Abstract—The goal of unpaired image captioning (UIC) is to describe images without using image-caption pairs in the training phase. Although challenging, we expect the task can be accomplished by leveraging images aligned with visual concepts. Most existing studies use off-the-shelf algorithms to obtain the visual concepts because the Bounding Box (BBox) labels or relationship-triplet labels used for training are expensive to acquire. To avoid exhaustive annotations, we propose a novel approach to achieve cost-effective UIC. Specifically, we adopt image-level labels to optimize the UIC model in a weakly-supervised manner. For each image, we assume that only the image-level labels are available without specific locations and numbers. The image-level labels are utilized to train a weakly-supervised object recognition model to extract object information (e.g., instance), and the extracted instances are adopted to infer the relationships among different objects using an enhanced graph neural network (GNN). The proposed approach achieves comparable or even better performance compared with previous methods without expensive annotations. Furthermore, we design an unrecognized object (UnO) loss to improve the alignment of the inferred object and relationship information with the images. It can effectively alleviate the issue encountered by existing UIC models when generating sentences with nonexistent objects. To the best of our knowledge, this is the first attempt to address the problem of Weakly-Supervised visual concept recognition for UIC (WS-UIC) based only on image-level labels. Extensive experiments demonstrate that the proposed method achieves inspiring results on the COCO dataset while significantly reducing the labeling cost.

Index Terms—Graph neural network, unpaired image captioning, weakly-supervised instance segmentation.

I. INTRODUCTION

IMAGE captioning [1], [2], [3] aims at describing the content and event of an image using a couple of words. We can also treat it as a mapping from image to corresponding natural language descriptions [4]. This task has been greatly promoted by the deep learning algorithms [1], which can learn from large-scale annotated image-sentence pairs [5]. Image captioning has been widely used in many practical applications [6], including human-robot interaction [7], [8], automatic driving [9], [10], visual assistance for impaired people [11], [12], etc. The essential practice of state-of-the-art image captioning approaches follows the encoder-decoder paradigm [13], [14]. Generally, an input image is first encoded into feature representation using Convolutional Neural Network (CNN). Then, the Recurrent Neural Network (RNN) is adopted to decode the features into multiple words one by one. This way, the natural description for the input image can be obtained.

In spite of the promising applications and various mature models developed, the standard image captioning models are mostly trained in a fully-supervised manner, which may require a tremendous number of manually annotated image-caption pairs [21], [22]. It is very difficult to obtain such image-caption pairs since the manual annotation is very costly [23], [24], [25]. In addition, the generalization ability of such models maybe limited, as the collected images and annotated captions are often biased and incomplete [26]. The existing image captioning datasets, such as Microsoft COCO [27], are relatively small in scales comparing with the most popular image classification datasets, such as OpenImages [28] and ImageNet [29]. The varieties of images and captions within these datasets are also limited in the order of 100 object categories [2]. As a result, it is difficult for the captioning models trained on such paired image-caption data to generalize them to images in the wild. Therefore, it is desirable to develop Unpaired Image Captioning (UIC) approaches that do not require image-caption training pairs [18], [30], [31].

In the absence of image-caption pairs, the learning of UIC models needs enormously additive labels. For example, the UIC model often needs to recognize the category and attributes of objects, and sometimes the relationships between different objects in the image. The clues obtained can be utilized to build the connections between the visual concepts and the

Manuscript received 4 March 2022; revised 19 July 2022 and 21 September 2022; accepted 26 September 2022. Date of publication 12 October 2022; date of current version 1 November 2023. This work was supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2019B011015502, in part by Peng Cheng Laboratory Research under Grant PCL2021A07, and in part by the National Natural Science Foundation of China under Grants 62102205 and 62276195. The Associate Editor coordinating the review of this manuscript and approving it for publication was Mrs. Si Liu. (Corresponding authors: Yaowei Wang; Changwen Chen.)
Fig. 1. Comparisons between the unpaired image captioning (UIC) approaches based on (a) object detection, (b) relationship-triplet recognition, and (c) the proposed weakly-supervised (WS) visual concept recognition that requires only image-level labels. Specifically, the object detection scheme utilizes BBox labels to recognize object concepts [2], [15], [16]; the approach termed relationship-triplet recognition combines object detection and relationship-triplet recognition [17], [18], [19] to differentiate the (subject, relationship, object) information, which relies on complicated relationship-triplet labels; and the proposed WS-UIC method, which is based on only image-level labels to train WS object recognition (WS-OR) and WS relationship recognition (WS-RR), can achieve better performance than most of the (a) and (b) approaches. According to the Amazon Mechanical Turk, the unit price of labeling BBox labels and relationship labels are three times and nine times that of annotating image-level labels, respectively [20].

![Image of Unpaired Image Captioning by Image-Level Weakly-Supervised Visual Concept Recognition]

Unit labeling price: 
- $0.036 for BBox labels; $0.108 for relationship-triplet labels; and $0.012 for image-level labels.

In this section, we first give a review of the image captioning and unpaired image captioning (UIC), which learns image captioner with paired image-sentence data or unpaired image/sentence data. Then, we discuss the core techniques for UIC, i.e., the object concept recognition and the relationship exploration.

A. Image Captioning

In the past few years, fully-supervised image captioning has been studied extensively [6]. The majority of the proposed models adopt the encoder-decoder paradigm where one Convolutional Neural Network (CNN) is leveraged to encode an input
image firstly and one Recurrent Neural Network (RNN) is utilized to output a description for the image subsequently [1], [33], [35]. These models are trained to maximize the probability of generating the ground-truth captions, depending on enormous image-caption pairs [26], [36]. As paired image-caption data is hard to collect, some researchers attempted to decouple the dependency on the paired annotations through other available datasets [13]. Hendricks et al. [25] trained a caption model of describing novel objects without relying on image-caption data containing the novel object concepts, which leverages large object recognition datasets and external text corpora by transferring knowledge between semantically similar concepts. Yao et al. [24] presented a Long Short-Term Memory with Copying Mechanism (LSTM-C) to describe novel objects in captions, incorporating copying mechanism into the CNN plus RNN image captioning framework. Chen et al. [37] proposed a semi-supervised image captioning model by artificially generating missing visual information conditioned on the textual data. Kim et al. [38] also proposed a semi-supervised learning method to assign pseudo-labels to unlabeled images via Generative Adversarial Networks, which in turn are utilized to train a fully-supervised captioner. Although promising captioning results have been achieved, the novel object captioning or semi-supervised image captioning methods still require expensive paired image-caption data for training. Different from these works, we aim to tackle unpaired image captioning without relying on any image-caption pair.

B. Unpaired Image Captioning

UIC is capable of producing captions for input images without adopting any paired image-sentence data and has attracted significant attention from researchers. Gu et al. [39] implemented language pivoting to achieve unpaired image captioning. However, the scheme requires the ground-truth pivot-image pairs and paired pivot-target language translation datasets. Feng et al. [2] proposed the first work that tackles the captioning task through training with totally unpaired image-sentence datasets. Recently, Laina et al. [15] employed a shared multi-modal embedding, structured by visual concepts, to bridge the gap between the image and sentence domains. Guo et al. [16] also proposed a novel Recurrent Relational Memory (R²M) Network which can be implemented to get rid of the complicated and sensitive adversarial learning. SCS [13] has proposed a semantic-constrained self-learning strategy that iteratively generates “pseudo” sentences and re-trains the captioner for UIC. In recent years, other researchers have adopted scene graph modeling in the model to utilize more semantic information, including the relationships between objects and the attributes of objects [17], [18], [19]. All of these approaches achieved better performance than the schemes proposed by Feng et al. and Laina et al. but required much more ground-truth information or more expensive annotations.

Although an unpaired image-sentence dataset is used, these approaches still depend on enormously expensive labels at the visual concept recognition stage, including the BBox labels or the relationship-triplet labels. As far as we know, the proposed approach in this research is the first work to tackle visual concept recognition for UIC aided by only image-level class labels. Besides, an unrecognized object loss is designed to guide the UIC model to exclude the unrecognized objects.

C. Captioning Based on Object Concept Recognition

The schemes of object concept recognition have been developed to recognize the objects contained in an image. The recognized object concepts ensure that the generated captions can
represent the main topic of an image by aligning these object words with images. Various schemes have been developed, including image classification, object detection \cite{40, 41, 42}, Multiple Instance Learning (MIL), and so on. In particular, Hendricks et al. \cite{25} and Venugopalan et al. \cite{43} utilized an image classifier to recognize various object categories. However, this approach can only be implemented to differentiate the object categories instead of the instances. Other related works include the use of object detection model \cite{40} to differentiate the instances of objects \cite{2, 15, 24, 44}. Specifically, Liu et al. \cite{18} have successfully employed the weakly-supervised MIL to build the semantic concept extractor. However, these object detection schemes and MIL approaches rely heavily on expensive BBox labels.

In this research, in order to obtain rich information of objects but free the model from the bondage of expensive labels, we adopt weakly-supervised instance segmentation to recognize the object concepts, which is a primary difference from existing UIC works.

D. Captioning Based on Relationship Exploration

For image captioning, many studies explore relationships between different objects to improve the performance of captioning. It is because these relationships provide rich semantic information of the input images. Among the existing works, the most popular model is graph representation which can represent the complex structural layout of both images and sentences. In supervised image captioning, Yang et al. \cite{45} adopted scene graph representation in the auto-encoder to obtain more human-like captions. For UIC, Gu et al. \cite{17} also applied scene graph to represent the object attributes and object relationships. Besides the scene graph representation, Cao et al. \cite{19} designed a mutual attention network to reason the object-object interactions in UIC. Although various works of vision-language tasks exhibit the value of relationships between objects, it is absolutely necessary to use costly triplet annotations (subject, relationship, object) with locations, thereby limiting the adoption of these schemes to broader applications.

Distinct from these previous works, we design a novel scheme using only image-level labels to differentiate the relationships between objects, properly aided by the results of weakly-supervised instance segmentation and image classification. More importantly, a batch normalization (BN) scheme and residual block are adopted to enhance the stability of the basic relationship recognition model.

III. OUR PROPOSED APPROACH

In this section, we will first give an overview of our newly proposed weakly-supervised (WS) visual concept recognition for unpaired image captioning (WS-UIC). Then, we discuss the problem formulation of WS-UIC. After that, we will describe the WS Object Recognition (WS-OR) and WS Relationship Recognition (WS-RR) module in our framework. Finally, we present the learning of UIC model under the guidance of the aforementioned modules.

A. Overview

The goal of WS-UIC is to learn an image captioning model using unpaired image-text samples in a weakly-supervised manner. To achieve the goal, our newly proposed WS-UIC framework contains three main modules, i.e., the WS-OR, WS-RR, and UIC model, as shown in Fig. 2. In particular, we have a WS-OR module, consisting of an image classifier and a WS instance segmentation head, to generate the object information supervised by the image-level object labels from an image. The image classifier is trained to generate the object category information. The WS instance segmentation head, conditioned on the object category information, is utilized to obtain information of each object instance. A WS-RR module is trained by the image-level relationship labels to output the relationship concepts of an input image, where we adopt the multi-scale feature maps and spatial relations of multiple-instance to construct a graph neural network (GNN) enhanced by batch normalization and residual block. In addition, we have a UIC model to generate the image descriptions supervised by these object information and relationship concepts via a designed unrecognized object (UnO) loss integrated with concept rewards.

B. Problem Formulation

To clearly and formally illustrate the proposed WS-UIC, we define some notations and formulate the problem of WS-UIC as a multi-stage process. Let \(D_i = \{i^n\}_{n=1}^{N_i-1} \) and \(D_s = \{s^n\}_{n=1}^{N_s-1} \) denote the image dataset with \(N_i \) images and the sentence dataset with \(N_s \) sentences, respectively. Let \(D_{i,o,r} = \{(i, o, r)\}_{i,o,r=1}^{N_{i,o,r}} \) denote the dataset with \(N_{i,o,r} \) tuples of images, image-level object labels, and image-level label relations, indexed by \(i, o, \) and \(r \), respectively. The dataset \(D_i \) and \(D_s \) are used for UIC, and \(D_{i,o,r} \) is used for WS-OR and WS-RR, simultaneously. Formally, the goal of the WS-UIC can be written as

\[
s \sim \arg \max_\theta \{P(s|\theta_{i\rightarrow s})\}, \tag{1}
\]

where \(\theta_{i\rightarrow s} \) are the model parameters to be learned in the absence of any paired \(i \) and \(s \), which are from independent datasets \(D_i \) and \(D_s \). We use the visual concepts to learn the mapping:

\[
i \xrightarrow{\theta_{i\rightarrow o}} o \xrightarrow{\theta_{i,o\rightarrow r}} s,
\]

where \(o \) denotes the object instance.

According to the defined notations, the multi-stage process of the WS-UIC, i.e., WS-OR, WS-RR, and UIC, can be formulated as:

\[
P(s, o, r|i) = \]

\[
P(s|i, o, r; \theta_{i,o\rightarrow r,s}) \quad \text{UIC} \tag{2}
\]

\[
\times P(r|i, o; \theta_{i,o\rightarrow r}) \quad \text{WS - RR} \tag{3}
\]

\[
\times P(o|i; \theta_{i\rightarrow o}). \quad \text{WS - OR} \tag{4}
\]

where \(P(o|i; \theta_{i\rightarrow o}), P(r|i, o; \theta_{i,o\rightarrow r}) \), and \(P(s|i, o, r; \theta_{i,o,r\rightarrow s}) \) represents the WS-OR module, WS-RR module, and UIC model, respectively. In the inference phase, only the UIC model
is needed to describe an unseen image:

\[
s \sim \arg \max_s \{ P(s|i; \theta_{\hat{o}, r \rightarrow s}) \}.
\]  

(5)

In the following subsections, we will depict the WS-OR, WS-RR, and UIC modules in a more detailed way. It is worth noting that only image-level labels are adopted in WS-UIC to train the WS-OR and WS-RR module in a weakly-supervised manner, but other works depend on enormous expensive labels, such as relationship-triplet labels, to obtain the object or/and relationship information in an image.

C. Exploring Weakly-Supervised Object Recognition

For UIC, image-text pairs are not available, therefore, the object concepts of images are taken as the crucial clues for generating accurate captions. Previous works utilize object detection models to identify these object concepts but rely on enormous Bounding Box (BBox) labels (as shown in Fig. 4(b)) to train the model. To overcome this drawback, a WS-OR model \( P(\hat{o}|i; \theta_{\hat{o}, r \rightarrow i}) \) is proposed to recognize the object concepts given only image-level labels, as shown in Fig. 4(d). The main structure of WS-OR includes an image classifier and an image-level instance segmentation module [32]. Therefore, utilizing only image-level labels is capable of differentiating individual objects.

1) Image Classification: plays essential roles in our proposed approach: First, it can enlarge the number of object categories of the captioning dataset to increase the vocabulary of the UIC model. Second, it can be implemented to generate the feature maps \( F \) and Class Attention Maps (CAMs) [46] which define the distinct areas of object instances and thus can be adopted to train the image-level instance segmentation model. Without loss of generality, in our experiments, we select the ResNet50 [47] as the backbone of the classification network. And we adopt the popular used multi-label soft margin loss [48].

To get the CAMs [46], we follow [49] by using the Global Average Pooling (GAP) in Convolutional Neural Networks (CNNs). Formally, the CAM of an object class \( \hat{o}_i \) is denoted by \( M_{\hat{o}_i} \),

\[
M_{\hat{o}_i}(x, y) = \frac{\phi_{\hat{o}_i}^T f(x, y)}{\max_{x, y} \phi_{\hat{o}_i}^T f(x, y)},
\]  

(6)

where \( f(x, y) \) represents the feature map with coordinate \( (x, y) \), and \( \phi_{\hat{o}_i} \) is the classification weights of object class \( \hat{o}_i \).

2) Image-level Instance Segmentation: is another critical component of WS-OR. It is utilized to recognize instance information for individual objects in an image, including instance category \( \hat{o}_i \), instance mask \( m_i \), and the corresponding confident score \( z_i \). We follow IRNet [32] to address the instance segmentation issue by adopting the CAMs as the source supervision. Fig. 3 illustrates some results of the WS-OR module.

The obtained object information can be formulated as below:

\[
\mathcal{O} = [(\hat{o}_1, m_1, z_1), \ldots, (\hat{o}_i, m_i, z_i), \ldots, (\hat{o}_{N_O}, m_{N_O}, z_{N_O})],
\]  

(7)

where \( N_O \) is the total number of object concepts in an image. The obtained object information is utilized to predicate the relationships between objects in Section III-D and train the UIC model to generate descriptions in Section III-E. It is worth to mention that we can adopt any image-level instance segmentation approach or image-level object detection scheme to differentiate objects.
of the graph are provide image-words pairs for as input and \( \times \) is the indicator \( V = (r, h, \alpha) \), \( \times \) as follows:

\[
\begin{align*}
\mathcal{F} &= (r, z_1, \ldots, r, z_k, \ldots, r, z_{N_R}), \quad \mathcal{O} = \{O_i\},
\end{align*}
\]

where \( h \) indicates the output of the former layer, \( k \) is the dimensional number of \( h \), \( E \) and \( Var \) mean the expectation and variance, respectively. The residual block is designed to solve the gradient degradation problem via adding the output of former layers with current layers [47]. Formally,

\[
\hat{F}(h) = F(h) + h,
\]

where \( F(h) \) is the output of the current layer. We adopt the residual scheme in each MLP block of the WS-RR network.

Through this section, we can obtain the relationship concepts of images:

\[
\mathcal{R} = \{r_1, z_1, \ldots, r_k, z_k, \ldots, r_{N_R}, z_{N_R}\},
\]

where \( r_k \) is the \( k \)-th inferred relationship class, \( z_k \) represents the corresponding confidence score, and \( N_R \) defines the total number of the recognized relationship concepts. The recognized relationship information can be utilized to guide UIC on generating relevant image captions in Section III-E.

D. Exploring Weakly-Supervised Relationship Recognition

In addition to the object category information, the relationships between these objects are another crucial part of the image captions. Most existing schemes [17], [30] explore these relationship concepts by using enormous relationship-triplet labels, which are expensive to be annotated and hard to collect in practice. To address this issue, we propose utilizing image-level relationship labels to train the model to recognize the relationship concepts. Fig. 4(c) and (d) show an example of relationship-triplet labels and image-level relationship labels respectively.

In our framework, the GNN is adopted to model the relations between multiple instances in an image. The GNN takes the graph representation of an image \( G = (V, E) \) as input and predict the relationships among objects. As the image-level relationship labels contain no object information, therefore, we need to use the object information \( \mathcal{O} \) obtained from Section III-C to construct the graph representation. Specifically, the nodes \( v_i = (v^o_i, v^e_i) \), indicating an object with its appearance features and spatial features, and edges \( e_{ij} \) of the graph are transformed through small networks separately. Then pairs of nodes and the edge connecting them are input into a several Multi-Layer Perceptron (MLP) network [30]. The graph is fully-connected since the relationships between all pairs of objects are considered. Later, a classifier is applied to each of the edges, and the softmax outputs are supervised by the image-level relationship labels via a cross-entropy loss. The whole architecture is illustrated in Fig. 5(a).

To make the training of our model more stable and easier to converge, the BN and residual schemes are introduced in GNN, as shown in Fig. 5(b) and (c) respectively. The BN with fixed means and expectations is adopted in the convolutional layer, which keeps the features from different layers following the same suitable distribution. It makes the network less sensitive to the initialization and can accelerate the convergence speed [50]. Formally,

\[
\hat{h}^k = \frac{h^k - E[h^k]}{\sqrt{Var[h^k]}},
\]

E. Unpaired Image Captioning

Similar to the standard image captioning models, the UIC model adopts the encoder-decoder framework [33] to encode an image into features and decode these features into captions. Any CNN backbone networks can be used for the feature extraction. In our experiments, we select the inception-V4 [51] as the image encoder, and obtain the feature representation \( f_m \). For the decoder network, we adopt the popular Long-short Term Memory (LSTM) network [52] which takes the encoded feature maps as input at the first time step. Then, it recurrently takes its hidden state and the word embedding of the previous time step as the input and output one word and the hidden state for subsequent time steps [35].

Unlike traditional captioning models, the ground-truth captions are not available for UIC, therefore, we have to find other useful information for the optimization of our network. Firstly, a discriminator with the LSTM architecture is designed to differentiate the true sentences and generated sentences via an adversarial loss [2]. Secondly, since the objects and the relationships of objects are the key components of a caption, they can be served as the supervision for UIC. In this work, we leverage both the concept reward loss and unrecognized object loss to train the model. More details of the two losses are given below.

1) Concept Reward: The recognized object concepts and relationship concepts \( \mathcal{V} \in (\mathcal{O}, \mathcal{R}) \) provide image-words pairs for UIC. If the \( t \)-th generated word \( s_t \) is in the recognized visual concepts, a concept reward \( R_t \) will be assigned to the \( t \)-th generated word \( s_t \) as follows:

\[
R_t = \alpha \times \sum_{i=1}^{N} I(s_t = o_i) \times z_i + \beta \sum_{k=1}^{K} I(s_t = r_k) \times z_k,
\]

where the \( \alpha \) and \( \beta \) are the weights of object concept rewards and relationship concept rewards, respectively. \( I(\cdot) \) is the indicator function [53].

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
For WS visual concept recognition, we use the object concepts and relationship concepts. Given a testing image, we can directly generate its captions using the image classification model. Formally,

\[ L_t^u = \lambda \times \sum_{i=1}^{N_q} \mathbb{I}(s_i = q_i) \times z_i, \]

where \( \lambda \) is the weight of the loss and \( N_q \) is the total number of unrecognized objects. The loss for \( s_i \) is a combination of the visual concept reward and unrecognized object loss:

\[ L_t = -R_t + L_t^u. \]

### IV. Experiments

In this section, we evaluate the effectiveness of the proposed Weakly-Supervised (WS) visual concept recognition for unpaired image captioning (WS-UIC).

#### A. Dataset and Evaluation Metric

1) **Dataset:** For WS visual concept recognition, we use the Visual Genome (VG) dataset [54] to train the WS object recognition (WS-OR) and WS relationship recognition (WS-RR) models, where 305 object categories and 64 relationship categories are utilized. Then the COCO images are adopted to derive the object and relationship concepts. Only the image-level class labels are utilized in these procedures. For UIC, we utilize COCO images paired with COCO captions to train the model in an unpaired way. We adopt the common splits of COCO images [55]: 113,287 training images, 5000 validation images, and 5000 test images. The COCO captions are also utilized to construct the vocabulary corpus with words presenting no less than 4 times.

To illustrate the robustness of visual concept recognition of WS-UIC, we use another dataset to train the WS-OR model and the WS-RR model, i.e., OpenImage V2 [28] with 545 object categories. Shutterstock sentence corpus [2] is implemented as another sentence corpus, and COCO images in the training set are taken as the image corpus in the experiments. The sentence corpus contains 2,282,444 distinct image descriptions. Flickr30k [56] is adopted as another image dataset for captioning, which contains 29,783 training images, 1,000 validation images, and 1,000 testing images.

2) **Evaluation Metric:** We report the BLEU-1 (B1), BLEU-2 (B2), BLEU-3 (B3), BLEU-4 (B4) [57], Meteor (M) [58], ROUGE (R) [59], CIDEr (C) [60] and SPICE (S) [61] scores which are calculated via the ground-truth captions of the test images. Among these metrics, BLEU and Meteor (M) were produced for machine translation in 2002 and 2014, ROUGE (R) was mainly proposed for automatic abstracting in 2004, CIDEr
(C) and SPICE (S) were mainly for image captioning proposed in 2015 and 2016, respectively.

B. Implementation Details

For WS-OR, we select the object categories for an image with logits larger than 2 in the image classification stage. The ResNet50 [47] is taken as the backbone. Then, we follow IR-Net [32] to tackle image-level instance segmentation.

In the WS-RR stage, the batch size is set as 256, the relationships with confident scores larger than 0.7 will be filtered out to train the UIC model. We choose GeLU [62] as the activation function. Four scales of image-level feature maps are used in our experiments.

For the UIC model, we set both the LSTM hidden dimension and the shared latent space dimension as 512. The learning rates are set to be 0.00001. We also adopt the initialization pipeline following [2]. The weighting hyper-parameters of the concept reward and unrecognized object loss are chosen roughly at the same scales. Specially, \(\alpha\), \(\beta\) and \(\lambda\) are set as 1, 0.5, 1, separately. The code of this paper is developed based on TensorFlow framework, and all the experiments are conducted on a server with 4 V100 GPUs.

C. Comparison on Benchmarks

Four types of experimental comparisons are conducted in the following parts. The first one is to compare the WS-UIC with Pivoting [39], Song et al. [63], Guo et al. [64], Feng et al. [2], Laina et al. [15], Cao et al. [19], Gu et al. [17], SCS [13], illustrated in Table I. These approaches implement the same dataset at the UIC stage. The second one is to compare the WS-UIC with Feng et al. [2] with the same dataset settings at both the visual concept recognition stage and the UIC stage, shown in Table II. The third one is to compare the WS-UIC with Feng et al. [2] via implementing the independent datasets, i.e., COCO images with Shutterstock sentences. The last one is to implement the Flickr30k test images to do experimental comparisons.

As shown in Table I, we report our results and compare with other state-of-the-art methods, including [2], [13], [15], [17], [19], [39], [63], [64]. Note that all these methods use the same dataset, i.e., the COCO dataset, for the training of the captioning model. From the Table, we can find that the performance of Cao et al. [19], Gu et al. [17], and SCS [13] are slightly better than ours. The reason is that these methods adopt fully-supervised methods for visual concept recognition, that is to say, many BBox and relationship-triplet labels are used for training which are very expensive to annotate. Therefore, our proposed image-level weakly-supervised methods for UIC is a more cost-effective approach. On the other hand, our proposed method is significantly better than [2], [15], [39], [63], [64], although these methods are also developed based on millions of much more costly labels, as shown in Table VI.

In addition, we compare our model with Feng et al. [2] by leveraging the same dataset settings at all stages. Specifically, we all adopt the OpenImage V2 with the same 545 object categories for the visual concept recognition, and utilize the COCO dataset for the training of UIC model. For the sake of fairness, we removed the WS-RR module and compared it with [2] under the same configuration. As shown in Table II, we can achieve better performance than Feng et al. [2] under all five evaluation metrics. More importantly, our model requires zero BBox labels, however, almost 0.7 million BBox labels are leveraged for the training of [2]. We also compare with Feng et al. [2] by adopting the independent datasets, i.e., COCO images with Shutterstock sentences. As illustrated in Table III, our method is also better than theirs considering the almost two times of labeling cost.

### Table I

| Supervision Method | B4 | M | R | C | S |
|-------------------|----|---|---|---|---|
| Pivoting [39]     | 5.4| 13.2| - | 17.7| - |
| Song et al. [63]  | 11.1| 14.2| - | 28.2| - |
| ConSen [2]        | 11.3| 15.7| 37.9| 33.9| 9.1 |
| Guo et al. [64]   | 16.8| 55.3| - | - | - |
| Feng et al. [2]   | 18.6| 15.7| 43.1| 54.9| 11.1 |
| Laina et al. [15] | 19.3| 20.2| 45.0| 61.8| 12.9 |
| Cao et al. [19]   | 21.9| 21.1| 46.5| 64.0| 14.5 |
| Gu et al. [17]    | 21.5| 20.9| 47.2| 69.5| 15.0 |
| SCS [13]          | 22.8| 21.4| 47.7| 74.7| 15.1 |

### Table II

| Supervision Method | B4 | M | R | C | S |
|-------------------|----|---|---|---|---|
| Fully-supervised  | ConSen [2] | 11.3| 15.7| 37.9| 33.9| 9.1 |
| Weakly-supervised | WS-UIC w/o relationship | 19.0| 18.6| 43.8| 55.7| 11.4 |

### Table III

| Sup | Method | B1 | B2 | B3 | B4 | M | R | C | S |
|-----|--------|----|----|----|----|---|---|---|---|
| Fully Sup | ConSen [2] | 37.2| 20.0| 9.6| 4.7| 12.3| 27.3| 22.5| 8.2 |
| Weakly Sup | WS-UIC | 41.3| 22.4| 11.4| 5.9| 12.0| 28.0| 26.9| 7.6 |

*“Sup.” means supervision.

### Table IV

| Sup | Method | B1 | B2 | B3 | B4 | M | R | C | C |
|-----|--------|----|----|----|----|---|---|---|---|
| Fully Sup | Phrase-based [65] | 6.0| 3.7| 2.2| 1.4| - | - | - | - |
| Weakly Sup | m-RNN-AlexNet [66] | 5.4| 3.6| 2.3| 1.5| - | - | - | - |
|              | m-RNN-VGGNet [66] | 6.0| 4.1| 2.8| 1.9| - | - | - | - |
|              | SCS [13] | - | - | 6.3| 11.7| 29.2| 11.4| - | - |
| Fully Sup | VL-T5 [67] | - | - | - | 2.0| - | - | 2.6| - |
| Weakly Sup | WS-UIC | 39.4| 20.1| 8.8| 4.0| 8.2| 24.8| 5.8| - |

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
TABLE V
Ablation Study to Evaluate the Effectiveness of Each Component in WS-UIC. The "obj" Means Using the Object Concepts to Train the UIC Model. The "rel" Indicates Utilizing the Relationship Concepts to Train the UIC Model. The "UnO" Represents Leveraging the Unrecognized Object Loss to Train the UIC Model. The WS-UIC Means Adopting All These Schemes Together to Train the UIC Model and is Optimized by the Pseudo Captions Finally

| Approach          | Evaluation Metrics |
|-------------------|--------------------|
|                   | B4     | M   | R   | C   | S   |
| Obj               | 19.3   | 19.3 | 43.6 | 60.7 | 12.9 |
| Obj + Rel         | 19.7   | 19.6 | 44.9 | 62.1 | 13.0 |
| Obj + UnO         | 19.9   | 19.6 | 44.8 | 61.8 | 13.0 |
| Obj + UnO + Rel   | 20.5   | 20.0 | 45.5 | 64.1 | 13.3 |
| WS-UIC            | 21.5   | 20.1 | 45.8 | 65.7 | 13.6 |

The bold values denote the best-performing results.

TABLE VI
The Number of Mainly Additive Labels, i.e., Image Caption Pairs, BBox Labels, Relationship-Triplet Labels, and Image-Level Labels Utilized in Different Approaches. All of the Previous Approaches Depended on Millions of at Least One of the Former Three Expensive Labels, But WS-UIC Requires Only Cheap Image-Level Labels

| Approach   | # Additive Labels (M) | Cost (Dollar) |
|------------|-----------------------|---------------|
|            | Image-captation pair  | BBox          | Relationship-triplet | Image-level |
| Pivoting   | 0.24                  | 0             | 0                    | 0           |
| Song et al. | 1.82                | 0             | 0                    | 0           |
| Guo et al. | 0.12                  | 0             | 0                    | 0           |
| Peng et al. | 0.71                | 0             | 0                    | 0           |
| Laina et al. | 1.05               | 0             | 0                    | 0           |
| Cao et al. | 0.38                  | 0             | 0                    | 0           |
| Gu et al. | 0.34                  | 0             | 0                    | 0           |
| SCS [13]   | 0.34                  | 0             | 0                    | 0           |
| WS-UIC     | 0.34                  | 0             | 0                    | 0           |

The experimental performance is improved compared with approaches Obj+Rel, Obj, and Obj+UnO. Moreover, the experiments with the Flickr30 k dataset are carried out and compared with the related methods for image captioning using the Flickr30 k dataset. The sentence dataset used in the experiments is the conceptual captions [36]. As shown in Table IV, the proposed method achieves 4.0, 8.1, 25.3, and 5.4 on B4, M, R, and C, respectively, even better than some fully-supervised captioning models, such as m-RNN-AlexNet [66] and m-RNN-VGGNet [66], which illustrates the effectiveness of the proposed WS-UIC.

D. Comparisons on the Labeling Cost of Related Approaches

We roughly compute the number of main additive labels adopted in different approaches, containing Pivoting [39], Song et al. [63], Guo et al. [64], Peng et al. [2], Laina et al. [15], Cao et al. [19], Gu et al. [17], SCS [13], and WS-UIC, shown in Table VI. There are mainly four types of additive labels adopted in the related approaches, including BBox labels, image-caption pairs, relationship-triplet labels, and image-level labels. From Table VI, we can clearly see that all of the previous approaches depended on enormous of the former three costly labels. In sharp contrast, WS-UIC relies on only cheap image-level labels. For example, Gu et al. [17] utilized around 3.84 million BBox labels and 2.35 million relationship-triplet labels, and SCS [13] employed about 3.84 million BBox labels.

To compare the labeling prices of all these related approaches, we roughly compute them according to the Built-in workflow with Amazon Mechanical Turk. The labeling price per object per review instance is clearly marked, including $0.012 for each image-label in the image classification task, $0.036 for each BBox label, etc. Since each relationship-triplet label consists of three BBox labels, i.e., (subject, relationship, object), so we regard its unit price as $0.036 * 3 = $0.108 although we know the actual price is higher than $0.108. And Amazon Mechanical Turk recommends using multiple labelers per object to improve label accuracy. Thus, we assume using three labelers and the labeling cost of each approach is shown in the last column of Table VI. For example, the labeling cost for Gu et al. [17] is (3.84 M * $0.036 + 2.35 M * $0.108) * 3 ≈ $1176.1 k. From the Table, we can see that the labeling price of our work is significantly less than other works. For instance, the labeling price of Gu et al. [17] is about 30 times that of ours, and the labeling price of SCS [13] is almost 10 times that of ours. To summarize, WS-UIC is cost-effective and has the best generalization ability among all these approaches.

E. Ablation Studies

1) Component Analysis: To help readers have a deeper understanding of our model, we conduct extensive experiments for the component analysis on the test split of the COCO dataset. As shown in Table V, we compare with the following models:

Obj: The WS-OR module alone is applied in this experiment, which is employed to recognize the object instances. Then, these object instances provide accurate object-word alignments with images for the UIC model. As a result, the approach leads to the worst results.

Obj+Rel: Both the WS-OR model and the WS-RR model are used in the experiment. The relationship classifier reasons the association information between different objects. Then, these relationship concepts can be implemented to provide more word alignments with images. The boosted experimental performance shows that the relationship concepts can be implemented to benefit the model training.

Obj+UnO: Besides the WS-OR, the unrecognized object loss is also leveraged to train the UIC. If one object concept is contained in the captions generated by the UIC model, but not recognized by the WS-OR, a loss will be given to the UIC model. The experimental performance is improved compared with approach Obj, which verifies the usefulness of the unrecognized object loss.

Obj+UnO+Rel: All the modules of WS-UIC are adopted in this approach, including the WS-OR model, the WS-RR model, and the unrecognized object loss. The experimental performance outperforms all aforementioned approaches under the same unpaired dataset settings, which illustrates the effectiveness of the combination of these models.
WS-UIC: Besides the WS-OR model, the WS-RR model, and the unrecognized object loss, we filtered out the captions with at least one object concept as the pseudo captions to train a fully-supervised image captioning model with the cross-entropy loss. The experimental performance surpasses all other approaches, which validates the promising ability of the proposed WS-UIC.

The performance of different UIC approaches exhibits that the proposed WS-OR and WS-RR have the ability to learn the alignments between the words and images. In addition, the experiments related to the unrecognized object loss also illustrate the capability of avoiding the captions containing wrong object concepts.

2) Effect of the GNN-BR in WS-RR: To illustrate the effectiveness of the GNN-BR in the WS-RR module, three experiments without using the unrecognized object loss are conducted to do comparisons, including the experiment removing GNN-BR (W/o GNN-BR), the experiment with standard GNN (GNN), and the experiment with GNN-BR (GNN-BR). The experiment of W/o GNN-BR means directly feeding the image features extracted from the CNN backbone into the classification network, without extracting the node and edge features as done in the GNN-BR module. The experiment of GNN-BR means directly feeding the image features extracted from the CNN backbone into the classification network, without extracting the node and edge features as done in the GNN-BR module. The experiment of GNN means to implement the standard GNN. Compared to the standard GNN [32], we add the batch normalization layer and the residual block, therefore, we can get the GNN-BR. As shown in Table VII, our newly proposed GNN-BR beats the W/o GNN-BR model and the GNN model on all the five evaluation metrics. For example, we get 19.7 and 62.1 in terms of B4 and C metrics, respectively, while the method W/o GNN-BR only obtains 13.1 and 31.7, and the GNN model only achieves 18.6 and 59.5. These experiments verify the effectiveness of our improved GNN model.

Moreover, the experiments of implementing the transformer layers, instead of the GNN-BR layers, in the WS-RR module are carried out. Specifically, a multi-head self-attention layer and a feed forward layer are used. The initialization pipeline is not utilized in these experiments. As shown in Table VIII, we can find that the performance of these two versions is comparable. The proposed GNN-BR obtains better results on the M, C, and S metrics. Besides, two examples of the qualitative results are exhibited in Fig. 7. From the figure, we can observe that the captioning results without utilizing the GNN-BR module are worse than the model using the GNN-BR module.

3) Effect of Unrecognized Object Loss: To illustrate the usefulness of the unrecognized object loss, the number of unrecognized objects in the training process are computed and exhibited in Fig. 8. We can find that the number of unrecognized object categories is decreasing slightly when the loss is utilized in the model, containing about 3.84 unrecognized object concepts on average. The number of unrecognized object concepts is not decreasing when the UIC model without the loss, containing about 3.93 unrecognized object concepts on average. The number of unrecognized object concepts is so high even we use the unrecognized object loss. It is because that we only filter out concepts with much high confident scores and ignore the concepts with low confident scores. All in all, our proposed unrecognized object loss contributes to the UIC model and generates fewer unrecognized object classes.

4) Influence of the WS-OR / WS-RR Module: The WS-OR and WS-RR are two key components in our model, which may affect the final results significantly. To further analyze their influence, two different settings are considered in this section. Specifically, we use different amount of training samples and different backbone networks for this experiment.

For the model trained with different amounts of training samples, we randomly select a subset of samples from the Visual Genome dataset to train our model and compare it with the full dataset version. As shown in Table IX, we can find that the corresponding results become worse. This may be because the
WS-OR module becomes weaker when fewer training samples are used. As wrongly recognized object concepts may lead to accumulated errors for the WS-RR module, the final UIC performance drops.

For the model equipped with different backbone networks, i.e., the ResNet-50 and ResNet-152. As shown in Table IX, the experimental results of ResNet-152 are significantly better than the ResNet-50 version. Specifically, the ResNet-50 based model achieves 16.6, 17.5, 41.7, 50.0, and 10.8 on the B4, M, R, C, and S metrics. When the ResNet-152 is adopted as the backbone, the performance can be improved to 18.0, 18.4, 42.8, 55.2, and 11.4, where the improvements are 1.4, 0.9, 1.1, 5.2, and 0.6, respectively. It is because the ResNet-152 has more parameters pre-trained using a classification task on the ImageNet dataset and stronger feature representation validated by many existing works. Therefore, more accurate object concepts and relationship concepts can be obtained using our proposed WS-OR and WS-RR, which further improved our final captioning performance.

From all the experimental results and analysis, we can draw the conclusion that stronger WS-OR/WS-RR modules are needed for the UIC task. One can further improve the final performance by introducing powerful backbone networks and larger scale training samples, such as the pre-trained big models.

### F. Qualitative Results

1) **Visualization of WS Visual Concept Recognition:** In this section, to give a qualitative evaluation on the WS visual concept recognition, we provide some visualization of object and relation information predicted by our WS-OR and WS-RR modules. As shown in Fig. 9, we can find that the WS-OR is able to differentiate the salient and main object categories with locations. These examples also illustrate the strong ability of WS-RR for recognizing the relations between two objects. Let us take the first image as an example, the object category “wave” and “man” are clearly and correctly identified by the WS-OR. For the relationship between the “wave” and “man,” the concept “in” is given by the WS-RR model considering the location information of the two objects, which demonstrates the promising differentiation capability of WS visual concept recognition. All these examples fully demonstrate the effectiveness of our model.

2) **Qualitative Results of WS-UIC:** As shown in Fig. 10, we provide some representative captions generated with multiple models, including Obj, Obj+Rel, Obj+UnO, Obj+UnO+Rel, and WS-UIC. From these results, we can find that these UIC models can generate reasonable captions by using the objects and the relationships recognized through the proposed WS visual concept recognition scheme. Take the first image as an example,
V. LIMITATIONS AND DISCUSSIONS

Despite that the unpaired image captioning is achieved by utilizing the proposed model of WS visual concept recognition, some issues are still required to be addressed. The recognized visual concepts can provide paired visual concepts for images, but the WS schemes cannot accurately capture each of the salient concept information. For instance, the “table” and “flower” are not recognized in the first image of Fig. 11. The accuracy of the visual concepts directly influences the performance of unpaired image captioning. A performance gap still exists between WS-UIC and some of the approaches with stronger supervision. It will be a promising direction to utilize more advanced weakly-supervised segmentation models for extracting visual concepts [68] or design a single-stage optimization paradigm [69] for UIC.

The second one is about the quality of the generated descriptions. Given the recognized concepts, the ability of the UIC model plays a vital role in describing images. A better model is
that it can generate a correctly structured description with reasonable rhetoric by aligning the image features and these concepts. However, our model needs to be improved at this aspect according to the simple sentences with wrong concepts exhibited in Fig. 12. In our future works, we will exploit strong network architectures (e.g., the Transformer [70], [71]), for better image caption generation.

VI. CONCLUSION

In this article, we have presented a novel cost-effective framework for weakly-supervised unpaired image captioning (WS-UIC). In contrast to the conventional UIC approaches that rely on expensive training labels, the proposed method is able to generate detailed descriptions for an image given only image-level labels. Besides, an enhanced graph neural network and a novel unrecognized object loss were designed to improve the performance of unpaired image captioning. Extensive experiments have been conducted to demonstrate the effectiveness of the proposed WS-UIC scheme by comparing it with several competitive counterparts that require much stronger supervision. The proposed scheme can be further enhanced by incorporating some more sophisticated strategies, such as cross-alignment between the vision domain and the language domain.

REFERENCES

[1] 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., 2015, pp. 3156–3164.

[2] Y. Feng, L. Ma, W. Liu, and J. Luo, “Unsupervised image captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 4125–4134.

[3] J. Zhang, K. Mei, Y. Zheng, and J. Fan, “Integrating part of speech guidance for image captioning,” IEEE Trans. Multimedia, vol. 23, pp. 92–104, 2021.

[4] J. Wu et al., “Fine-grained image captioning with global-local discriminative objective,” IEEE Trans. Multimedia, vol. 23, pp. 2413–2427, 2021.

[5] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 652–663, Apr. 2017.

[6] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, “A comprehensive survey of deep learning for image captioning,” ACM Comput. Surv., vol. 51, no. 6, pp. 1–36, 2019.

[7] A. Das et al., “Visual dialog,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 326–335.

[8] S. Fidler, “Teaching machines to describe images with natural language feedback,” Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[9] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, “Textual explanations for self-driving vehicles,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 563–578.

[10] X. Wang, W. Xiong, H. Wang, and W. Y. Wang, “Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 37–53.

[11] S. Wu, J. Wieland, O. Farivar, and J. Schiller, “Automatic alt-text: Computer-generated image descriptions for blind users on a social network service,” in Proc. ACM Conf. Comput. Supported Cooperative Work Soc. Comput., 2017, pp. 1180–1192.

[12] D. Gurari et al., “Vizwiz grand challenge: Answering visual questions from blind people,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 3608–3617.

[13] H. Ben et al., “Unpaired image captioning with semantic-constrained self-learning,” IEEE Trans. Multimedia, vol. 24, pp. 904–916, 2022.

[14] C. Gao et al., “Room-and-object aware knowledge reasoning for remote embodied referring expression,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 3064–3073.

[15] I. Laina, C. Rupprecht, and N. Navab, “Towards unsupervised image captioning with shared multimodal embeddings,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 7414–7424.

[16] D. Guo, Y. Wang, P. Song, and M. Wang, “Recurrent relational memory network for unsupervised image captioning,” in Proc. 29th Int. Conf. Int. Joint Conf. Artif. Intell., 2021, pp. 920–926.

[17] J. Gu et al., “Unpaired image captioning via scene graph alignments,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 10323–10332.

[18] F. Liu, M. Gao, T. Zhang, and Y. Zou, “Exploring semantic relationships for image captioning without parallel data,” in Proc. IEEE Int. Conf. Data Mining, 2019, pp. 439–448.

[19] S. Cao, G. An, Z. Zheng, and Q. Ruan, “Interactions guided generative adversarial network for unsupervised image captioning,” Neurocomputing, vol. 417, pp. 419–431, 2020.

[20] P. G. Ipeirotis, F. Provost, and J. Wang, “Quality management on amazon mechanical turk,” in Proc. ACM SIGKDD Workshop Hum. Comput. Interact., 2010, pp. 64–67.

[21] 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.

[22] M. Yang et al., “Multitask learning for cross-domain image captioning,” IEEE Trans. Multimedia, vol. 21, no. 4, pp. 1047–1061, Apr. 2019.

[23] J. Mao et al., “Learning like a child: Fast novel visual concept learning from sentence descriptions of images,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 2533–2541.

[24] T. Yao, Y. Pan, Y. Li, and T. Mei, “Incorporating copying mechanism in image captioning for learning novel objects,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6580–6588.

[25] L. A. Hendricks et al., “Deep compositional captioning: Describing novel object categories without paired training data,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 1–10.

[26] X. Yang, H. Zhang, and J. Cai, “Deconfounded image captioning: A causal retrospective,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 1, 2021.

[27] T.-Y. Lin et al., “Microsoft coco: Common objects in context,” Proc. ACM Int. Conf. Comput. Vis., 2014, pp. 740–755.

[28] A. Kuznetsova et al., “The open images dataset v4,” Int. J. Comput. Vis., vol. 128, no. 7, pp. 1956–1981, 2020.

[29] J. Deng et al., “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Int. Conf. Comput. Vis., 2009, pp. 248–255.

[30] F. Baldassarre, K. Smith, J. Sullivan, and H. Aizpurua, “Explanation-based weakly-supervised learning of visual relations with graph networks,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 612–630.

[31] F. Liu et al., “Bridging the gap between vision and language domains for improved image captioning,” in Proc. 26th ACM Int. Conf. Multimedia, 2020, pp. 4153–4162.

[32] J. Ahn, S. Cho, and S. Kwak, “Weakly supervised learning of instance segmentation with inter-pixel relations,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 2209–2218.

[33] L. Wu, M. Xu, J. Wang, and S. Perry, “Recall what you see continually using GridLSTM in image captioning,” IEEE Trans. Multimedia, vol. 22, no. 3, pp. 808–818, Mar. 2020.

[34] L. Guo, J. Liu, S. Lu, and H. Lu, “Show, tell, and polish: Ruminant de-coding for image captioning,” IEEE Trans. Multimedia, vol. 22, no. 8, pp. 2149–2162, Aug. 2020.

[35] N. Xu et al., “Multi-level policy and reward-based deep reinforcement learning framework for image captioning,” IEEE Trans. Multimedia, vol. 22, no. 5, pp. 1372–1383, May 2020.

[36] 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, 2018, pp. 2556–2565.
[37] W. Chen, A. Lucchi, and T. Hofmann, “A semi-supervised framework for image captioning,” 2016, arXiv:1611.05321.

[38] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon, “Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach,” in Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP), 2019, pp. 2012–2023.

[39] J. Gu, S. Joty, J. Cai, and C. Wang, “Unpaired image captioning by language pivoting,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 503–519.

[40] J. Huang et al., “Speed/accuracy trade-offs for modern convolutional object detectors,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 7310–7311.

[41] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[42] B. Singh and L. S. Davis, “An analysis of scale invariance in object detection-snip,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 3578–3587.

[43] S. Venugopalan et al., “Captioning images with diverse objects,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5755–5761.

[44] G. Lample, A. Conneau, L. Denoyer, and M. Ranzato, “Unsupervised machine translation using monolingual corpora only,” Int. Conf. Learn. Representations, 2018.

[45] X. Yang, K. Tang, H. Zhang, and J. Cai, “Auto-encoding scene graphs for image captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 10685–10694.

[46] N. Komodakis and S. Zagoruyko, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer,” Int. Conf. Learn. Representations, 2017.

[47] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[48] P. Yang et al., “SGM: Sequence generation model for multi-label classification,” in Proc. 27th Int. Conf. Comput. Linguistics, 2018, pp. 3915–3926.

[49] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Deep features for discriminative localization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 2921–2929.

[50] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 448–456.

[51] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, Inception-ResNet and the impact of residual connections on learning,” in Proc. 31st AAAI Conf. Artif. Intell., 2017, pp. 4278–4284.

[52] Z. Zhang, Q. Wu, Y. Wang, and F. Chen, “High-quality image captioning with fine-grained and semantic-guided visual attention,” IEEE Trans. Multimedia, vol. 21, no. 7, pp. 1681–1693, Jul. 2019.

[53] Q. Y. Kenny et al., “Indicator function and its application in two-level factorial designs,” Ann. Statist., vol. 31, no. 3, pp. 984–994, 2003.

[54] R. Krishna et al., “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” Int. J. Comput. Vis., vol. 123, no. 1, pp. 32–73, 2017.

[55] K. Andrej and F.-F. Li, “Deep visual-semantic alignments for generating image descriptions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 3128–3137.

[56] B. A. Plummer et al., “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 2641–2649.

[57] 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.

[58] M. Denkowski and A. Lavie, “Meteor universal: Language specific translation evaluation for any target language,” in Proc. 9th Workshop Stat. Mach. Transl., 2014, pp. 376–380.

[59] C.-Y. Lin and F. Och, “Looking for a few good metrics: ROUGE and its evaluation,” in Proc. 4th NTITC Workshops, 2004, pp. 1–8.

[60] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, “Cider: Consensus-based image description evaluation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 4566–4575.

[61] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “Spice: Semantic propositional image caption evaluation,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 382–398.

[62] I. Tolstikhin et al., “Mlp-mixer: An mlp-mixer for all: Architecture for vision,” Adv. Neural Inf. Process. Syst., vol. 34, pp. 24261–24272, 2021.

[63] Y. Song, S. Chen, Y. Zhao, and Q. Jin, “Unpaired cross-lingual image caption generation with self-supervised rewards,” in Proc. 27th ACM Int. Conf. Multimedia, 2019, pp. 784–792.
Zhenglong Sun (Senior Member, IEEE) received the B.E. degree in biomedical engineering, the M.E. and Ph.D. degrees in mechatronic system and design from Nanyang Technological University, Singapore. He is currently an Assistant Professor with the School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen), Shenzhen, China. He is also the Deputy Director of the Healthcare Robotics Center, Shenzhen Institute of Artificial Intelligence and Robotics for Society. He has authored or coauthored more than 70 journal and conference papers. His research interests include medical/surgical robot, haptic sensing and feedback, and human-machine interaction. He is a Reviewer of a number of journals and conferences, such as IEEE TRO, TMECH, TMRB, ICRA, IROS, IJCAI, and AAAI.

Wei-Shi Zheng received the Ph.D. degree in applied mathematics from Sun Yat-sen University, Guangzhou, China, in 2008. He is currently a Full Professor with Sun Yat-sen University. His research interests include person or object association and activity understanding in visual surveillance, and the related large-scale machine learning algorithm. He is an Associate Editor for the Pattern Recognition Journal and the Area Chair of a number of top conferences. He joined Microsoft Research Asia Young Faculty Visiting Programme. He was the recipient of the Excellent Young Scientists Fund of the National Natural Science Foundation of China and Royal Society-Newton Advanced Fellowship of the United Kingdom.

Yaowei Wang (Member, IEEE) received the Ph.D. degree in computer science from the Graduate University of Chinese Academy of Sciences, Beijing, China, in 2005. He is currently a tenured Associate Professor with Peng Cheng Laboratory, Shenzhen, China. From 2005 to 2019, he was with the Department of Electronics Engineering, Beijing Institute of Technology, Beijing, China. In 2019, he was a Professor with the National Engineering Laboratory for Video Technology Shenzhen, Peking University Shenzhen Graduate School, Shenzhen, China. From 2014 to 2015, he was an Academic Visitor with the Vision Lab, Queen Mary University of London, London, U.K. He is the author or co-author of more than 100 refereed journals and conference papers. His research interests include machine learning, multimedia content analysis, and understanding. He was the recipient of the second prize of the National Technology Invention in 2017 and the first prize of the CIE Technology Invention in 2015. His team was ranked as one of the best performers in the TRECVID CCD/SED tasks from 2009 to 2012 and in PETS 2012. He is a Member of CCF, CIE, and CSIG.

Changwen Chen (Fellow, IEEE) received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 1983, the M.S.E.E. degree from the University of Southern California, Los Angeles, CA, USA, in 1986, and the Ph.D. degree from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 1992. He is currently the Chair Professor of visual computing with The Hong Kong Polytechnic University, Hong Kong. Before his current position, he was the Dean of the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China, from 2017 to 2020. From 2008 to 2021, he was also an Empire Innovation Professor with the University at Buffalo, Buffalo, NY, USA, and the State University of New York, Albany, NY, USA. From 2003 to 2007, he was an Allen Henry Endow Chair Professor with the Florida Institute of Technology, Melbourne, FL, USA. He was with the Faculty of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA, from 1992 to 1996 and the Faculty of Electrical and Computer Engineering, University of Missouri-Columbia, Columbia, MO, USA, from 1996 to 2003. He is the Editor-in-Chief of IEEE TRANSACTIONS ON MULTIMEDIA from January 2014 to December 2016, and the Editor-in-Chief of IEEE TRANSACTIONS ON VIDEO TECHNOLOGY from January 2006 to December 2009. He is the Editor of several other major IEEE Transactions and Journals, including the PROCEEDINGS OF IEEE, IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS, and IEEE JOURNAL OF EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS. He was the Conference Chair of several major IEEE, ACM, and SPIE conferences related to multimedia communications and signal processing. His research has been funded by both government agencies and industrial corporations. He and his students have been the recipient of nine Best Paper Awards or Best Student Paper Awards over the past two decades. He has also been the recipient of several research and professional achievement awards. These include the Sigma Xi Excellence in Graduate Research Mentoring Award in 2003, Alexander von Humboldt Research Award in 2009, University at Buffalo Exceptional Scholar – Sustained Achievement Award in 2012, SUNY System Chancellor’s Award for Excellence in Scholarship and Creative Activities in 2016, and University of Illinois ECE Distinguished Alumni Award in 2019. He is an SPIE Fellow, and a Member of the Academia Europaea.