language in the scene graph encoding space and decode sentences in the target language, and (*ii*) a cross-modal unsupervised feature mapping, which seeks to map the encoded scene graph features from image modality to language modality. We verify the effectiveness of our proposed method on the Chinese image caption generation task. The comparisons against several existing methods demonstrate the effectiveness of our approach.

1 Introduction

Image captioning has attracted a lot of attention in recent years due to its emerging applications, including image indexing, virtual assistants, etc. Despite the impressive results achieved by the existing captioning techniques, most of them focus on English because of the availability of image-caption paired datasets, which can not generalize for languages where such paired dataset is not available. In reality, there are more than 7,100 different languages spoken by billions of people worldwide (source: Ethnologue(2019)). Building visual-language technologies only for English would deprive a significantly large population of non-English speakers of AI benefits and also leads to ethical concerns, such as unequal access to resources. Therefore, similar to other NLP tasks (*e.g.* parsing, question answering) (Hu et al. 2020; Conneau et al. 2018; Gu et al. 2020), visual-language tasks should also be extended to multiple languages. However, creating paired captioning datasets for each target language is infeasible, since the labeling process is very time consuming and requires excessive human labor.

To alleviate the aforementioned problem, there have been several attempts in relaxing the requirement of image-caption paired data in the target language (Gu et al. 2018; Song et al. 2019), which rely on paired image-caption data in a pivot language to generate captions in the target language via sentence-level translation. However, even for English, the existing captioning datasets (e.g., MS-COCO (Lin et al. 2014)) are not sufficiently large and comprise only limited object categories, making it challenging to generalize the trained captioners to scenarios in the wild (Tran et al. 2016). In addition, sentence-level translation relies purely on the text description and can not observe the entire image, which may ignore important contextual semantics and lead to inaccurate translation. Thus, such pivot-based methods fail to fully address the problem.

Recently, a few works explore image captioning task in an unpaired setting (Feng et al. 2019; Gu et al. 2019; Laina, Rupprecht, and Navab 2019). Nevertheless, these methods still rely on manually labeled caption corpus. For example, Gu et al. (2019) train their model based on shuffled image-caption pairs of MS-COCO; Feng et al. (2019) use an image descriptions corpus from Shutterstock; Lania et al. (2019) create training dataset by sampling the images and captions from different image-caption datasets. Despite they belong to *unpaired* methods *in spirit*, one could still argue that they depend heavily on the collected caption corpus to get a reasonable cross-modal mapping between vision and language distributions – a resource that is not always practical to assume. It therefore remains questionable how these methods would perform when there is no caption data at all. To the best of our knowledge, there is yet no work that investigates image captioning without relying on any caption corpus.

Despite the giant gap between images and texts, they are essentially different mediums to describe the same entities and how they are related in the objective world. Such internal logic is the most essential information carried by the medium, which can be leveraged as the bridge to connect data in different modalities. Scene graph(Wang et al. 2018), a structural representation that contains 1) the objects, 2) their respective attributes and 3) how they are related as described by the medium (image or text), which has been developed into a mature technique for visual un-

*Corresponding author.

derstanding tasks in recent years(Johnson, Gupta, and Fei-Fei 2018). Previous researches on scene graph generation have demonstrated its effectiveness in aiding cross-modal alignment (Yang et al. 2019; Gu et al. 2019). However, existing scene graph generators are only available in English, which poses challenges for extending its application on other languages. One naive approach is to perform cross-lingual alignment to other target languages by conducting a superficial word-to-word translation on the scene graphs nodes, which neglects the contextual information of the sentence or the image. Since a word on a single node can carry drastically different meanings in various contexts, such an approach often leads to sub-optimal cross-lingual mapping. To address this issue, we propose a novel Cross-lingual Hierarchical Graph Mapping (HGM) to effectively conduct the alignment between languages in the scene graph encoding space, which benefits from contextual information by gathering semantics across different levels of the scene graph. Notably, the scene graph translation process can be enhanced by the large-scale parallel corpus (bi-text), which is easily accessible for many languages (Esplà et al. 2019).

In this paper, we propose **UN**pa**I**red cros**S**-lingual image capti**ON**ing (**UNISON**), a novel approach to generate image captions in the target language without relying on any caption corpus. Our UNISON framework consists of two phases: (*i*) a cross-lingual auto-encoding process and (*ii*) a cross-modal unsupervised feature mapping (Fig. 1). Using the parallel corpus, the cross-lingual auto-encoding process aims to train the HGM to map a scene graph derived from the source language (English) sentence to the space of the target language (Chinese), and learns to generate a sentence in the target language based on the mapped scene graph. Then, a cross-modal feature mapping (CMM) function is learned in an unsupervised manner, which aligns the image scene graph features from image modality to language modality. The features in language modality is subsequently mapped by HGM, and then fed to the decoder in phase (*i*) to generate image captions in the target language. Our experiments show 1) the effectiveness of the proposed **HGM** when conducting cross-lingual alignment(§5.2) in the scene graph encoding space and 2) the superior performance of our UNISON framework as a whole(§5.1).

2 Related Work

Paired Image Captioning. Previous studies on supervised image captioning mostly follow the popular encoder-decoder framework (Vinyals et al. 2015; Rennie et al. 2017; Anderson et al. 2018), which mostly focus on generating captions in English since the neural image captioning models require large-scale data of annotated image-caption pairs to achieve good performance. To relax the requirement of human effort in caption annotation, Lan, Li, and Dong (2017) propose a fluency-guided learning framework to generate Chinese captions based on pseudo captions, which are translated from English captions. Yang et al. (2019) adopt the scene graph as the structured representation to connect image-text domains and generate captions. Zhong et al. (2020) propose a method to select the important subgraphs of scene graphs to generate comprehensive captioning. Nguyen et al. (2021) further close the semantic gap between image and text scene graphs by HOI labels.

Unpaired Image Captioning. The main challenge in unpaired image captioning is to learn the captioner without any image-caption pairs. Gu et al. (2018) first propose an approach based on pivot language. They obviate the requirement of paired image-caption data in the target language but still rely on paired image-caption data in the pivot language. Feng et al. (2019) use a concept-to-sentence model to generate pseudo-image-caption pairs, and align image features and text features in an adversarial manner. Song et al. (2019) introduce a self-supervised reward to train the pivot-based captioning model on pseudo image-caption pairs. Gu et al. (2019) propose a scene graph-based method for unpaired image captioning on disordered images and captions.

Summary. While several attempts have been made towards unpaired image captioning, they require caption corpus to learn a reasonable cross-modal mapping between vision and language distributions, *e.g.* the corpus in (Feng et al. 2019) is collected from Shutterstock image descriptions, Gu et al. (2019) use the MSCOCO corpus after shuffling the image-caption pairs. Thus, arguably these approaches are not entirely "unpaired" as they rely on the labelled corpus, limiting their applicability to different languages. Meanwhile, our method generates captions in target language without relying on any caption corpus.

3 Methods

3.1 Preliminary and Our Setting

In the conventional paired paradigm, image captioning aims to learn a captioner which can generate an image caption $\hat{S}$ for a given image I, such that $\hat{S}$ is similar to the ground-truth (GT) caption. Given the image-caption pairs $\{I_i, S_i\}_{i=1}^{N_I}$, the popular encoder-decoder framework is formulated as:

$$\mathcal{I} \rightarrow \mathcal{S}: \quad I \rightarrow \boldsymbol{v} \rightarrow \hat{S} \tag{1}$$

where $\boldsymbol{v}$ denotes the encoded image feature. The training objective for Eq. 1 is to maximize the probability of words in the GT caption given the previous GT words and the image.

Compared with paired setting, which relies on paired image-caption data and can not generalize beyond the language used to label the caption, our unpaired setting does not depend on any image-caption pairs and can be extended to other target languages. Specifically, we assume that we have an image dataset $\{I_i\}_{i=1}^{N_I}$ and a source-target parallel corpus dataset $\{(S_i^x, S_i^y)\}_{i=1}^{N_S}$. Our goal is to generate caption $\hat{S}^y$ in the target language y (Chinese) for an image I with the help of unpaired images and parallel corpus.

3.2 Overall Framework

As shown in Fig. 1, there are two phases in our framework: (*i*) a **cross-lingual auto-encoding process** and (*ii*) a **cross-modal unsupervised feature mapping**, which can be formulated as the following equations, respectively:

$$\mathcal{S}^x \rightarrow \mathcal{S}^y : S^x \rightarrow \mathcal{G}^x \Rightarrow \mathcal{G}^y \rightarrow \boldsymbol{z}^y \rightarrow \hat{S}^y \tag{2}$$

$$\mathcal{I} \rightarrow \mathcal{S}^y \ : I \rightarrow \mathcal{G}^{x,I} \Rightarrow \mathcal{G}^{y,I} \rightarrow \boldsymbol{z}^{y,I} \Rightarrow \boldsymbol{z}^y \rightarrow \hat{S}^y \tag{3}$$

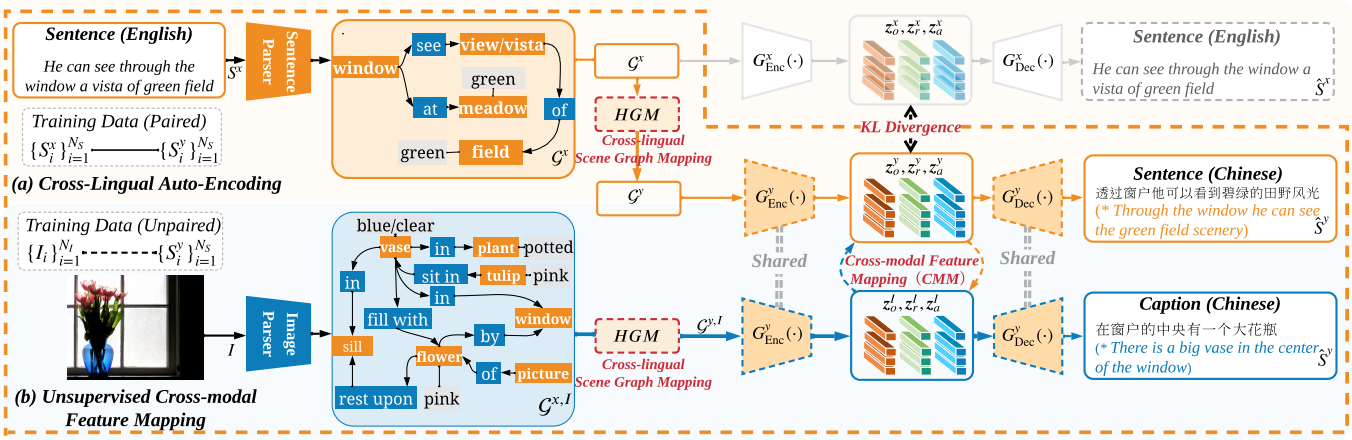


Figure 1: Overview of our UNISON framework. It has two phases: *cross-lingual auto-encoding process* and *unsupervised cross-modal feature mapping*. The cross-lingual scene graph mapping in the first phase (Top) is designed to map the scene graph from the source language (*e.g.* English) to the target language (*e.g.* Chinese) without relying on a scene graph parser in the target language. The unsupervised cross-modal feature mapping in the second phase (Bottom) is designed to align the visual modality to textual modality. We mark the object, relationship, and attribute nodes yellow, blue, and grey in the scene graph. The English sentences (marked in gray) in parentheses are translated by google translator for better understanding.

where $l \in \{x, y\}$ is the source or target language; $\mathcal{S}^l$ is the sentence in laugnage l; $\mathcal{G}^l$ and $\mathcal{G}^{l,I}$ are the scene graphs for the language l in sentence modality and image modality (I), respectively; $\boldsymbol{z}^l$ and $\boldsymbol{z}^{l,I}$ are the encoded scene graph features for $\mathcal{G}^l$ and $\mathcal{G}^{l,I}$, respectively.

The **cross-lingual auto-encoding process** (shown in top of Fig. 1) aims to generate a sentence in the target language given a scene graph in the source language: we first extract the sentence scene graph $\mathcal{G}^x$ from each (English) sentence S^x using a sentence scene graph parser, and map it to $\mathcal{G}^y$ via our proposed HGM (detail in later section). Then we feed $\mathcal{G}^y$ to the encoder to produce the scene graph features $\boldsymbol{z}^y$, which the decoder then takes as inputs to generate $\hat{S}^y$. Note that the mapping from $\mathcal{G}^x$ to $\mathcal{G}^y$ is done at the embedding level, *i.e.* no symbolic $\mathcal{G}^y$ is constructed. This phase addresses the misalignment among different language domains.

The **cross-modal unsupervised feature mapping** (shown in the bottom part of Fig. 1) closes the gap between image modality and language modality: we first extract the image scene graph $\mathcal{G}^{x,I}$ from image I, which is in source language x (English). After that, we map the $\mathcal{G}^{x,I}$ to $\mathcal{G}^{y,I}$ with the HGM (shared with the first phase). As shown in Eq. 3, a cross-modal mapping function ($\boldsymbol{z}^{y,I} \Rightarrow \boldsymbol{z}^y$) is learned, which maps the encoded image scene graph features from image modality to language modality. Once mapped to $\boldsymbol{z}^y$, we can use the sentence decoder to generate $\hat{S}^y$. We further elaborate each phase in detail below.

3.3 Cross-Lingual Auto-Encoding Process

Scene Graph. A scene graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ contains three kinds of nodes: object, relationship and attribute nodes. Let object o_i denote the i-th object. The triplet $\langle o_i, r_{i,j}, o_j \rangle$ in $\mathcal{G}$ is composed of two objects: o_i (as subject role) and o_j (as object role), along with their relation $r_{i,j}$. As each object may have a set of attributes, we denote a_i^k as the k-th attribute of object o_i. To generate an image scene graph $\mathcal{G}^I$, we build the image scene graph generator based on Faster-RCNN (Ren et al. 2015) and MOTIFS (Zellers et al. 2018). To generate sentence scene graph $\mathcal{G}^x$, we first convert each sentence into a dependency tree with a syntactic parser (Anderson et al. 2016), and then apply a rule-based method (Schuster et al. 2015) to build the graph. The $\mathcal{G}^y$ is mapped from $\mathcal{G}^x$ through our HGM module.

Cross-Lingual Hierarchical Graph Mapping (HGM). Our hierarchical graph mapping contains three levels: (*i*) word-level mapping, (*ii*) sub-graph mapping, and (*iii*) full-graph mapping. The semantic information from all three levels are fused in an self-adaptive manner via a self-gated mechanism, which effectively takes into account the structures and relations from the context.

The proposed HGM is illustrated in Fig. 2. Let $\langle \boldsymbol{e}^l_{o_i}, \boldsymbol{e}^l_{r_{i,j}}, \boldsymbol{e}^l_{o_j} \rangle \in \mathcal{G}^l$ denote the triplet for relation $r^l_{i,j}$ in language l, where $\boldsymbol{e}^l_{o_i}$, $\boldsymbol{e}^l_{o_j}$ and $\boldsymbol{e}^l_{r_{i,j}}$ are the embeddings representing subject o^l_i, object o^l_i, and relationship $r^l_{i,j}$. Formally, our hierarchical graph mapping from language x to language y can be expressed as:

$$\begin{aligned}\langle \boldsymbol{e}^y_{o_i}, \boldsymbol{e}^y_{r_{i,j}}, \boldsymbol{e}^y_{o_j} \rangle =& \langle f_{\text{HGM}}(\boldsymbol{e}^x_{o_i}, \mathcal{G}^x), f_{\text{Word}}(\boldsymbol{e}^x_{r_{i,j}}), \\ & f_{\text{HGM}}(\boldsymbol{e}^x_{o_j}, \mathcal{G}^x) \rangle \end{aligned} \tag{4}$$

$$\begin{aligned} f_{\text{HGM}}(\boldsymbol{e}^x_{o_i}, \mathcal{G}^x) =& \alpha_w f_{\text{Word}}(\boldsymbol{e}^x_{o_i}) + \alpha_s f_{\text{Sub}}(\boldsymbol{e}^x_{o_i}, \mathcal{G}^x) \\ & + \alpha_f f_{\text{Full}}(\boldsymbol{e}^x_{o_i}, \mathcal{G}^x) \end{aligned} \tag{5}$$

$$\langle \alpha_w, \alpha_s, \alpha_f \rangle = \text{softmax}\big(f_{\text{MLP}}(f_{\text{Word}}(\boldsymbol{e}^x_{o_i}))\big) \tag{6}$$

where $\boldsymbol{e}^y_{o_i}$, $\boldsymbol{e}^y_{o_j}$ and $\boldsymbol{e}^y_{r_{i,j}}$ are the mapped embeddings in target language y; α_w, α_s, and α_f are the level-wise importance weights calculated by Eq. 6; $f_{\text{MLP}}(\cdot)$ represents a multi-layer perception (MLP) composed of three fully-connected (FC) layers with ReLU activations.

Word-level Mapping. The word-level mapping relies on a retrieval function $f_{\text{Word}}(.)$: after obtaining an embedding in language x, $f_{\text{Word}}(.)$ retrieves the most similar embedding in language y from a cross-lingual word embedding space as illustrated in Fig. 2(a), where cosine similarity is used to measure the distance. In practice, we adopt the pre-trained common space trained on Wikipedia following (Joulin et al. 2018). The retrieved embedding is then passed to an FC layer to obtain a high-dimension embedding in language y.

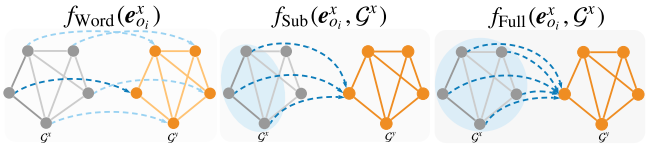
Graph-level Mapping. Since the relation and structure of surrounding nodes encode crucial context information, we introduce the node mapping with graph-level information (as illustrated by Fig. 2(b) and 2(c)): namely sub-graph mapping (f_{Sub}) and full-graph mapping (f_{Full}), which first construct the contextualized embedding in graph-level and then conduct the cross-lingual mapping on the produced embedding. More specifically, for sub-graph mapping, the contextualized embedding is computed by: $\sum_{k=1}^{N'_o} sconv(\boldsymbol{e}^x_{o_i}, \boldsymbol{e}^x_{o_k})/N'_o$, where N'_o is the total number of nodes directly connected to node o_i, and $sconv(\cdot)$ is the spatial convolution operation (Yang et al. 2019). For full-graph mapping, the contextualized embedding is calculated by an attention module: $\sum_{k=1}^{N_o} \alpha_k \boldsymbol{e}^x_{o_k}$, where α_k is calculated by softmax over all the object embeddings $\boldsymbol{e}^x_{o_{1:N_o}}$. Both f_{Sub} and f_{Full} use a linear mapping to project the resulted contextualized (English) embedding to the target (Chinese) embedding space. We consider graph-level mapping only for the object nodes since relationships only exist between objects. For relationship and attribute nodes, only word-level mapping is performed.

Self-gated Adaptive Fusion. To leverage the complementary advantages of information in different levels, we propose a self-gate mechanism to adaptively adjust the importance weights when fusing the embeddings. Specifically, the importance scores are calculated based on the word-level embeddings by passing it through a three-class MLP and a softmax function (Eq. 6). Compared with directly concatenating the embeddings from different levels, which assigns them with equal importance, our fusing mechanism adaptively concentrates on important information and suppress the noises when the context becomes sophisticated.

Scene Graph Encoder. We encode the $\mathcal{G}^x$ and $\mathcal{G}^y$(mapped by the HGM) with two scene graph encoders $G^x_{\text{Enc}}(\cdot)$ and $G^y_{\text{Enc}}(\cdot)$, which are implemented by spatial graph convolutions. The output of each scene graph encoder can be formulated as:

$$\boldsymbol{f}^l_{o_{1:N^l_o}}, \boldsymbol{f}^l_{r_{1:N^l_r}}, \boldsymbol{f}^l_{a_{1:N^l_a}} = G^l_{\text{Enc}}(\mathcal{G}^l), \quad l \in \{x, y\} \tag{7}$$

where $\boldsymbol{f}^l_{o_{1:N^l_o}}$, $\boldsymbol{f}^l_{r_{1:N^l_r}}$, and $\boldsymbol{f}^l_{a_{1:N^l_a}}$ denote the set of encoded object embeddings, relationship embeddings, and attribute embeddings, respectively. Each object embedding $\boldsymbol{f}^l_{o_i}$ is calculated by considering relationship triplets $\langle \boldsymbol{e}^l_{\text{sub}(o_i)}, \boldsymbol{e}^l_{r_{\text{sub}(o_i),i}}, \boldsymbol{e}^l_{o_i} \rangle$ and $\langle \boldsymbol{e}^l_{o_j}, \boldsymbol{e}^l_{r_{j,\text{obj}(o_i)}}, \boldsymbol{e}^l_{\text{obj}(o_i)} \rangle$; $\text{sub}(o_i)$ represents the subjects where o_i acts as an object, and $\text{obj}(o_i)$ represents the objects where o_i plays the subject role. $\boldsymbol{f}^l_{r_i}$ is calculated based on relationship triplet $\langle \boldsymbol{e}^l_{o_i}, \boldsymbol{e}^l_{r_{i,j}}, \boldsymbol{e}^l_{o_j} \rangle$. $\boldsymbol{f}^l_{a_i}$ is the attribute embedding calculated by object o_i and its associated attributes.



(a) Word-level Mapping (b) Sub-Graph Mapping (c) Full-Graph Mapping

Figure 2: Illustration of HGM. Sub-graph mapping only considers those directly connected nodes, while full-graph mapping considers all the nodes in the scene graph.

Sentence Decoder. As shown in Fig. 1, we have two decoders: $G^x_{\text{Dec}}(\cdot)$ and $G^y_{\text{Dec}}(\cdot)$. Each decoder is composed of three attention modules and an LSTM-based decoder. It takes the encoded scene graph features as input and generates the captions. The decoding process is defined as:

$$\boldsymbol{o}^l_t, \boldsymbol{h}^l_t = \text{G}^l_{\text{Dec}}\Big(f_{\text{Triplet}}\big([\boldsymbol{z}^l_o, \boldsymbol{z}^l_r, \boldsymbol{z}^l_a]\big), \boldsymbol{h}^l_{t-1}, \hat{s}^l_{t-1}\Big) \tag{8}$$

$$\hat{s}^l_t \sim \text{softmax}(\boldsymbol{W}_o \boldsymbol{o}^l_t) \tag{9}$$

where $l \in \{x, y\}$, $\hat{s}^l_t$ is the t-th decoded word drawn from the dictionary according to the softmax probability, $\boldsymbol{W}_o$ is a learnable weight matrix, $\boldsymbol{o}^l_t$ is the cell output of the decoder, $\boldsymbol{h}^l_t$ is the hidden state. $f_{\text{Triplet}}(\cdot)$ is a non-linear mapping function that takes the concatenated features as input and outputs the triplet level feature. $\boldsymbol{z}^l_{o_i}$ is calculated by the attention module defined as: $\sum_i^{N^l_o} \alpha^l_{o_i} \boldsymbol{f}^l_{o_i}$, where $\alpha^l_{o_i}$ is the attention weight calculated by the softmax operation over $\boldsymbol{f}^l_{o_{1:N^l_o}}$. $\boldsymbol{z}^l_{r_i}$ and $\boldsymbol{z}^l_{a_i}$ are calculated in a similar way.

Joint-training Mechanism. Inspired by the fact that common structures exist in the encoded scene graph space that are language-agnostic, which may be leveraged to benefit the encoding process, we propose joint training mechanism to enhance the features in target language with the help of features in the source language. In practice, we train a separate scene graph encoder for each language in parallel, then align the encoded scene graph features by enforcing them to be semantically close.

Specifically, we train the scene graph encoders (G^x_{Enc} and G^y_{Enc}), sentence decoders (G^x_{Dec} and G^y_{Dec}), and the cross-lingual HGM module ($\mathcal{G}^x \Rightarrow \mathcal{G}^y$), supervised by a parallel corpus. The two graph encoders encode $\mathcal{G}^x$ and $\mathcal{G}^y$ into feature representations and predict sentences ($\hat{S}^x$ and $\hat{S}^y$) with the decoders. We minimize the following loss:

$$\begin{aligned} \mathcal{L}_{\text{XE}} = &- \sum_t \log P_{\theta_{\mathcal{G}^x \to S^x}}(s^x_t | s^x_{0:t-1}, \mathcal{G}^x) \\ &- \sum_t \log P_{\theta_{\mathcal{G}^y \to S^y}}(s^y_t | s^y_{0:t-1}, \mathcal{G}^y) \end{aligned} \tag{10}$$

where the s^x_t and s^y_t are the ground truth words, $\mathcal{G}^x$ and $\mathcal{G}^y$ are the sentence scene graphs in different languages with $\mathcal{G}^y$ being derived from $\mathcal{G}^x$ using our HGM, $\theta_{\mathcal{G}^x \to S^x}$ and $\theta_{\mathcal{G}^y \to S^y}$ are the parameters of two encoder-decoder models.

To close the semantic gap between the encoded scene graph features $\{\boldsymbol{z}_o^x, \boldsymbol{z}_a^x, \boldsymbol{z}_r^x\}$ and $\{\boldsymbol{z}_o^y, \boldsymbol{z}_a^y, \boldsymbol{z}_r^y\}$, we introduce a Kullback–Leibler (KL) divergence loss:

$$\mathcal{L}_{\text{KL}} = \exp\big(\text{KL}(p(\boldsymbol{z}_o^x)||p(\boldsymbol{z}_o^y)\big) + \exp\big(\text{KL}(p(\boldsymbol{z}_a^x)||p(\boldsymbol{z}_a^y)\big) + \exp\big(\text{KL}(p(\boldsymbol{z}_r^x)||p(\boldsymbol{z}_r^y)\big) \quad (11)$$

where $p(\cdot)$ is composed of a linear layer that maps the input features to a low-dimension d_c, followed by a softmax to get a probability distribution. The overall objective of our joint training mechanism is as follows: $\mathcal{L}_{\text{Phase 1}} = \mathcal{L}_{\text{XE}} + \mathcal{L}_{\text{KL}}$.

3.4 Unsupervised Cross-Modal Feature Mapping

To adapt the learned model from sentence modality to image modality, we drew inspiration from (Gu et al. 2019) and adopt CycleGAN (Zhu et al. 2017) to align the features. For each type $p \in \{o, r, a\}$ of triplet embedding in Eq. 8, we have two mapping functions: $g^p_{I\to y}(\cdot)$ and $g^p_{y\to I}(\cdot)$, where $g^p_{I\to y}(\cdot)$ maps the features from image modality to the sentence modality, and $g^p_{y\to I}(\cdot)$ maps from sentence modality to the image modality. Note that we freeze the cross-lingual mapping module trained in the first phase. The training objective for cross-modal feature mapping is defined as:

$$\mathcal{L}^p_{\text{CycleGAN}} = \mathcal{L}^{I\to y}_{\text{GAN}} + \mathcal{L}^{y\to I}_{\text{GAN}} + \lambda\mathcal{L}^{I\leftrightarrow y}_{\text{cyc}} \quad (12)$$

where $\mathcal{L}^{I\leftrightarrow y}_{\text{cyc}}$ is a cycle consistency loss, $\mathcal{L}^{I\to y}_{\text{GAN}}$ and $\mathcal{L}^{y\to I}_{\text{GAN}}$ are the adversarial losses for the mapping functions with respect to the discriminators.

Specifically, the objective of the mapping function $g^p_{I\to y}(\cdot)$ is to fool the discriminator D^p_y through adversarial learning. We formulate the objective function for cross-modal mapping as:

$$\mathcal{L}^{I\to y}_{\text{GAN}} = \mathbb{E}_S[\log D^p_y(\boldsymbol{z}^y_p)] + \mathbb{E}_I[\log(1 - D^p_y(g^p_{I\to y}(\boldsymbol{z}^I_p))] \quad (13)$$

where $\boldsymbol{z}^y_p$ and $\boldsymbol{z}^I_p$ are the encoded embeddings for sentence scene graph $\mathcal{G}^y$ and image scene graph $\mathcal{G}^{y,I}$, respectively. The adversarial loss for sentence to image mapping $\mathcal{L}^{y\to I}_{\text{GAN}}$ is similarly defined. The cycle consistency loss $\mathcal{L}^{I\leftrightarrow y}_{\text{cyc}}$ is designed to regularize the training and make the mapping functions cycle-consistent:

$$\mathcal{L}^{I\leftrightarrow y}_{\text{cyc}} = \mathbb{E}_I[\|g^p_{S\to I}\big(g^p_{I\to S}(\boldsymbol{z}^I_p)\big) - \boldsymbol{z}^I_p\|_1] + \mathbb{E}_y[\|g^p_{I\to y}\big(g^p_{y\to I}(\boldsymbol{z}^y_p)\big) - \boldsymbol{z}^y_p\|_1] \quad (14)$$

The overall training objective for phase 2 becomes: $\mathcal{L}_{\text{Phase 2}} = \mathcal{L}^o_{\text{CycleGAN}} + \mathcal{L}^a_{\text{CycleGAN}} + \mathcal{L}^r_{\text{CycleGAN}}$.

3.5 Inference of the UNISON Framework

During inference, given an image I, we first extract the image scene graph $\mathcal{G}^{x,I}$ with a pre-trained image scene graph generator and then map the $\mathcal{G}^{x,I}$ in x (English) to $\mathcal{G}^{y,I}$ in y (Chinese) with our HGM module. After that, we encode $\mathcal{G}^{y,I}$ with $G^y_{\text{Enc}}(\cdot)$ and map the encoded features to the language domain through $g^p_{I\to y}(\cdot)$. The mapped features are then fed to the LSTM-based sentence decoder $G^y_{\text{Dec}}(\cdot)$ to generate the image caption $\hat{S}^y$ in target language y.

4 Experiments

4.1 Datasets and Setting

Datasets. For cross-lingual auto-encoding, we collect a paired English-Chinese corpus from existing MT datasets, including WMT19 (Barrault et al. 2019), AIC_MT (Wu et al. 2017), UM (Tian et al. 2014), and Trans-zh (Brightmart 2019)[1]. We filter the sentences in MT datasets according to an existing caption-style dictionary containing 7,096 words in Li et al. (2019). For the first phase, we use 151,613 sentence pairs for training, 5,000 sentence pairs for validation, and 5,000 pairs for testing. For the second phase, following Li et al. (2019), we use 18,341 training images from MS-COCO and randomly select 18,341 Chinese sentences from the training split of the MT corpus. During evaluation, we use the validation and testing splits in COCO-CN.

Corpus	0 Obj/$\mathcal{G}$	1 Obj/$\mathcal{G}$	2 Obj/$\mathcal{G}$	$\geqslant$3 Obj/$\mathcal{G}$
Raw	17.7%	42.6%	24.4%	15.4%
Back-Trans.	12.3%	13.3%	15.1%	59.3%

Table 1: Statistics of the English sentence scene graphs, where n Obj/$\mathcal{G}$ denotes the number of object in a scene graph, $\geqslant$ means greater than or equal to 3.

Preprocessing. We extract the image scene graph with MOTIFS (Zellers et al. 2018) pretrained on VG (Krishna et al. 2017). We tokenize and lowercase the English sentences, then replace the tokens appeared less than five times with UNK, resulting in a vocabulary size of 13,194. We segment the Chinese sentences with *Jieba*[2], resulting in a vocabulary size of 11,731. The English sentence scene graphs are extracted with the parser proposed by (Anderson et al. 2016). We augment the English sentences with the pre-trained back-translators (Ng et al. 2019), resulting in 808,065 English sentences in total, which helps enrich the English sentence scene graphs. Specifically, the statistics in Table 1 shows that the percentage of scene graphs containing more than 3 objects is increased from 15.4% to 59.3%.

4.2 Implementation Details

During cross-lingual auto-encoding phase, we set the dimension of scene graph embeddings to 1,000 and d_c to 100. LSTM with 2 layers is adopted to construct the decoder, whose hidden size is 1000. We start by initializing the graph mapping from a pre-trained common space (Joulin et al. 2018) to stabilize training. The cross-lingual encoder-decoder is firstly trained with the $\mathcal{L}_{\text{XE}}$ for 80 epochs, then with joint loss $\mathcal{L}_{\text{Phase 1}}$ for 20 epochs.

During unsupervised cross-modal mapping phase, we learn the cross-modal feature mapping on the unpaired MS-COCO images and translation corpus. Specifically, we inherit and freeze the parameters of the Chinese scene graph encoder, HGM, and Chinese sentence decoder from cross-lingual auto-encoding process. The cross-modal mapping

[1] https://doi.org/10.5281/zenodo.3402023

[2] https://github.com/fxsjy/jieba

	Method	B@1	B@2	B@3	B@4	METEOR	ROUGH	CIDEr
Setting w/o caption corpus								
Un.	Graph-Aligh(En)(Gu et al. 2019) +GoogleTrans.	39.2	16.7	6.5	2.3	13.2	26.5	9.3
	UNISON	**44.9**	**19.9**	**8.6**	**3.3**	**16.5**	**29.6**	**12.7**
Setting w/ caption corpus								
Pair	FC-2k (En)(Rennie et al. 2017)+GoogleTrans.	58.9	38.0	23.5	14.3	23.5	40.2	47.3
	FC-2k (Cn, Pseudo COCO)(Rennie et al. 2017)	60.4	40.7	26.8	17.3	24.0	43.6	52.7
Un.	UNISON	**63.4**	**43.2**	**29.5**	**17.9**	**24.5**	**45.1**	**53.5**

Table 2: Performance comparisons on the test split of COCO-CN. 'Un.' is short for Unpaired. B@n is short for BLEU-n. 'En' and 'Cn' in the parentheses represent English and Chinese, respectively. 'GoogleTrans' stands for google translator.

functions and discriminators are learned with $\mathcal{L}_{\text{Phase 2}}$. We optimize the model with Adam, batch size of 50, and learning rate of 5×10^{-5}. The discriminators are implemented with a linear layer of dimension 1,000 and a LeakyReLU activation. We set λ to 10. During inference, we use beam search with a beam size of 5. We use the popular BLEU (Papineni et al. 2002), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), METEOR (Denkowski and Lavie 2014) and ROUGE (Lin 2004) for evaluation.

4.3 Model Statement

To gain insights into the effectiveness of our HGM, we construct ablative models by progressively introducing cross-lingual graph-mappings in different levels:

GM_{BASE} is our baseline model, which adopts Google's MT system (Wu et al. 2016) to symbolically map the scene graph from English to Chinese in a node-to-node manner.

GM_{WORD} maps the English scene graph to Chinese through word-level mapping in the scene graph encoding space.

$\text{GM}_{\text{WORD+SUB.}}$ considers both word-level and subgraph-level mappings by directly concatenating them.

HGM_{BASE} considers mappings across all levels, which are directly concatenated and passed through an FC layer.

HGM is similar to HGM-base, except that it adopts a self-gated fusion to adaptively fuse the three features, as illustrated by Eq. 5 and Eq. 6.

5 Results and Analysis

5.1 Overall Results

We demonstrate the superior performance of the proposed UNISON framework on Chinese image captions generation task. We first compare UNISON with the SOTA unpaired method Graph-Align(Gu et al. 2019) under the setting without using any caption corpus. More specifically, we run the Graph-Align[3] and translate the generated English captions to Chinese by google translator for comparison. From the result in Table 2, we can find that our method significantly surpasses Graph-Align with translation, demonstrating that translation in graph level is superior to translation in sentence level. This is reasonable since the graph level alignment is able to consider structural and relational information of the whole image, while sentence level translation suffers from information loss as it can only observe the predicted sentences, and can be severely affected if the translation tools perform poorly. We do not compare with the other unpaired method (Song et al. 2019) here, as the dataset and codes are not publicly available.

To further verify the effectiveness of our framework, we compare UNISON with the supervised pipeline methods: **(*i*)** *FC-2k(En)+Trans.* We train the FC-2k model on image-caption pairs(En) of MS-COCO and translate the generated captions(En) to caption(Cn) using Google translator; **(*ii*)** *FC-2k(Pseudo).* We train the FC-2k model on pseudo Chinese image-caption pairs of MS-COCO, where the captions(Cn) are translated by Google translator from captions(En). For such comparisons, we fine-tune our cross-lingual mapping on the unpaired captions. The results show that our method significantly and consistently outperforms the *FC-2k(En)+Trans.* and *FC-2k(Pseudo)* models in all metrics, despite our unpaired setting is much weaker.

Method	*SG-m*	*S-gate*	B@1	B@4	M	R
G_{EN}	✗	✗	25.0	5.2	14.4	27.3
GM_{BASE}	✓	✗	26.6	7.3	15.1	28.1
GM_{WORD}	✓	✗	28.1	8.0	15.4	28.2
$\text{GM}_{\text{WORD+SUB.}}$	✓	✗	29.2	9.9	16.3	30.2
HGM_{BASE}	✓	✗	29.6	9.9	16.5	30.4
HGM	✓	✓	**30.4**	**11.1**	**17.0**	**31.9**

Table 3: Performance comparison between variants of HGM on Chinese sentence generation task. Test split of MT corpus is used for evaluation. '*SG-m*' is cross-lingual scene graphs mapping. '*S-gate*' is self-gate fusion mechanism.

5.2 Effectiveness of Cross-Lingual Alignment

Analyzing the superior performance of HGM. We conduct experiments on MT task to demonstrate our HGM's effectiveness in cross-lingual alignment, which is shown in Table 3. The advantage of HGM lies in four aspects: (1) The cross-lingual graph translation is effective. Our HGM and its variants achieve considerably higher performance compared with G_{EN}, which directly generates Chinese sentences based on English scene graphs. (2) The cross-lingual alignment in the encoding space is superior than direct symbolic translation. GM_{WORD} achieve considerably higher performance compared with GM_{BASE}, which proves that the scene graph

[3]Code is acquired from the first author of (Gu et al. 2019).

encoding contains richer information and is more suitable for cross-lingual alignment. (3) Performing node mapping considering features across different graph levels boosts the performance. When we consider full-graph and sub-graph level features, the cross-lingual alignment starts achieving significant performance improvement, which verifies the importance of structural and relational information in the context. E.g., HGM outperforms GM_{WORD} in B@1, B@4, METEOR, and ROUGE metrics by 8.2%, 38.8%, 10.4%, 13.1%, respectively. (4) The adaptive self-gate fusion mechanism is beneficial. We can observe that HGM_{BASE} is surpassed by HGM by a large margin. As shown in Table 5, the role of self-gate fusion becomes more essential when HGM is applied to image scene graphs.

Joint training benefits the enconding process. We train our models using the joint loss $\mathcal{L}_{\text{Phase 1}}$, where $\mathcal{L}_{\text{KL}}$ enforces the distributions of latent scene graph embeddings between different languages to be close. Table 4 shows that the models trained with joint loss consistently outperforms their counterparts with only $\mathcal{L}_{\text{XE}}$ for all metrics, which indicates that the encoding process of the target language can benefit from the source language.

Method	$\mathcal{L}_{\text{KL}}$	B@1	B@2	B@4	M	R
GM_{WORD}	✓	28.1	16.2	8.0	15.4	28.2
w/o joint	✗	-0.4	-0.4	-0.3	-0.1	-0.2
$\text{GM}_{\text{WORD+SUB.}}$	✓	29.2	17.9	9.9	16.3	30.2
w/o joint	✗	-0.3	-0.3	-0.4	-0.2	-0.1
HGM	✓	**30.4**	**19.1**	**11.1**	**17.0**	**31.9**
w/o joint	✗	-0.5	-0.3	-0.3	-0.2	-0.4

Table 4: Effectiveness of joint training. Results are report on Chinese sentence generation task (test set).

5.3 Effectiveness of Cross-Modal Mapping

Table 5 shows the performance of Chinese image captioners with and without CMM. We can see that adversarial training can consistently improve the model's performance. Specifically, CMM can boost the performance of our HGM by 3.8%(B@1), 0.7%(B@4), 1.2%(ROUGE), 3.0%(CIDER), respectively. Notably, $\text{GM}_{\text{WORD+SUB.}}$ and HGM_{BASE} perform even worse than GM_{BASE}, which is because the generated image scene graphs are noisy with repeated relation triples (as explained in §5.5), leading to degradation on contextualized cross-lingual graph mapping (sub-graph and full-graph), whereas self-gated fusion can tackle this problem by decreasing the importance of noisy graph-level mapping.

5.4 Human Evaluation

Table 6 shows human evaluation results. The caption quality is measured by *fluency* and *relevancy* metrics. The *fluency* measures whether the generated caption is fluent. The *relevancy* measures whether the caption correctly describes relevant information of the image. Metrics are graded by: 1-Very poor, 2-Poor, 3-Adequate, 4-Good, 5-Excellent. We invite 10 Chinese native speakers from diverse professional backgrounds to participate in the evaluation. Table 6 reports the mean scores, which illustrate that our method can generate relevant and human-like captions.

Method	B@1	B@4	M	R	C
GM_{WORD}	40.1	2.2	15.6	28.4	9.5
GM_{WORD}+CMM	43.1	3.0	16.5	29.4	12.6
$\text{GM}_{\text{WORD+SUB.}}$	37.3	2.5	14.3	27.0	7.9
$\text{GM}_{\text{WORD+SUB.}}$+CMM	40.6	2.8	15.2	28.3	10.8
HGM_{BASE}	38.0	2.4	14.4	27.3	8.0
HGM_{BASE}+CMM	39.8	2.6	14.8	27.7	10.2
HGM	41.1	2.6	15.7	28.4	9.7
HGM+CMM	**44.9**	**3.3**	**16.5**	**29.6**	**12.7**

Table 5: Effectiveness of CMM. Results are reported on test set of COCO-CN. C is short for CIDEr.

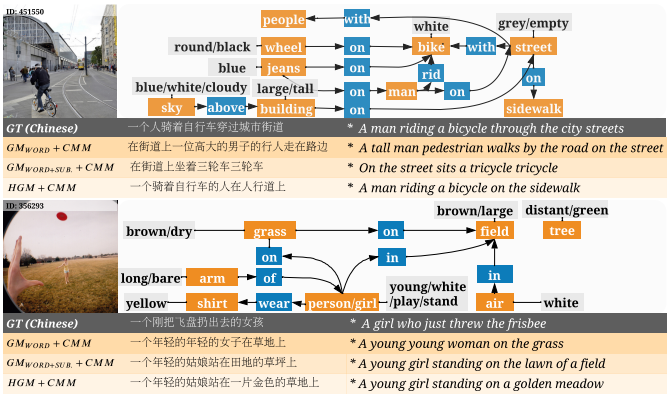


Figure 3: Qualitative results of different unsupervised cross-lingual caption generation models.

Metric	GM_{WORD}	$\text{GM}_{\text{WORD+SUB.}}$	HGM	HGM*	GT
Rel.	2.78	2.96	3.22	3.96	4.86
Flu.	2.49	2.76	3.05	4.06	4.91

Table 6: Human evaluation on COCO-CN test split. HGM* represents fine-tuned HGM. Models are trained with CMM. *Rel.* and *Flu.* is short for relevancy and fluency, respectively.

5.5 Qualitative Results

We provide some Chinese captioning examples for MS-COCO images in Fig. 3. We can see that our method can generate reasonable image descriptions without using any paired image-caption data. Also, we observe that the image scene graphs are quite noisy, which potentially explains the performance degradation when introducing graph-mappings without self-fusion mechanism (see Table 5).

6 Conclusion

In this paper, we propose a novel framework to learn a cross-lingual image captioning model without any image-caption pairs. Extensive experiments demonstrate the effectiveness of our proposed methods. We hope our work can provide inspiration for unpaired image captioning in the future.

Acknowledgments

We would like to thank Lingpeng Kong, Renjie Pi and the anonymous reviewers for insightful suggestions that have significantly improved the paper. This work was supported by TCL Corporate Research (Hong Kong). The research of Philip L.H. Yu was supported by a start-up research grant from the Education University of Hong Kong (#R4162).

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Barrault, L.; Bojar, O.; Costa-Jussà, M. R.; Federmann, C.; Fishel, M.; Graham, Y.; Haddow, B.; Huck, M.; Koehn, P.; Malmasi, S.; et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *WMT*.

Brightmart. 2019. NLP Chinese Corpus: Large Scale Chinese Corpus for NLP. *Zenodo*.

Conneau, A.; Rinott, R.; Lample, G.; Williams, A.; Bowman, S.; Schwenk, H.; and Stoyanov, V. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *EMNLP*.

Denkowski, M.; and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *ACL*.

Eberhard, D. M.; Simons, G. F.; and Fennig, C. D., eds. 2019. *Ethnologue: Languages of the World*. SIL International, 22 edition.

Esplà, M.; Forcada, M.; Ramírez-Sánchez, G.; and Hoang, H. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, 118–119. Dublin, Ireland: European Association for Machine Translation.

Feng, Y.; Ma, L.; Liu, W.; and Luo, J. 2019. Unsupervised image captioning. In *CVPR*.

Gu, J.; Joty, S.; Cai, J.; and Wang, G. 2018. Unpaired image captioning by language pivoting. In *ECCV*.

Gu, J.; Joty, S.; Cai, J.; Zhao, H.; Yang, X.; and Wang, G. 2019. Unpaired image captioning via scene graph alignments. In *ICCV*.

Gu, J.; Kuen, J.; Joty, S.; Cai, J.; Morariu, V.; Zhao, H.; and Sun, T. 2020. Self-supervised relationship probing. *NeurIPS*.

Hu, J.; Ruder, S.; Siddhant, A.; Neubig, G.; Firat, O.; and Johnson, M. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. *CoRR*, abs/2003.11080.

Johnson, J.; Gupta, A.; and Fei-Fei, L. 2018. Image generation from scene graphs. In *CVPR*.

Joulin, A.; Bojanowski, P.; Mikolov, T.; Jégou, H.; and Grave, E. 2018. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In *EMNLP*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*.

Laina, I.; Rupprecht, C.; and Navab, N. 2019. Towards Unsupervised Image Captioning with Shared Multimodal Embeddings. In *ICCV*.

Lan, W.; Li, X.; and Dong, J. 2017. Fluency-guided cross-lingual image captioning. In *ACMMM*.

Li, X.; Xu, C.; Wang, X.; Lan, W.; Jia, Z.; Yang, G.; and Xu, J. 2019. COCO-CN for Cross-Lingual Image Tagging, Captioning, and Retrieval. *TMM*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Ng, N.; Yee, K.; Baevski, A.; Ott, M.; Auli, M.; and Edunov, S. 2019. Facebook FAIR's WMT19 News Translation Task Submission. *arXiv preprint arXiv:1907.06616*.

Nguyen, K.; Tripathi, S.; Du, B.; Guha, T.; and Nguyen, T. Q. 2021. In Defense of Scene Graphs for Image Captioning. In *ICCV*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *CVPR*.

Schuster, S.; Krishna, R.; Chang, A.; Fei-Fei, L.; and Manning, C. D. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *ACL*.

Song, Y.; Chen, S.; Zhao, Y.; and Jin, Q. 2019. Unpaired Cross-lingual Image Caption Generation with Self-Supervised Rewards. In *ACMMM*.

Tian, L.; Wong, D. F.; Chao, L. S.; Quaresma, P.; Oliveira, F.; and Yi, L. 2014. UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation. In *LREC*.

Tran, K.; He, X.; Zhang, L.; Sun, J.; Carapcea, C.; Thrasher, C.; Buehler, C.; and Sienkiewicz, C. 2016. Rich image captioning in the wild. In *CVPRW*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Wang, Y.-S.; Liu, C.; Zeng, X.; and Yuille, A. 2018. Scene graph parsing as dependency parsing. *arXiv preprint arXiv:1803.09189*.

Wu, J.; Zheng, H.; Zhao, B.; Li, Y.; Yan, B.; Liang, R.; Wang, W.; Zhou, S.; Lin, G.; Fu, Y.; et al. 2017. AI Challenger: A Large-scale Dataset for Going Deeper in Image Understanding. *arXiv preprint arXiv:1711.06475*.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yang, X.; Tang, K.; Zhang, H.; and Cai, J. 2019. Auto-encoding scene graphs for image captioning. In *CVPR*.

Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural motifs: Scene graph parsing with global context. In *CVPR*.

Zhong, Y.; Wang, L.; Chen, J.; Yu, D.; and Li, Y. 2020. Comprehensive Image Captioning via Scene Graph Decomposition. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *ECCV*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.