

# Cross-Domain Image Captioning via Cross-Modal Retrieval and Model Adaptation

Wentian Zhao, Xinxiao Wu<sup>✉</sup>, Member, IEEE, and Jiebo Luo

**Abstract**—In recent years, large scale datasets of paired images and sentences have enabled the remarkable success in automatically generating descriptions for images, namely image captioning. However, it is labour-intensive and time-consuming to collect a sufficient number of paired images and sentences in each domain. It may be beneficial to transfer the image captioning model trained in an existing domain with pairs of images and sentences (i.e., source domain) to a new domain with only unpaired data (i.e., target domain). In this paper, we propose a cross-modal retrieval aided approach to cross-domain image captioning that leverages a cross-modal retrieval model to generate pseudo pairs of images and sentences in the target domain to facilitate the adaptation of the captioning model. To learn the correlation between images and sentences in the target domain, we propose an iterative cross-modal retrieval process where a cross-modal retrieval model is first pre-trained using the source domain data and then applied to the target domain data to acquire an initial set of pseudo image-sentence pairs. The pseudo image-sentence pairs are further refined by iteratively fine-tuning the retrieval model with the pseudo image-sentence pairs and updating the pseudo image-sentence pairs using the retrieval model. To make the linguistic patterns of the sentences learned in the source domain adapt well to the target domain, we propose an adaptive image captioning model with a self-attention mechanism fine-tuned using the refined pseudo image-sentence pairs. Experimental results on several settings where MSCOCO is used as the source domain and five different datasets (Flickr30k, TGIF, CUB-200, Oxford-102 and Conceptual) are used as the target domains demonstrate that our method achieves mostly better or comparable performance against the state-of-the-art methods. We also extend our method to cross-domain video captioning where MSR-VTT is used as the source domain and two other datasets (MSVD and Charades Captions) are used as the target domains to further demonstrate the effectiveness of our method.

**Index Terms**—Cross-domain image captioning, cross-modal retrieval, model adaptation.

## I. INTRODUCTION

AUTOMATICALLY generating natural language descriptions for images, i.e. image captioning, has attracted much attention in recent years. Different from other computer vision tasks, image captioning is a multi-modal learning task

Manuscript received May 1, 2020; revised September 28, 2020 and November 1, 2020; accepted November 24, 2020. Date of publication December 11, 2020; date of current version December 17, 2020. This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant 61673062 and Grant 62072041. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dong Tian. (*Corresponding author: Xinxiao Wu.*)

Wentian Zhao and Xinxiao Wu are with the Media Computing and Intelligent Systems Laboratory, Beijing Institute of Technology, Beijing 100081, China (e-mail: wentian\_zhao@bit.edu.cn; wuxinxiao@bit.edu.cn).

Jiebo Luo is with the Department of Computer Science, University of Rochester, Rochester, NY 14627 USA (e-mail: jluo@cs.rochester.edu).

Digital Object Identifier 10.1109/TIP.2020.3042086

that requires both understanding the visual information in the image and generating natural language descriptions that are semantically coherent and syntactically correct. Image captioning can be applied to many scenarios, including content-based image retrieval, visual question answering and visual dialog.

Inspired by the great success of deep neural networks in computer vision [1] and the remarkable performance of encoder-decoder framework in machine translation, many recent studies [2], [3] [4] employ the encoder-decoder framework based on deep neural networks for image captioning, where a convolutional neural network (CNN) serves as the encoder to encode the input images and a recurrent neural network (RNN) is used as the decoder to generate descriptions for the images. However, these image captioning models are trained in a supervised learning scheme that requires large scale image captioning datasets consisting of image-sentence pairs, such as MSCOCO [5] and Flickr30k [6]. Given a new domain, such a supervised learning scheme would be difficult to implement since it is costly to annotate each image with a corresponding sentence. In the mean time, unpaired multimedia data is easy to acquire from the web, including images and text descriptions. Therefore, it would be beneficial to transfer an image captioning model trained in an existing source domain with paired images and sentences to a new target domain with unpaired data, referred to as cross-domain image captioning [7].

In the task of cross-domain image captioning, we are given a source domain with image-sentence pairs and a target domain with unpaired images and sentences. Our goal is to adapt an image captioning model trained on the source domain to the target domain, i.e. the adapted captioning model can describe the images in the target domain by generating sentences that are similar to the sentences in the target domain. It is a challenging task since there exists a large gap between the source and the target domains. This domain gap is not only caused by the appearance variance between source and target images, but also caused by the linguistic difference between source and target domain sentences. For example, as shown in Fig. 1, the source domain (MSCOCO) contains images about realistic scenes as well as the corresponding sentences that describe the salient objects and their relationships. Different from the source domain, the images in the target domain (CUB-200) are about flowers, and the sentences describe the appearance of the flowers in detail. Moreover, the image-sentence pairs are not available in the target domain, further making it difficult to adapt the captioning model to the target domain.

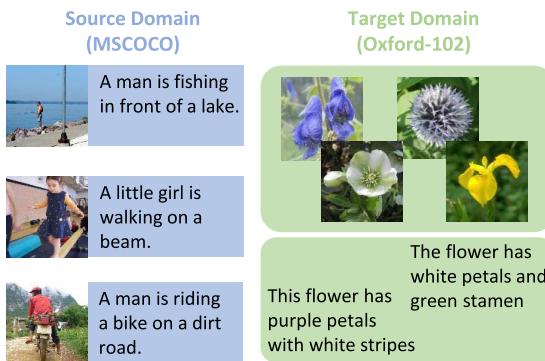


Fig. 1. An example of the images and sentences in the source domain (MSCOCO) and target domain (Oxford-102). The images and sentences in MSCOCO describe realistic scenes, while the images and sentences in Oxford-102 mainly focus on the details of flowers.

In this paper, we propose a novel cross-domain image captioning approach that couples an adaptive image captioning model with a cross-modal retrieval model. The cross-modal retrieval model guides the adaptation of the image captioning model by generating pseudo image-sentence pairs in the target domain. Specifically, we propose an iterative process to discover and refine the pseudo image-sentence pairs. The retrieval model is first pre-trained using the paired data in the source domain to provide a good initialization for the iterative process. The retrieval model is then applied to the unpaired target domain data to generate an initial set of pseudo image-sentence pairs. The initial pseudo image-sentence pairs are further refined by iteratively performing the following two steps: fine-tuning the retrieval model with the pseudo image-sentence pairs, and using the fine-tuned retrieval model to update the pseudo image-sentence pairs. Finally, the refined pseudo image-sentence pairs are used to adapt the image captioning model to the target domain.

In cross-domain image captioning, the sentences in the source domain and the target domain may follow different patterns, which causes difficulty for the adaptation process of the language model. To address this issue, we design an adaptive image captioning model based on a variant of a two-layer LSTM, which better adapts to the language patterns in the target domain. Specifically, two groups of parameters are devised for a traditional LSTM to capture the linguistic patterns of the source and target sentences, respectively. With a self-attention module, the adaptation of the language model is implemented by automatically adjusting the corresponding attention weights of the source and target linguistic patterns.

The contributions of this paper are summarized as follows:

- We propose a cross-modal retrieval guided approach to cross-domain image captioning. The adaptation of the image captioning model is facilitated by the pseudo image-sentence pairs discovered and iteratively refined by a cross-modal retrieval model.
- We propose an adaptive language LSTM that can be effectively adapted from the source domain to the target domain by learning transferable linguistic patterns of the sentences.

- Our method outperforms the state-of-the-art methods across five diverse settings between MSCOCO and other publicly available datasets.

The remainder of this paper is organized as follows. Section II discusses the related work. In Section III, our proposed method is described in detail. In Section IV, we present the experimental settings and the results. Section V makes a conclusion and discusses the future work.

## II. RELATED WORK

### A. Image Captioning

Existing image captioning methods can be roughly divided into two categories: traditional machine learning based methods and deep learning based methods. Traditional machine learning based methods generate captions by either completing pre-defined sentence templates [8], [9] or retrieving existing captions [10]. However, traditional machine learning based methods suffer when generalizing to new images since these methods rely heavily on existing templates or sentences.

In recent years, deep learning based image captioning methods have been extensively studied with superior performance to the traditional methods. Most deep learning based methods [2], [3], [11] follow the encoder-decoder framework, where a CNN is employed as the encoder to extract the vector representation of the input image, and an RNN is used as the decoder to generate word sequence according to the vector representation of the image. Several different attention mechanisms [12]–[14] are proposed to further improve the performance of the encoder-decoder framework. In [13], a two-layer LSTM network is proposed where the first layer calculates the attention weights for different image regions, and the second layer outputs the probability of each word according to the attended image regions. Some methods [15]–[17] attempt to model the relationship between the objects in the images using scene graphs. Motivated by the recent progress in natural language processing, the Transformer model [18] is also applied to image captioning [19], [20]. Feng *et al.* [21] first introduce a new paradigm of unsupervised image captioning, where the image captioning model is trained using unpaired image set and sentence corpus. Several subsequent methods explore using scene graph alignment [22], shared embedding space [23] or memory network [24] to learn from unpaired images and sentences.

These encoder-decoder based models are optimized with the cross-entropy loss function. However, these models are evaluated with non-differentiable metrics including Bleu [25], METEOR [26], ROUGE\_L [27] and CIDEr [28], which are inconsistent with the cross-entropy loss. In addition, the models trained with the cross-entropy loss suffer from the exposure bias [29] problem. To address the above issues, reinforcement learning is applied in some methods [7], [30] [31]. For instance, [30] proposes to optimize the image captioning model using the REINFORCE algorithm with an estimated baseline. The above mentioned methods follow a fully-supervised training scheme, which requires large scale image captioning datasets consisting of image-sentence pairs.

Recently, domain adaptation has also been applied to image captioning. Several methods [32], [33] aim to describe the

objects that are absent in the training data, which is referred to as novel object captioning. Other methods use unpaired images and sentences in the target domain for cross-domain image captioning. An adversarial training procedure is proposed in [7] for cross-domain image captioning model. Two different critics, namely the domain critic and the multimodal critic, serve as the discriminators and the image captioning model serves as the generator. [34] and [35] propose a dual-learning mechanism to learn the knowledge in unpaired images and sentences, which consists of an image captioning model and an image synthesis model. In these methods, image captioning and image synthesis are simultaneously optimized via the dual learning mechanism, which could enhance the performance of image captioning in the target domain. Different from the aforementioned methods that utilize unpaired images and sentences in the target domain separately, our method exploits the intrinsic semantic correlation between images and sentences in the target domain with the help of a cross-modal retrieval model.]

### B. Cross-Modal Retrieval

The task of cross-modal retrieval focuses on retrieving the most relevant instances in one modality for the query in another modality. Existing cross-modal retrieval methods can be generally divided into binary representation learning methods and real-valued representation learning methods. Binary representation learning is also termed as cross-modal hashing, which aims to map the representation of samples in different modalities into a common Hamming space [36]–[39]. Real-valued representation learning methods expect to learn a common latent space where the distance of samples in different modalities can be directly measured [40]–[42]. According to the information utilized to learn the common latent space, the real-valued retrieval methods can be further categorized as: (1) supervised methods that utilize the label information [43], (2) unsupervised methods that learn from co-occurrence information [40]–[42], [44], (3) rank based methods that utilize the rank lists, and (4) pairwise based methods that learn from similar sample pairs in different domains. The cross-modal retrieval method involved in our framework belongs to the unsupervised methods.

In recent years, the task of natural language moment retrieval has attracted more and more attention. Given an untrimmed video and a natural language query, this task focuses on retrieving the video segments that are most relevant to the query. The pioneer of such methods [45] learn to project the video segments and the language queries to a common embedding space. To fully exploit the relationship between natural language and visual content, an iterative graph adjustment method [46] as well as the methods that fuse visual and textual features using cross-modal interaction modules [47]–[49] are proposed. Most recently, Chen *et al.* [50] propose to conduct fine-grained video-text matching by first constructing hierarchical semantic role graphs for sentences and then reasoning over the graph. Some moment retrieval methods [51], [52] attempt to utilize the weakly annotated data, i.e. the training data only contains pairs of untrimmed videos and video-level sentence annotations and the temporal

boundary of the video segments are unknown. Compared to the existing cross-modal retrieval methods that are trained using the image-sentence pairs, the cross-modal retrieval model in our method utilizes both the image-sentence pairs in the source domain and the unpaired images and sentences in the target domain. With the help of an iterative refining process, the retrieval model is able to discover the semantic correlation between the images and sentences in the target domain.

### C. Visual Question Answering

Visual question answering (VQA) is an important research topic in the field of vision and language. It aims at answering questions about visual content, including images [53], videos [54]–[57] or personal albums [58]. To answer complex questions that require effective reasoning, some methods [53], [59] incorporate external knowledge into VQA models. The inverse visual question answering (iVQA) task [60], which focuses on generating reasonable questions for image-answer pairs, is proposed to diagnose the existing VQA models. Both image captioning and the iVQA task take images as input and generates natural language. However, image captioning aims at describing the salient objects and their relationships in the image, while the output of iVQA is conditioned on both the image and the answer.

Most recently, some methods [61], [62] attempt to narrow the modality gap between visual content and natural language questions by generating dense image captions using pre-trained captioning models. However, the domain gap between the dense captioning dataset used for pre-training and the VQA dataset may affect the quality of the generated sentences and degrade the VQA model's performance. For instance, in [62], the appearance of images in Visual Genome significantly differs from the video frames in the TVQA dataset. In this case, it is favorable to adapt a pre-trained captioning model to the existing VQA datasets, since the adapted model generates sentences of higher quality without using additional paired caption annotations.

## III. OUR METHOD

### A. Problem Formulation

For the cross-domain image captioning, we are given the source domain data  $D_s = \{x_i^s, y_i^s\}_{|i|}$  with  $x_i^s$  representing the  $i$ -th image and its corresponding sentence  $y_i^s$  describing  $x_i^s$ . Let  $I_s = \{x_i^s\}_{|i|}$  and  $S_s = \{y_i^s\}_{|i|}$  denote the source image set and the source sentence set, respectively. In the target domain, we are given two separate sets: a set of images  $I_t = \{x_i^t\}_{|i|}$  and a set of sentences  $S_t = \{y_i^t\}_{|i|}$ . Each sentence  $y$  with length  $T$  is represented by a sequence of words, i.e.,  $y = [w_1, w_2, \dots, w_T]$ . The vocabulary used by the captioning model is defined by  $V_{s,t} = \{w | w \in y, y \in S_s \cup S_t\}$  that contains the words in the source domain as well as the words in the target domain.

### B. Overview

To leverage the unpaired images and sentences in the target domain, we propose a novel approach consisting of an adaptive image captioning model and a cross-modal retrieval model. These two models are first pre-trained using the paired data

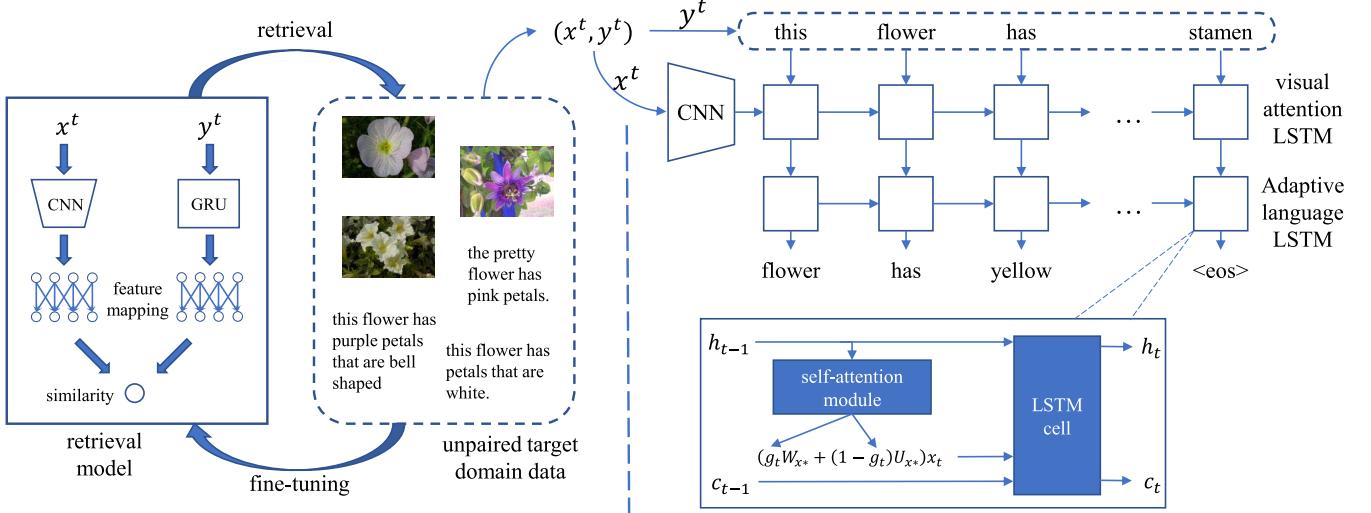


Fig. 2. Overview of the proposed method. The left part illustrates the cross-modal retrieval model, and the right part illustrates the adaptive LSTM model. The refined pseudo image-sentence pairs generated by the retrieval model guides the adaptation process of the adaptive LSTM model.

$D_s$  in the source domain. We then acquire a set of pseudo image-sentence pairs in the target domain using an iterative algorithm. In the  $i$ -th iteration, the cross-modal retrieval model is applied to the unpaired target domain data to obtain pseudo image-sentence pairs  $\hat{D}_t^i$ . The pseudo image-sentence pairs are used to fine-tune the retrieval model, which is applied to the target domain data in the next iteration. Finally, the adaptive image captioning model is fine-tuned using the final pseudo image-sentence pair set, denoted as  $\hat{D}_t^K$ . An overview of our proposed model is shown in Fig. 2.

### C. Cross-Modal Retrieval Model

Our cross-modal retrieval first maps both the images and the sentences to fixed-dimension vectors. Let  $\hat{s}_x$  and  $\hat{s}_y$  denote the feature representations of image  $x$  and sentence  $y$ , respectively. We use the Resnet-101 model pre-trained on ImageNet to encode the image  $x$  into a  $14 \times 14 \times 512$  feature map that contains  $L$  vectors, where  $L = 196$ . We denote these feature vectors as  $F = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_L\}$ , where  $\mathbf{v}_i \in \mathbb{R}^D$  and  $D = 512$ . Each feature  $\mathbf{v}_i$  corresponds to an image region divided by a  $14 \times 14$  grid. The sentence  $y = [w_1, w_2, \dots, w_T]$  is first embedded into a sequence of vectors  $[\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_T]$  with a word embedding matrix  $\mathbf{W}_e^r$ , and the word embeddings are then used as the input of a single-layer gated recurrent unit (GRU) network. The output of the GRU network at the last time step is used as the feature representation  $\hat{s}_y$  of the sentence.

In order to establish the connection between the source domain and the target domain, we propose domain-shared latent attributes to learn the common representation between the two domains. The images and sentences in the target domain are represented by both the domain-shared latent attributes and the domain-specific latent attributes that only appear in the target domain. We represent each latent attribute using a dictionary atom. In the pre-training process, the retrieval model learns the domain-shared attributes by reconstructing the features of the source domain images and sentences using the linear combinations of the dictionary atoms. In the model adaptation process, the retrieval model

only has to learn the domain-specific attributes by reconstructing the images and sentences in the target domain. We use a dictionary  $\mathbf{M}_{s,i} \in \mathbb{R}^{d \times n}$  with  $n$  atoms to encode the attributes in the source domain images, and another dictionary  $\mathbf{M}_{t,i} \in \mathbb{R}^{d \times n}$  is used to encode the domain-specific attributes in the target domain images. The process of reconstructing the feature of the source domain image  $x^s$  can be formulated as

$$\begin{aligned}\hat{\alpha}_s &= \mathbf{M}_{s,i}^\top \mathbf{W}_i \hat{s}_{x^s}, \\ \alpha_s &= \text{softmax}(\hat{\alpha}_s), \\ s_{x^s} &= \mathbf{M}_{s,i} \hat{\alpha}_s,\end{aligned}\quad (1)$$

where  $\mathbf{W}_i \in \mathbb{R}^{d \times d}$  is a mapping matrix and  $\alpha_s \in \mathbb{R}^n$  denotes the weights of the dictionary atoms. The feature  $s_{y^s}$  of the source domain sentence  $y^s$  is reconstructed in a similar manner with another dictionary  $\mathbf{M}_{s,s} \in \mathbb{R}^{d \times n}$ . The process of reconstructing the feature of a target domain image  $x^t$  can be formulated as

$$\begin{aligned}\hat{\alpha}_t &= [\mathbf{M}_{s,i}; \mathbf{M}_{t,i}]^\top \mathbf{W}_i \hat{s}_{x^s}, \\ \alpha_t &= \text{softmax}(\hat{\alpha}_t), \\ s_{x^t} &= [\mathbf{M}_{s,i}; \mathbf{M}_{t,i}] \hat{\alpha}_t,\end{aligned}\quad (2)$$

where  $\alpha_t \in \mathbb{R}^{2n}$  denotes the weights of dictionary atoms in the joint dictionary of  $\mathbf{M}_{s,i}$  and  $\mathbf{M}_{t,i}$ , and the operator  $[;]$  denotes matrix concatenation. The feature  $s_{y^t}$  of target domain sentence  $y^t$  is reconstructed with two dictionaries, namely  $\mathbf{M}_{s,s}$  and  $\mathbf{M}_{t,s} \in \mathbb{R}^{d \times n}$ . We calculate the similarity between the image  $x$  and the sentence  $y$  using the reconstructed features:

$$\text{sim}(x, y) = \frac{s_x \cdot s_y}{\|s_x\| \cdot \|s_y\|}, \quad (3)$$

where the operator  $\|\cdot\|$  denotes the L2 norm of vectors. During pre-training, the source domain dictionaries and the mapping matrices  $\theta_s^r = \{\mathbf{M}_{s,i}, \mathbf{M}_{s,s}, \mathbf{W}_i, \mathbf{W}_s\}$  are learned. In the fine-tuning process, the source domain parameters  $\theta_s^r$  are fixed and the target domain dictionaries  $\theta_t^r = \{\mathbf{M}_{t,i}, \mathbf{M}_{t,s}\}$  are updated to learn the target domain attributes.

#### D. Adaptive Image Captioning Model

An adaptive image captioning model based on the encoder-decoder framework is proposed to bridge the gap between different domains. Since the source domain sentences differ from the target domain sentences in many aspects (e.g. the word usage and the syntactic structure), it is crucial for the captioning model to transfer the linguistic patterns learned in the source domain to the target domain. To this end, we design an adaptive captioning model that encodes the knowledge about the sentences in the two domains using two groups of parameters and weighs these parameters for knowledge adaptation with the help of an attention mechanism.

Inspired by [13], we devise a new variant of a two-layer LSTM as the decoder to generate sentences. The first LSTM layer acts as a visual attention model to weigh each feature. The second LSTM layer is characterized as an adaptive language model to transfer the language patterns learned in the source domain to the target domain.

The input to the visual attention LSTM at each time step  $t$  is the concatenation of the previous output of the adaptive language LSTM, an encoding of the average-pooled image feature  $\bar{\mathbf{v}} = \frac{1}{L} \sum_{i=1}^L \mathbf{v}_i$  and an encoding of the previously generated word:

$$\mathbf{x}_t^1 = [\mathbf{h}_{t-1}^2; \mathbf{W}_x^c \bar{\mathbf{v}} + \mathbf{b}_x^c; \mathbf{W}_e^c \Pi_{t-1}], \quad (4)$$

where  $\mathbf{W}_x^c \in \mathbb{R}^{H \times D}$  and  $\mathbf{b}_x^c \in \mathbb{R}^{H \times 1}$  denotes the weight and bias of the linear transform applied to  $\bar{\mathbf{v}}$ .  $\mathbf{W}_e^c$  represents the word embedding matrix and  $\Pi_{t-1}$  is a one-hot vector representing the previously generated word. The output of the visual attention LSTM is given by

$$\mathbf{h}_t^1 = \text{LSTM}^1(\mathbf{h}_{t-1}^1, \mathbf{x}_t^1). \quad (5)$$

At each time step  $t$ , the normalized weight  $\alpha_{i,t}$  for each image feature  $\mathbf{v}_i$  is generated by

$$\begin{aligned} \hat{\alpha}_{i,t} &= \boldsymbol{\omega}_a \tanh(\mathbf{W}_{va} \mathbf{v}_i + \mathbf{W}_{ha} \mathbf{h}_t^1) \\ \alpha_{i,t} &= \frac{\exp(\hat{\alpha}_{i,t})}{\sum_{j=1}^L \exp(\hat{\alpha}_{j,t})}, \end{aligned} \quad (6)$$

where  $\boldsymbol{\omega}_a \in \mathbb{R}^{1 \times H}$ ,  $\mathbf{W}_{va} \in \mathbb{R}^{H \times D}$  and  $\mathbf{W}_{ha} \in \mathbb{R}^{H \times H}$  are learned parameters. The weighted sum of the image features  $\hat{\mathbf{v}}_t = \sum_{i=1}^L \alpha_{i,t} \mathbf{v}_i$  and the output  $\mathbf{h}_t^1$  of the visual attention LSTM are concatenated as the input to the adaptive language LSTM:

$$\mathbf{x}_t^2 = [\mathbf{h}_t^1; \hat{\mathbf{v}}_t]. \quad (7)$$

The output of the adaptive language LSTM is given by

$$\mathbf{h}_t^2 = \text{LSTM}^2(\mathbf{h}_{t-1}^2, \mathbf{x}_t^2). \quad (8)$$

Different from the language LSTM in the Top-Down model [13], our adaptive language LSTM incorporates two groups of parameters to capture the knowledge from the sentences in the source domain and the target domain. Specifically, two groups of weight matrices, denoted by  $\{\mathbf{W}_{x*}\}$  and  $\{\mathbf{U}_{x*}\}$ , are designed to learn the linguistic patterns of the sentences in the source domain and the target domain, respectively. An additional attention mechanism calculates the weights  $g_t$

and  $1 - g_t$  for these parameters at the  $t$ -th time step, where  $g_t \in (0, 1)$ . The  $t$ -th word in the sentence is more likely to be related to the linguistic patterns of the source domain sentences when the value of  $g_t$  is close to 0, and vice versa. Compared to the top-down attention mechanism that is used to attend to the crucial parts of images, the attention mechanism in the adaptive LSTM determines whether the model should use more knowledge from the source domain or the target domain when predicting a word. Formally, the adaptive language LSTM is defined as

$$\begin{aligned} i &= \text{sigmoid}((g_t \mathbf{W}_{xi} + (1 - g_t) \mathbf{U}_{xi}) \mathbf{x}_t^2 + \mathbf{W}_{hi} \mathbf{h}_{t-1}^2 + \mathbf{b}_i) \\ f &= \text{sigmoid}((g_t \mathbf{W}_{xf} + (1 - g_t) \mathbf{U}_{xf}) \mathbf{x}_t^2 + \mathbf{W}_{hf} \mathbf{h}_{t-1}^2 + \mathbf{b}_f) \\ o &= \text{sigmoid}((g_t \mathbf{W}_{xo} + (1 - g_t) \mathbf{U}_{xo}) \mathbf{x}_t^2 + \mathbf{W}_{ho} \mathbf{h}_{t-1}^2 + \mathbf{b}_o) \\ \hat{c}_t &= \tanh((g_t \mathbf{W}_{xc} + (1 - g_t) \mathbf{U}_{xc}) \mathbf{x}_t^2 + \mathbf{W}_{hc} \mathbf{h}_{t-1}^2 + \mathbf{b}_c) \\ c_t &= f \odot c_{t-1} + i \odot \hat{c}_t \\ h_t &= o \odot \tanh(c_t), \end{aligned} \quad (9)$$

where  $\odot$  denotes element-wise multiplication, the parameters  $\theta_s^c = \{\mathbf{W}_{i*}, \mathbf{W}_{h*}, \mathbf{b}_*\}$  represent the source domain language LSTM and the parameters  $\theta_t^c = \{\mathbf{U}_{i*}, \mathbf{W}_{i*}, \mathbf{W}_{h*}, \mathbf{b}_*\}$  represent the target domain language LSTM. The attention weight  $g_t$  is calculated by

$$\begin{aligned} \hat{g}_t &= \text{ReLU}(\mathbf{W}_{g1} \mathbf{h}_{t-1} + \mathbf{b}_{g1}) \\ g_t &= \text{sigmoid}(\mathbf{W}_{g2} \hat{g}_t + \mathbf{b}_{g2}). \end{aligned} \quad (10)$$

Note that at each time step  $t$ , the value of  $g_t$  is different. During pre-training, the value of  $g_t$  is fixed to 1 and the parameters  $\theta_s^c = \{\mathbf{W}_{i*}, \mathbf{W}_{h*}, \mathbf{b}_*\}$  are learned. The parameters  $\theta_t^c = \{\mathbf{U}_{i*}, \mathbf{W}_{i*}, \mathbf{W}_{h*}, \mathbf{b}_*\}$  as well as the parameters  $\mathbf{W}_{g*}, \mathbf{b}_{g*}$  in the attention submodule are learned in the process of fine-tuning. Intuitively,  $\theta_s^c$  captures the semantic relationship between images and words, while  $\theta_t^c$  learns the specific linguistic pattern in the target domain.

Finally, the probability of words in the vocabulary at time step  $t$  is calculated as

$$p(w_t | w_1, w_2, \dots, w_{t-1}) = \text{softmax}(\mathbf{W}_p \mathbf{h}_t^2 + \mathbf{b}_p), \quad (11)$$

where  $\mathbf{W}_p$  and  $\mathbf{b}_p$  represent the learned parameters.

#### E. Model Pre-Training

In the process of model pre-training, the cross-modal retrieval model and the adaptive image captioning model are trained using the paired training data in the source domain by maximizing the probability of the ground truth captions and minimizing the distance between the images and the corresponding sentences, respectively. Given an image  $x$  and its corresponding ground truth sentence  $y = [w_1, w_2, \dots, w_T]$ , the captioning model with parameters  $\theta^c$  is optimized by minimizing the cross-entropy loss function:

$$L_c = - \sum_{t=1}^T \log p_{\theta^c}(w_t | w_1, w_2, \dots, w_{T-1}). \quad (12)$$

When training the cross-modal retrieval model, we consider negative samples that are the most similar to the queries, namely hard negatives, following the practice in [42].

In the two processes of querying images with sentences and querying sentences with images, the hard negative samples are given by  $x' = \operatorname{argmax}_{\hat{x}} \operatorname{sim}(\hat{x}, y)$ ,  $\langle \hat{x}, y \rangle \notin D_s$  and  $y' = \operatorname{argmax}_{\hat{y}} \operatorname{sim}(x, \hat{y})$ ,  $\langle x, \hat{y} \rangle \notin D_s$ , respectively. Accordingly, the loss for training the retrieval model is defined as

$$L_r = \max_{x'} (\delta + \operatorname{sim}(x', y) - \operatorname{sim}(x, y))_+ + \max_{y'} (\delta + \operatorname{sim}(x, y') - \operatorname{sim}(x, y))_+, \quad (13)$$

where  $(x)_+ = \max(x, 0)$  and  $\delta$  is a tunable hyper parameter.

#### F. Model Adaptation

The model adaptation process aims to adapt an image captioning model pre-trained in the source domain well to the target domain by fine-tuning the captioning model with the pseudo image-sentence pairs generated by the retrieval model. In order to further improve the performance of the image captioning model in the target domain, a policy-gradient based training algorithm is employed during fine-tuning. We show the processes of generating pseudo image-sentence pairs and fine-tuning the image captioning model in this section.

1) *Generating Pseudo Image-Sentence Pairs*: To adapt the image captioning model to the target domain, we leverage the pre-trained cross-modal retrieval model to generate a set of pseudo image-sentence pairs. The retrieval is performed in two directions, namely retrieving sentences with images and retrieving images with sentences. Specifically, we propose the following iterative algorithm to acquire the pseudo image sentence pairs:

- The pre-trained cross-modal retrieval model is applied to the unpaired target domain data to acquire an initial set of pseudo image-sentence pairs, denoted as  $\hat{D}_t^0$ .
- In the  $i$ -th iteration, the retrieval model is fine-tuned using the current set of pseudo image sentence pairs  $\hat{D}_t^i$ . The loss function for fine-tuning the retrieval model is defined in Eq. 13.
- The fine-tuned retrieval model is applied to the target domain data to acquire an updated set of pseudo image-sentence pairs, denoted as  $\hat{D}_t^{i+1}$ . For each query image (or sentence), the top  $k$  most similar sentences (or images) are used to construct the pseudo image-sentence pairs.
- Repeat the previous two steps for  $P$  times.

The pseudo image-sentence pairs  $\hat{D}_t^P$  acquired in the final iteration are used to fine-tune the image captioning model.

2) *Fine-Tuning Image Captioning Model*: In the first few epochs, the cross-entropy loss function  $L_c$  is used to fine-tune the captioning model with the pseudo paired data  $\hat{D}_t$ . In the next epochs, the captioning model is fine-tuned with a policy-gradient based algorithm to better learn the linguistic patterns of sentences in the target domain. The captioning model is regarded as an *agent*, which interacts with external *environment* composed of the input image and previously generated words. Each generated word is considered as an *action*. Upon generating a whole sentence, the *agent* receives a *reward*, denoted by  $r$ , indicating the quality of the generated sentence. For a sampled sentence  $\hat{y}$ , the reward  $r(\hat{y})$  is

---

**Algorithm 1** Cross-Modal Retrieval Guided Cross-Domain Image Captioning

---

**Input:** source domain paired data  $D_s$ , target domain images  $I_t$ , target domain sentences  $S_t$   
**Output:** parameters of image captioning model  $\theta_c$ , parameters of cross-modal retrieval model  $\theta_r$

```

1: procedure PRE-TRAIN( $D_s, \theta_s^c, \theta_s^r$ )
2:   for  $x_i^s, y_i^s$  in  $D_s$  do
3:     optimize  $\theta_s^c$  with Eq.12
4:     optimize  $\theta_s^r$  with Eq.13
5:   end for
6:   return  $\theta_s^c, \theta_s^r$ 
7: end procedure
8: procedure RETRIEVAL( $I_t, S_t, \theta^r$ )
9:   for  $i$  in  $0, 1, \dots, P$  do
10:    retrieve  $\hat{D}_t^i$  on  $I_t$  and  $S_t$  using  $\theta^r$ 
11:    for  $x^t, y^t$  in  $\hat{D}_t^i$  do
12:      optimize  $\theta_t^r$  with Eq.13
13:    end for
14:    fine-tune  $\theta^r$  with  $\hat{D}_t^i$ 
15:   end for
16:   return  $\hat{D}_t^P$ 
17: end procedure
18: for  $i = 1, 2, \dots, M$  do
19:   optimize  $\theta_s^c$  and  $\theta_s^r$  with PRE-TRAIN( $D_s, \theta_s^c, \theta_s^r$ )
20: end for
21:  $\hat{D}_t^P \leftarrow$  RETRIEVAL( $I_t, S_t, \theta^r$ )
22: for  $i = 1, 2, \dots, N$  do
23:   optimize  $\theta_t^c$  with Eq.12 on  $\hat{D}_t^P$ 
24: end for
25: while  $\theta_c$  not converged do
26:   optimize  $\theta_t^c$  with Eq.16 on  $\hat{D}_t^P$ 
27: end while

```

---

calculated with two evaluation metrics:

$$r(\hat{y}) = \frac{\text{Bleu4}(\hat{y}) + \text{CIDEr}(\hat{y})}{2}. \quad (14)$$

During training, the following negative expected long-term reward is minimized:

$$J(\theta_c) = -\mathbb{E}(r(\hat{y})), \hat{y} \sim p_{\theta_c}. \quad (15)$$

Following [30], the gradient is approximated by

$$\nabla_{\theta_c} J(\theta_c) \approx -(r(\hat{y}) - r(y^*)) \nabla_{\theta_c} \log_{\theta_c}(\hat{y}), \quad (16)$$

where  $\hat{y}$  is a sentence sampled using  $\theta_c$  and  $y^*$  is a sentence obtained by greedy decoding, whose score serves as a baseline. The whole training process is summarized in algorithm 1.

## IV. EXPERIMENTAL SETUP

### A. Datasets

In our experiments, the MSCOCO dataset [5] is used as the source domain, while the Flickr30k dataset [6], the TGIF dataset [63], the CUB-200 dataset [64], the Oxford-102 dataset [65] and a newly collected “Conceptual” dataset are used as target domains. In general, it is believed that Flickr30k and TGIF have a moderate domain gap from MSCOCO, while

CUB-200, Oxford-102 and Conceptual have a larger domain gap from MSCOCO. The details about the datasets are as follows:

- **MSCOCO:** The MSCOCO dataset includes 117,283 images in total, each annotated with 5 manually written sentences. For fair comparison with the baseline methods, we adopt the data split in [3] where the training, validation and testing splits include 82,783 images, 5,000 images and 5,000 images, respectively.
- **Flickr30k:** There are 31,783 images in the Flickr30k dataset and each image is also annotated with 5 sentences. The data split is adopted from [3], where the training split and testing split includes 29,000 images and 1,000 images, respectively.
- **CUB-200:** The CUB-200 dataset includes 6,033 images of birds in total, and the sentence annotation for images are adopted from [66]. We follow the data split in [7], where 4,000 images are used for training and 2,033 images are used for testing.
- **Oxford-102:** The Oxford-102 dataset includes 8,189 images of flowers in total and the sentences are also adopted from [66]. We adopt the same data splitting as in [7], where 7,189 images are used for training and 1,000 images are used for testing.
- **TGIF:** The TGIF dataset contains 100,000 animated GIF images, where 90,000 images are for training and 10,000 images are for testing. Following the strategy in [7], we sample the first frame of each GIF image as input.
- **Conceptual:** To further evaluate the performance of our method, we conduct an additional experiment on a target domain denoted as “Conceptual”, where the images and sentences are unpaired and from different sources. Specifically, the target domain images are from the Conceptual Captions dataset [67], which are collected from various websites. The sentences are from the Shutterstock Image Description Corpus [21], which is harvested from the Shutterstock website. For training, we randomly sample 30,000 images and 30,000 sentences from the Conceptual Captions dataset and the Shutterstock Image Description Corpus, respectively. We test our model using the validation set of the Conceptual Captions dataset, which contains 28,355 images and each image is annotated with one sentence.

### B. Baseline Methods

To evaluate the effectiveness of our method, we compare our proposed method with the following cross-domain image captioning methods. Both our method and the comparison methods are first pre-trained on the source domain and then fine-tuned on the target domain.

- **Source Pre-trained:** The captioning model that is only pre-trained on the source domain is directly evaluated on the test set of the target domain.
- **DCC [32]:** This model combines a lexical classifier based on CNN and a language LSTM trained on unpaired text with a fully connected layer.
- **SAdT [7]:** This model performs cross-domain captioning by utilizing adversarial learning. A domain critic is

designed to assess whether the generated sentences are indistinguishable from the target domain sentences and a multi-modal critic is designed to evaluate the semantic consistency between input image and generated sentence.

- **DL [34]:** A dual learning mechanism is designed to utilize unpaired training data. Two objectives are simultaneously optimized: generating descriptions from images and generating images from text.
- **MLADIC [35]:** An improved version of [34] which also employs the dual learning mechanism. The policy gradient method is further refined in MLADIC.
- **Graph-Enc-Dec [22]:** An unsupervised image captioning method that learns to align the scene graphs extracted from images and sentences. This method includes two feature mapping functions that map features from one modality to another modality and two discriminators that distinguish the real features from the mapped features. For fair comparison, we pre-train the feature mapping functions and the discriminators using the source domain paired data, and then fine-tune the model using unpaired target domain data.
- **Paired:** The image captioning model in our method is fine-tuned directly using the paired training data in the target domain. This serves as an empirical upper bound of our experiments (shown in rows with a gray background). Note that the actual obtained performance numbers sometimes are not the highest among the results because using paired training data only theoretically should lead to the upper bound.

### C. Evaluation Metrics

We adopt the evaluation metrics that are widely used in previous image captioning methods [2]–[4], including Bleu-n [25], METEOR [26], ROUGE\_L [27] and CIDEr [28]. Bleu-n calculates the fraction of n-grams of the candidate sentence that appear in reference sentences, and the value of n varies from 1 to 4 in our method. METEOR evaluates a candidate sentence by calculating several scores between words and phrases in the candidate sentence and the reference sentences. ROUGE\_L is based on the longest common subsequence between a candidate sentence and the reference sentences. CIDEr is specifically designed for evaluating image descriptions and calculates the relevancy between the candidate sentence and the reference sentences using human consensus.

### D. Implementation Details

**1) Image Feature Extraction:** The ResNet-101 model [68] pre-trained on ImageNet is used to extract the representation of images. The image is used as the input of the CNN without re-sizing or cropping, and the output of the last convolutional layer is further processed for different tasks. We perform average pooling over all spatial locations of the feature map to obtain a 2,048-dimensional feature vector. For the image captioning task, adaptive average pooling is applied to obtain a feature map with the size of  $14 \times 14 \times 2048$ .

**2) Corpus and Language Model:** The sentences of the source domain and the target domains are pre-processed following [3]. Words appearing less than 5 times are replaced

TABLE I

QUANTITATIVE EVALUATION RESULTS OF CROSS-DOMAIN IMAGE CAPTIONING ON DIFFERENT TARGET DOMAINS. NOTE THAT THE ACTUAL OBTAINED PERFORMANCE NUMBERS WITH PAIRED TRAINING DATA (IN ROWS WITH A GRAY BACKGROUND) SOMETIMES ARE NOT THE HIGHEST AMONG THE RESULTS BECAUSE USING PAIRED TRAINING DATA ONLY THEORETICALLY SHOULD LEAD TO THE UPPER BOUND

Method	Target	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE_L	CIDEr
<i>Source Pre-trained</i>	Flickr30k	60.6	41.1	26.5	17.8	16.1	42.2	30.5
DCC [32]	Flickr30k	54.3	34.6	21.8	13.8	16.1	38.8	27.7
SADT [7]	Flickr30k	62.1	41.7	27.6	17.9	16.7	42.1	32.6
DL [34]	Flickr30k	63.9	44.1	31.8	17.3	16.3	44.2	33.6
MLADIC [35]	Flickr30k	68.2	45.4	32.7	18.7	17.5	45.3	50.5
Graph-Enc-Dec [22]	Flickr30k	66.0	46.1	31.5	17.0	17.7	44.3	45.7
<i>Ours</i>	Flickr30k	<b>69.0</b>	<b>49.3</b>	<b>34.7</b>	<b>24.1</b>	<b>19.5</b>	<b>46.5</b>	<b>52.8</b>
<i>Paired (upper bound)</i>	Flickr30k	66.6	48.4	34.5	24.5	20.3	46.5	53.3
<i>Source Pre-trained</i>	Oxford-102	85.5	75.3	68.4	61.4	37.3	70.2	39.5
DCC [32]	Oxford-102	51.0	33.8	24.1	16.7	21.5	38.3	6.0
SADT [7]	Oxford-102	85.6	76.9	67.4	60.5	36.4	72.1	29.3
DL [34]	Oxford-102	91.2	84.4	77.1	71.6	43.0	82.4	79.7
MLADIC [35]	Oxford-102	92.5	85.6	78.2	73.5	<b>46.1</b>	<b>84.5</b>	<b>90.6</b>
Graph-Enc-Dec [22]	Oxford-102	93.4	87.2	79.9	75.4	40.5	76.0	80.8
<i>Ours</i>	Oxford-102	<b>96.9</b>	<b>91.8</b>	<b>86.0</b>	<b>80.2</b>	42.2	77.8	87.3
<i>Paired (upper bound)</i>	Oxford-102	96.8	92.1	86.6	81.3	43.8	80.1	87.5
<i>Source Pre-trained</i>	CUB-200	82.1	66.1	51.7	40.1	27.4	56.8	28.1
DCC [32]	CUB-200	68.6	47.3	31.4	21.4	23.8	46.4	11.9
SADT [7]	CUB-200	91.4	73.1	51.9	32.8	27.6	58.6	24.8
Graph-Enc-Dec [22]	CUB-200	90.8	78.7	65.3	52.7	35.2	67.0	68.2
<i>Ours</i>	CUB-200	<b>95.3</b>	<b>83.9</b>	<b>72.0</b>	<b>61.6</b>	<b>36.6</b>	<b>69.3</b>	<b>76.7</b>
<i>Paired (upper bound)</i>	CUB-200	91.3	82.9	75.7	66.8	39.5	74.9	78.0
<i>Source Pre-trained</i>	TGIF	38.7	20.6	11.1	6.4	13.2	32.8	15.6
DCC [32]	TGIF	34.6	17.5	9.3	4.1	11.8	29.5	7.1
SADT [7]	TGIF	47.5	29.2	17.9	10.3	14.5	37.0	22.2
Graph-Enc-Dec [22]	TGIF	50.1	31.4	18.6	11.5	15.6	39.5	35.5
<i>Ours</i>	TGIF	<b>53.8</b>	<b>34.3</b>	<b>21.1</b>	<b>12.7</b>	<b>16.7</b>	<b>40.0</b>	<b>37.2</b>
<i>Paired (upper bound)</i>	TGIF	52.9	34.7	21.7	13.3	17.0	39.7	39.6
<i>Source Pre-trained</i>	Conceptual	12.1	5.2	2.7	0.8	10.3	16.2	8.4
<i>Ours</i>	Conceptual	<b>13.2</b>	<b>6.4</b>	<b>3.6</b>	<b>1.7</b>	<b>12.7</b>	<b>16.5</b>	<b>11.2</b>

with a special token. To simplify the implementation, a joint vocabulary containing words in both the source domain and the target domain is used. The word embedding vectors are set to 300 dimension and initialized randomly.

3) *Model Pre-Training Details*: The adaptive image captioning model is pre-trained on the source domain by Adam optimizer [69] with the learning rate of  $2 \times 10^{-3}$ . Dropout is applied with a dropout probability of 0.5 to avoid overfitting. The value of hyper parameter  $\delta$  in Eq. 13 is set to 0.2, following [42].

4) *Model Adaptation Details*: When fine-tuning the image captioning model using the pseudo image-sentence pairs in the target domain, the Adam optimizer [69] is applied and the learning rate is set to  $5 \times 10^{-4}$ . The number of top retrieving results (i.e.,  $k$ ) in both retrieving directions is set to 5 for the best performance. The detailed analysis of this parameter will be given in Section 4.4.

#### E. Cross-Domain Image Captioning

We show the automatic evaluation results of cross-domain image captioning in Table I. From the results demonstrated in Table I, we can draw the following observations:

- Our method achieves mostly better or comparable results for most evaluation metrics on all four target domain datasets, which clearly demonstrates the benefit of designing a cross-modal retrieval model to generate pseudo paired target data as an auxiliary for cross-domain image captioning.

- From Table I, we observe that our method performs slightly inferior to “DL” and “MLADIC” on the Oxford-102 dataset in terms of METEOR, ROUGE\_L and CIDEr. The large domain gap between MSCOCO and Oxford-102 may lead to the pseudo image-sentence pairs that are less accurate and affects the performance of the captioning model. Another possible reason is that the training set in Oxford-102 contains 7,192 images and is relatively small, so the performance of our method degrades due to the overfitting problem. While in the dual learning framework proposed by “DL” and “MLADIC”, the image captioning model is trained together with an image synthesis model. The input of the captioning model includes both the images in the target domain and the synthesized images, which indicates that the captioning model learns from more training data and the overfitting problem is alleviated.
- On the target domain denoted as “Conceptual” that contains images and sentences from different sources and has a large domain gap from MSCOCO, our method outperforms the “Source Pre-trained” baseline. The results validate the effectiveness of our method when handling the challenging cross-domain image captioning scenario.

#### F. Ablation Studies

To evaluate the effect of each individual component, we evaluate several variants of our method on Flickr30k

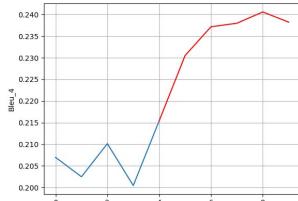
TABLE II  
QUANTITATIVE EVALUATION RESULTS OF CROSS-DOMAIN VIDEO CAPTIONING

Method	Target	Bleu_1	Bleu_2	Bleu_3	Bleu_4	METEOR	ROUGE_L	CIDEr
Source Pre-trained	MSVD	74.8	59.2	47.1	35.4	30.2	63.9	60.3
Ours	MSVD	<b>81.2</b>	<b>69.2</b>	<b>59.1</b>	<b>48.6</b>	<b>33.3</b>	<b>70.0</b>	<b>84.2</b>
Paired (upper bound)	MSVD	81.6	71.3	62.5	52.6	34.4	70.7	87.6
Source Pre-trained	Charades Captions	12.0	4.2	1.6	0.5	7.3	20.8	1.0
Ours	Charades Captions	<b>60.0</b>	<b>39.2</b>	<b>25.6</b>	<b>15.8</b>	<b>18.0</b>	<b>38.5</b>	<b>22.2</b>
Paired (upper bound)	Charades Captions	62.4	41.8	26.8	16.4	18.6	38.7	22.3

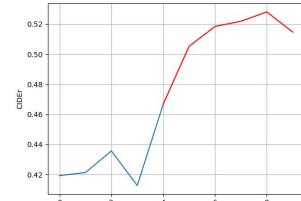
TABLE III

RESULTS OF ABLATION STUDIES ON THE FLICKR30K DATASET AND THE OXFORD-102 DATASET. THE COLUMNS 'ADAPTIVE' AND 'REFINE' DENOTES USING THE ADAPTIVE LSTM AND USING THE REFINED RETRIEVAL RESULTS, RESPECTIVELY

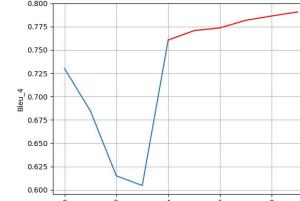
adaptive	refine	Target	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE_L	CIDEr
✓		Flickr30k	68.0	48.6	34.0	23.5	19.0	45.8	49.9
	✓	Flickr30k	68.5	49.1	34.2	23.6	19.3	46.1	50.9
	✓	Flickr30k	68.9	49.0	34.3	23.8	19.4	46.5	52.1
	✓	Flickr30k	69.0	49.3	34.7	24.1	19.5	46.5	52.8
✓		Oxford-102	93.7	89.6	84.4	80.3	43.0	83.7	82.3
	✓	Oxford-102	93.7	89.9	84.7	80.6	43.2	83.8	83.1
	✓	Oxford-102	96.9	91.7	85.8	80.0	43.1	77.7	86.8
	✓	Oxford-102	96.9	91.8	86.0	80.2	42.2	77.8	87.3



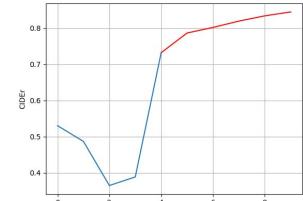
(a) Bleu-4 on Flickr30k



(b) CIDEr on Flickr30k



(c) Bleu-4 on Oxford-102



(d) CIDEr on Oxford-102

Fig. 3. Convergence analysis during model adaptation on Flickr30k and Oxford-102. The blue curve indicates training with cross-entropy and the red curve indicates training with reinforcement learning.

and Oxford-102. The details of these variants are as follows:

- *without adaptive LSTM*: The adaptive language LSTM is replaced by a normal LSTM model.
- *without refining retrieve results*: The pseudo image-sentence pairs in the initial iteration  $\hat{D}_t^0$  is used for fine-tuning the captioning model.
- *without adaptive LSTM or refined retrieve results*: The adaptive language LSTM is replaced by a normal LSTM model, and the pseudo image-sentence pairs in the initial iteration  $\hat{D}_t^0$  is used for fine-tuning the captioning model.

The results are shown in Table III. In general, both the adaptive LSTM and the pseudo image-sentence pairs contribute to the performance of the model.

1) *Parameter Analysis*: Besides, we also conduct additional experiments to show how the quantity of the pseudo image-sentence pairs affects the performance of cross-domain image captioning. In the parameter analysis experiments, the value of  $k$  defined in Section III-F.1 varies from 1 to 7 and the results are illustrated in Table IV. As the value of  $k$  increases, it is interesting to observe that most of the evaluation metrics show a trend of increasing, reaching the maximum value when  $k = 5$  and then decrease. The reason might be that when  $k$  is small, the retrieval results are accurate but the

TABLE IV  
RESULTS OF USING DIFFERENT NUMBERS OF RETRIEVED SAMPLES ON FLICKR30K AND OXFORD-102. THE COLUMNS B-N, M, R, C ARE ABBREVIATIONS FOR BLEU-N, METEOR, ROUGE\_L AND CIDEr, RESPECTIVELY

$k$	Target	B-1	B-2	B-3	B-4	M	R	C
1	Flickr30k	67.4	47.5	32.7	22.4	18.8	45.5	48.6
3	Flickr30k	68.0	48.6	34.0	23.5	19.0	45.8	49.9
5	Flickr30k	69.0	49.3	34.7	24.1	19.5	46.5	52.8
7	Flickr30k	68.5	48.8	34.1	23.7	19.5	46.2	51.8
1	Oxford-102	93.2	89.3	83.2	78.3	41.0	75.6	78.7
3	Oxford-102	94.7	90.8	84.7	80.0	41.7	75.9	83.1
5	Oxford-102	96.9	91.8	86.0	80.2	42.2	77.8	87.3
7	Oxford-102	94.3	90.5	84.3	79.3	41.2	77.0	82.8

quantity of pseudo image-sentence pairs is not sufficient to reflect the semantic correlations in the target domain data. When the value of  $k$  is too large, some inaccurate retrieval results are introduced, which reduces performance slightly. Thus, we fix the value of  $k$  to 5 in our experiments.

2) *Convergence Analysis*: To analyse the convergence of our method, we visualize the performance change in the Flickr30k dataset and the Oxford-102 dataset. We report the learning curves of Bleu-4 and CIDEr in Fig. 3, since Bleu-4 and CIDEr evaluates the quality of the sentence well and other evaluation metrics show a similar changing trend

TABLE V  
QUANTITATIVE RESULTS OF CROSS-MODAL RETRIEVAL ON THE TARGET DOMAIN DATASETS

method	target	Caption Retrieval (i2t)			Image Retrieval (t2i)		
		R@1	R@5	R@10	R@1	R@5	R@10
VSE++ [42]	Flickr30k	5.66	16.31	23.48	3.62	10.87	16.10
VSE++_refined	Flickr30k	7.13	18.32	25.74	4.71	12.76	18.38
VSE++ [42]	Oxford-102	0.10	0.46	0.93	0.09	0.40	0.80
VSE++_refined	Oxford-102	0.13	0.65	1.24	0.12	0.47	0.92

Image	Ground truth sentence	Generated sentence
	<ul style="list-style-type: none"> <li>A girl in a polka dotted blue jean dress walks barefoot on a balance beam</li> <li>A girl in a jean dress is walking along a raised balance beam.</li> <li>The little girl is walking a beam in gymnastics class.</li> <li>A little girl balances on a gymnastics beam.</li> <li>A little girl is walking on a balance beam.</li> </ul>	<b>Ours:</b> a little girl in a blue dress is jumping in a gym <b>Ours w/o refine:</b> a little girl in a blue shirt is playing <b>Ours w/o adapt:</b> a little girl in a blue dress is playing on a floor
	<ul style="list-style-type: none"> <li>A shirtless man is fishing, standing on large rocks that rest next to a boardwalk with a tall lamp post in the forefront.</li> <li>A shirtless man in shorts is fishing while standing on some rocks.</li> <li>A shirtless man in jeans shorts fishing off the rocks.</li> <li>A solitary fisherman stands along a riverbank.</li> <li>A man with no shirt fishing.</li> </ul>	<b>Ours:</b> a man in a blue shirt is fishing on a body of water <b>Ours w/o refine:</b> a man in a white shirt is standing on a water <b>Ours w/o adapt:</b> a man in a blue shirt is fishing on a lake
	<ul style="list-style-type: none"> <li>The boy is wearing a blue shirt, and the girl is wearing black clothes.</li> <li>A young boy wearing a blue shirt is playing at a park.</li> <li>The boy and girl are playing on the outdoor gym.</li> <li>The two kids are playing at the playground.</li> <li>Children playing on a playground.</li> </ul>	<b>Ours:</b> a young boy in a blue shirt is playing with a yellow toy on the street <b>Ours w/o refine:</b> a young boy in a blue shirt is playing on a playground <b>Ours w/o adapt:</b> a little girl in a blue dress is standing on the sidewalk
	<ul style="list-style-type: none"> <li>the flower shown has blue petals and several white filament</li> <li>this flower has blue purple petals with white stripes</li> <li>this flower is purple in color with petals that are wavy and ruffled</li> <li>the petals on this flower are purple with green stamen</li> <li>this flower has petals that are blue and folded together</li> </ul>	<b>Ours:</b> this flower is blue in color and has petals that are purple <b>Ours w/o refine:</b> this flower has petals that are purple and has yellow stamen <b>Ours w/o adapt:</b> this flower is blue in color and has petals that are purple
	<ul style="list-style-type: none"> <li>this bird is grey with white and has a long pointy beak</li> <li>this bird has wings that are brown and has a thick bill</li> <li>this bird has wings that are brown and has a silver bill</li> <li>a big bird with gray feathers and a big gray beak</li> <li>this bird has wings that are brown with a big beak</li> </ul>	<b>Ours:</b> this bird has wings that are brown and has a long bill <b>Ours w/o refine:</b> this bird has wings that are grey and has a long bill <b>Ours w/o adapt:</b> this bird has wings that are black and has a long pointy beak
	<ul style="list-style-type: none"> <li>this bird has a long tan neck a mottled black brown colored body and a red eyering</li> <li>this bird has a blood colored eye with a long neck and black bill</li> <li>a large brown bird with a long neck red eyes and black beak</li> <li>a bird with black and brown feathers and a red eye ring</li> <li>this bird has wings that are brown and has red eyes</li> </ul>	<b>Ours:</b> this bird has wings that are brown and has a long neck and red eyes <b>Ours w/o refine:</b> this bird has wings that are red and has a red belly <b>Ours w/o adapt:</b> this bird has wings that are brown and red eyes

Fig. 4. Sentences generated by our method and baseline methods conditioned on the images from the Flickr30k dataset (the first three rows), the Oxford-102 dataset (the fourth row) and the CUB-200 dataset (the last two rows). The sentences labeled with ‘Ours w/o refine’ are generated with the captioning model trained with the pseudo image-sentence pairs that are not refined, namely  $\hat{D}_t^0$ . The sentences marked with ‘Ours w/o adapt’ are generated with our full model that replaces the adaptive LSTM with normal LSTM model.

with these metrics during training. As is shown in the figure, after training with reinforcement learning, the performance largely increases and the model converges after about 8 epochs.

#### G. Extension to Cross-Domain Video Captioning

To evaluate the effectiveness and extensibility of our method, we further apply the proposed method to cross-domain video captioning. In these experiments, the MSR-VTT dataset [70] is used as the source domain, while the MSVD dataset [71] and the Charades Captions dataset [72] are used as the target domains. MSR-VTT contains 10,000 video clips

and 200,000 sentence annotations. The training, validation and testing splits of MSR-VTT contains 6,513, 2,990 and 497 video clips, respectively. The MSVD dataset contains 1,970 video clips from YouTube, where each clip is annotated with 10 sentences and the training, validation and testing splits include 1,200, 100 and 670 videos, respectively. There are 9,848 videos in the Charades Captions dataset, where 7,985 videos are for training and 1,863 videos are for testing. Compared to the MSVD dataset, the domain gap between the MSR-VTT dataset and the Charades Captions dataset is larger since the videos in Charades Captions are longer than the videos in MSR-VTT, and the descriptions in

Query	Retrieved samples	Retrieved samples (w/o refining)	Ground truth sample
A man and a little girl in a grassy area are planting a tree while a little boy off to the side is holding a hoe.			
	<ul style="list-style-type: none"> <li>· A group of people are climbing in cold weather.</li> <li>· Five people wearing winter jackets and helmets stand in the snow, with snowmobiles in the background.</li> <li>· A group of snowmobile riders gather in the snow.</li> <li>· Climbers with hiking boots and blue helmets ascend a snow covered mountain.</li> </ul>	<ul style="list-style-type: none"> <li>· Five people wearing winter jackets and helmets stand in the snow, with snowmobiles in the background.</li> <li>· A group of snowmobile riders gather in the snow.</li> <li>· Climbers with hiking boots and blue helmets ascend a snow covered mountain.</li> </ul>	<ul style="list-style-type: none"> <li>· The people are quietly listening while the story of the ice cabin was explained to them.</li> <li>· A group of people standing in front of an igloo.</li> </ul>

Fig. 5. Some examples retrieved by our cross-domain retrieval model on the Flickr30k dataset.

Charades Captions are more complex and includes multiple sentences.

For videos in all the video captioning datasets, we sample frames at 2 fps and extract the features of the frames using the pre-trained Inception-Resnet model [73]. The feature vector before the last fully-connected layer is extracted for each frame.

As is shown in Table II, we observe that our method outperforms the baseline by a large margin. Compared to the model pre-trained on the source domain, our method improves the performance significantly on Charades Captions dataset, which has a large domain gap from the source domain. The results indicate that our method generalizes well to the cross-domain video captioning task, validating the effectiveness of our method.

#### H. Results of Cross-Modal Retrieval

We also quantitatively evaluate the cross-modal retrieval performance of our method on the target domain datasets. We compare the cross-modal retrieval method proposed in [42], denoted as VSE++, with our method. The VSE++ model is pre-trained on MSCOCO and is evaluated on the test sets of Flickr30k and Oxford-102. The performance is measured using recall at N, namely the portion of queries whose top N nearest samples in another modality contain the correct sample, denoted as R@N. The results are shown in Table V. Compared with the baseline method, our method achieves higher recall on both datasets, indicating that our method is able to transfer to the target domains with a moderate domain gap as well as a large domain gap.

#### I. Qualitative Results

In this section, we show some sentences generated by our method and the baseline methods in Fig. 4. As shown in the figure, our captioning model generates captions that describe the content of the image in detail. We can also observe that the quality of the sentences generated by the full model is superior to that of the sentences generated by the captioning model trained with pseudo image-sentence pairs that are not refined.

To obtain an intuitive understanding of our cross-modal retrieval model, we also show some examples of retrieved image-sentence pairs in Fig. 5. By comparing the retrieved samples and the query, we observe that the retrieved samples are semantically relevant to the query, indicating that our retrieval model is able to capture the correlation between the images and sentences in the target domains. Furthermore, the retrieved samples closest to the query share similar content. For instance, all the retrieved images in the first row involve multiple persons and have trees and grassy field in the background, which illustrates that the retrieved results are reasonable.

#### V. CONCLUSION

In this paper, we have presented a novel cross-domain image captioning approach that couples a cross-modal retrieval model and an adaptive image captioning model. The retrieval model is designed to generate pseudo image-sentence pairs in the target domain which can enhance the performance of captioning. The adaptive image captioning model is responsible for adapting the source domain knowledge to the target domain which can bridge the gap between different domains. Extensive experiments on four public datasets can validate the effectiveness of our method. In our future work, we are going to reduce the domain shift in image spaces to improve the captioning performance.

#### REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
- [2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.
- [3] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.
- [4] K. Xu *et al.*, “Show, attend and tell: Neural image caption generation with visual attention,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
- [5] T.-Y. Lin *et al.*, “Microsoft COCO: Common objects in context,” in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
- [6] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” *Trans. Assoc. for Comput. Linguistics*, vol. 2, pp. 67–78, Dec. 2014.

- [7] T.-H. Chen, Y.-H. Liao, C.-Y. Chuang, W.-T. Hsu, J. Fu, and M. Sun, "Show, adapt and tell: Adversarial training of cross-domain image captioner," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 521–530.
- [8] A. Farhadi *et al.*, "Every picture tells a story: Generating sentences from images," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 15–29.
- [9] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proc. 15th Conf. Comput. Natural Lang. Learn.*, 2011, pp. 220–228.
- [10] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artif. Intell. Res.*, vol. 47, pp. 853–899, Aug. 2013.
- [11] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain images with multimodal recurrent neural networks," 2014, *arXiv:1410.1090*. [Online]. Available: <http://arxiv.org/abs/1410.1090>
- [12] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4651–4659.
- [13] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.
- [14] B. Zhao, X. Li, and X. Lu, "CAM-RNN: co-attention model based RNN for video captioning," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5552–5565, Nov. 2019.
- [15] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10685–10694.
- [16] J. Gu, H. Zhao, Z. Lin, S. Li, J. Cai, and M. Ling, "Scene graph generation with external knowledge and image reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1969–1978.
- [17] X. Li and S. Jiang, "Know more say less: Image captioning based on scene graphs," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 2117–2130, Aug. 2019.
- [18] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [19] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8927–8936.
- [20] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10575–10584.
- [21] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4125–4134.
- [22] J. Gu, S. Joty, J. Cai, H. Zhao, X. Yang, and G. Wang, "Unpaired image captioning via scene graph alignments," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10322–10331.
- [23] I. Laina, C. Rupprecht, and N. Navab, "Towards unsupervised image captioning with shared multimodal embeddings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7413–7423.
- [24] D. Guo, Y. Wang, P. Song, and M. Wang, "Recurrent relational memory network for unsupervised image captioning," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 920–926.
- [25] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
- [26] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. summarization*, 2005, pp. 65–72.
- [27] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.
- [28] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.
- [29] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," 2015, *arXiv:1511.06732*. [Online]. Available: <http://arxiv.org/abs/1511.06732>
- [30] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1179–1195.
- [31] J. Gao, S. Wang, S. Wang, S. Ma, and W. Gao, "Self-critical N-Step training for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6300–6308.
- [32] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, "Deep compositional captioning: Describing novel object categories without paired training data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1–10.
- [33] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko, "Captioning images with diverse objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1170–1178.
- [34] W. Zhao *et al.*, "Dual learning for cross-domain image captioning," in *Proc. ACM Conf. Inf. Knowl. Manage.*, Nov. 2017, pp. 29–38.
- [35] M. Yang *et al.*, "Multitask learning for cross-domain image captioning," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1047–1061, Apr. 2019.
- [36] J. Tang, K. Wang, and L. Shao, "Supervised matrix factorization hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3157–3166, Jul. 2016.
- [37] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3893–3903, Aug. 2018.
- [38] J. Chen, W. K. Cheung, and A. Wang, "Learning deep unsupervised binary codes for image retrieval," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 613–619.
- [39] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4242–4251.
- [40] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.
- [41] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," 2014, *arXiv:1411.2539*. [Online]. Available: <http://arxiv.org/abs/1411.2539>
- [42] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," in *Proc. Brit. Mach. Vis. Conf.*, 2018, p. 12.
- [43] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning coupled feature spaces for cross-modal matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2088–2095.
- [44] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3441–3450.
- [45] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with natural language," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5804–5813.
- [46] D. Zhang, X. Dai, X. Wang, Y.-F. Wang, and L. S. Davis, "MAN: Moment alignment network for natural language moment retrieval via iterative graph adjustment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1247–1257.
- [47] D. Cao, Z. Yu, H. Zhang, J. Fang, L. Nie, and Q. Tian, "Video-based cross-modal recipe retrieval," in *Proc. 27th ACM Int. Conf. Multimedia*, Nice, France, Oct. 2019, pp. 1685–1693.
- [48] Z. Lin, Z. Zhao, Z. Zhang, Z. Zhang, and D. Cai, "Moment retrieval via cross-modal interaction networks with query reconstruction," *IEEE Trans. Image Process.*, vol. 29, pp. 3750–3762, 2020.
- [49] Z. Zhang, Z. Lin, Z. Zhao, and Z. Xiao, "Cross-modal interaction networks for query-based moment retrieval in videos," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 655–664.
- [50] S. Chen, Y. Zhao, Q. Jin, and Q. Wu, "Fine-grained video-text retrieval with hierarchical graph reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10635–10644.
- [51] N. C. Mithun, S. Paul, and A. K. Roy-Chowdhury, "Weakly supervised video moment retrieval from text queries," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11592–11601.
- [52] Z. Lin, Z. Zhao, Z. Zhang, Q. Wang, and H. Liu, "Weakly-supervised video moment retrieval via semantic completion network," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11539–11546.
- [53] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1367–1381, Jun. 2018.
- [54] H. Xue, W. Chu, Z. Zhao, and D. Cai, "A better way to attend: Attention with trees for video question answering," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5563–5574, Nov. 2018.
- [55] Z. Zhao, Z. Zhang, X. Jiang, and D. Cai, "Multi-turn video question answering via hierarchical attention context reinforced networks," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3860–3872, Aug. 2019.

- [56] T. Yu, J. Yu, Z. Yu, and D. Tao, "Compositional attention networks with two-stream fusion for video question answering," *IEEE Trans. Image Process.*, vol. 29, pp. 1204–1218, 2020.
- [57] Z. Zhao, S. Xiao, Z. Song, C. Lu, J. Xiao, and Y. Zhuang, "Open-ended video question answering via multi-modal conditional adversarial networks," *IEEE Trans. Image Process.*, vol. 29, pp. 3859–3870, 2020.
- [58] J. Liang, L. Jiang, L. Cao, Y. Kalantidis, L.-J. Li, and A. G. Hauptmann, "Focal visual-text attention for memex question answering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1893–1908, Aug. 2019.
- [59] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel, "FVQA: fact-based visual question answering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2413–2427, Oct. 2018.
- [60] F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun, "Inverse visual question answering: A new benchmark and VQA diagnosis tool," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 460–474, Feb. 2020.
- [61] H. Li, P. Wang, C. Shen, and A. V. D. Hengel, "Visual question answering as reading comprehension," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6319–6328.
- [62] H. Kim, Z. Tang, and M. Bansal, "Dense-caption matching and frame-selection gating for temporal localization in VideoQA," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4812–4822.
- [63] Y. Li *et al.*, "TGIF: A new dataset and benchmark on animated GIF description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4641–4650.
- [64] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.
- [65] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2008, pp. 722–729.
- [66] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 49–58.
- [67] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 2556–2565.
- [68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [69] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
- [70] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5288–5296.
- [71] D. Chen and W. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proc. Annu. Meet. Assoc. Comput. Linguistics*, 2011, pp. 190–200.
- [72] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 510–526.
- [73] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 2017, pp. 4278–4284.

**Wentian Zhao** received the B.S. degree in computer science from the Beijing Institute of Technology, Beijing, in 2017, where he is currently pursuing the Ph.D. degree with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science and Technology. His research interests include computer vision and natural language processing.



**Xinxiao Wu** (Member, IEEE) received the B.S. degree from the Nanjing University of Information Science and Technology, Nanjing, in 2005, and the Ph.D. degree from the Beijing Institute of Technology, China, in 2010. She is currently an Associate Professor with the Beijing Institute of Technology. Her research interests include computer vision, machine learning, and video content analysis.



**Jiebo Luo** joined the Department of Computer Science, University of Rochester, in 2011, after a prolific career of more than 15 years with Kodak Research. He has authored more than 400 technical articles and holds more than 90 U.S. patents. His research interests include computer vision, machine learning, data mining, social media, and biomedical informatics. He has served as the Program Chair for the ACM Multimedia 2010, the IEEE CVPR 2012, ACM ICML 2016, and IEEE ICIP 2017, and on the Editorial Boards for the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON BIG DATA, *Pattern Recognition, Machine Vision and Applications*, and *ACM Transactions on Intelligent Systems and Technology*. He is a fellow of ACM, AAAI, SPIE, and IAPR.

