Automatically Generating Natural Language Descriptions of Images by a Deep Hierarchical Framework

Lin Huo, Lin Bai, and Shang-Ming Zhou, *Member, IEEE*

***Abstract*—Automatically generating an accurate and meaningful description of an image is very challenging. However, the recent scheme of generating an image caption by maximizing the likelihood of target sentences lacks the capacity of recognizing the human–object interaction (HOI) and semantic relationship between HOIs and scenes, which are the essential parts of an image caption. This article proposes a novel two-phase framework to generate an image caption by addressing the above challenges: 1) a hybrid deep learning and 2) an image description generation. In the hybrid deep-learning phase, a novel factored three-way interaction machine was proposed to learn the relational features of the human–object pairs hierarchically. In this way, the image recognition problem is transformed into a latent structured labeling task. In the image description generation phase, a lexicalized probabilistic context-free tree growing scheme is innovatively integrated with a description generator to transform the descriptions generation task into a syntactic-tree generation process. Extensively comparing state-of-the-art image captioning methods on benchmark datasets, we demonstrated that our proposed framework outperformed the existing captioning methods in different ways, such as significantly improving the performance of the HOI and relationships between HOIs and scenes (RHIS) predictions, and quality of generated image captions in a semantically and structurally coherent manner.**

***Index Terms*—Human–object interaction (HOI), hybrid deep learning, image captioning, image context, natural language processing.**

Manuscript received February 6, 2020; revised May 30, 2020 and September 10, 2020; accepted November 23, 2020. This work was supported in part by the Major Project of National Social Science Foundation of China under Grant 16ZDA0092; in part by the National Natural Science Foundation of China under Grant 61966003; in part by the Guangxi Natural Science Foundation under Grant 2018GXNSFAA138085; and in part by the Guangxi High-Level Innovation Team in Higher Education Institutions "Digital ASEAN Cloud Big Data Security and Mining Technology" Innovation Team. This article was recommended by Associate Editor F. Wu. *(Corresponding authors: Shang-Ming Zhou; Lin Bai.)*

Lin Huo is with the China-ASEAN Research Institute, Guangxi University, Nanning 530004, China (e-mail: lhuo@gxu.edu.cn).

Lin Bai is with the School of Computer, Electronics and Information, Guangxi University, Nanning 530004, China (e-mail: bailin@gxu.edu.cn).

Shang-Ming Zhou was with the Scottish Digital Health and Care Institute, University of Strathclyde, Glasgow G1 1XQ, U.K., and also with the Department of Computer and Information Sciences, University of Strathclyde, Glasgow G1 1XQ, U.K. He is now with the Centre for Health Technology, University of Plymouth, Plymouth PL4 8AA, U.K. (e-mail: shangming.zhou@plymouth.ac.uk; smzhou@ieee.org).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TCYB.2020.3041595.

Digital Object Identifier 10.1109/TCYB.2020.3041595

I. Introduction

AUTOMATICALLY generating accurate and meaningful descriptive sentences of an image, referred to as image captioning, could have great impacts in different domains, for instance, by assisting visually impaired people in various aspects of life (e.g., shopping and walking), improving visual data retrieval, providing more general knowledge about the visual world implicitly encoded in human language, etc. However, image captioning is challenging, which involves machine learning, computer vision, and natural language processing. Specifically, image captioning entails a tradeoff between several design objectives [1]: how to represent an image, how to represent the sentences (i.e., language modeling), and how to fuse the visual and textual information, encoded by the images and the sentences, respectively. This involves the tasks of not only identifying the objects contained in an image (corresponding to the *nouns* in the caption) but also identifying the relationships between the objects and the appearance of the objects (corresponding to linguistic constituents, such as *verbs* in the caption), moreover expressing the above semantic knowledge in a natural language. A core issue of such relationships is how to identify the human–object interaction (HOI) and the associated RHIS [2], [3].

Some pioneering approaches have been developed to address these image captioning challenges. Template-based models [4] rely on explicitly predefined templates and hard-coded visual concepts to generate sentences by filling detected visual elements, such as objects. Retrieval-based models [5], first learn joint embedding to map both sentences and images into the same semantic space and then retrieve the similar sentences to describe the query image. However, the generated sentences by these methods often have a very limited variety and cannot describe specific contents in an image. Moreover, these methods cannot generate descriptive sentences that should depict the verb and the adverbial compositions [6], [7]. As a result, they cannot always generate very realistic sentences that capture all image concepts.

Recently, a surge of interest focused on a scheme of inferring the most likely sentence of words $S = \{S_1, S_2, \ldots\}$ from an image I by maximizing the likelihood $P(S|I)$, where each word S_t comes from a given dictionary that describes the image adequately [8]. However, this scheme cannot precisely identify the HOI and the relationship between an HOI and a scene [9]–[11].

In this article, we propose a novel deep-learning framework to generate natural language descriptions of images automatically by addressing the challenges of the HOI identification, RHIS prediction, and descriptive sentence generation. The novelty of this framework lies in that it consists of two phases to generate image caption: 1) a hybrid deep learning and 2) a description generation. The hybrid deep-learning phase first uses a faster region-based convolutional neural network (RCNN) to learn the features of objects at the low level. Then, based on the low-level object features, we proposed a novel factored three-way interaction machine (FTWIM) to learn the relational features of the human–object pairs hierarchically with their 3-D spatial configuration at the high level. In this way, the recognition problems are transformed into latent structured labeling tasks in a unified max-margin learning problem. The description generation phase uses the HOI and the RHIS as key ingredients and innovatively integrated with a syntactic tree scheme to build an image description generator. In this way, the description generation becomes a tree-based lexicalized syntactic derivation process integrating both the tree structure and sequence cohesion into tree growth inference, so that a substantially higher level of linguistic expressiveness, precision, and flexibility than the previous studies can be guaranteed.

In the remainder of this article, we first review the related work in Section II. Section III gives an overview of our proposed two-phase framework. The hybrid deep-learning model, model inference, and learning are presented in Section IV. Section V presents the details of the description generator, including the tree-growth rules and process. Finally, experimental results are provided in Section VI, followed by conclusions.

II. Related Work

Generating descriptive sentences of an image is a fundamental task in computer vision and machine learning [8], [12], [13]. Some approaches combine tree algorithms with object detection methods to generate descriptive sentences, but these models are prone to the problem caused by false human or object detection [14]. In order to alleviate issues, some research relies heavily on extensive labor-intensive labels of objects during the training phrase [15]. A number of approaches treat this task as a retrieval task, which formulates descriptive sentences generation process as a ranking learning problem [16]. However, these approaches can only annotate the query image with descriptive sentences of the similar images already existing in the training dataset. Obviously, they cannot precisely depict the semantic relationship between the human and objects in the query image. Inspired by recent studies in machine translation, some researchers strove to build a joint probability over a large image–sentence corpus to learn the correct description. They generated the descriptive sentences word-by-word based on the order of the objects learned from experience [8], [9], [17]. However, these models focus on the compatibility between objects and words; thus making them powerless to predict the action, and the relationship between HOI and other image parts (e.g., scene).

Emerging evidence indicated that the HOIs are the core element of image contents [2], [3], [18], [19]. Recently, increasing attention is being paid to the use of image contexts to aid visual recognition [20], [21]. Some studies used objects interacting with human poses to improve action classifiers [22]. However, human poses parts may profoundly change or be self-occluded when they participate in relations. Some approaches exploited 2-D spatial context between objects to alleviate human pose issues. For example, the studies in [23] and [24] used 2-D spatial context between human and object to alleviate the shortage of human pose. Fei-Fei's team modeled the mutual context between the 2-D spatial arrangement of objects and human poses to facilitate the object detection and action recognition [25]. Some approaches encode the interaction activities by a set of 2-D spatial layout between human–object pairs [11], [26]. However, the methods based on the 2-D spatial layout share a common shortage: 2-D spatial context cannot precisely describe the spatial consistency. For example, when a person and an object are far away from each other, but look like a close or occluding one from another in a special viewpoint, 2-D spatial context would misunderstand and misinterpret their relationship in this human–object pair.

In order to generate more accurately spatial context analysis, some researchers attempted to learn 3-D locations of objects from video to improve spatial co-occurrence arrangement between humans and objects [27], while others focused on modeling the movement trajectories of objects, and used this contextual information to aid the object prediction and spatial layout analysis [28]. The most relevant research to ours is the work of [29] and [30]. They reconstructed a 3-D spatial layout of the indoor scene with the help of manual annotation, and then modeled the consistent spatial relationship between humans and objects across images [30]. However, such a method heavily depends on manual annotations, and can only work in limited domains [29].

Different from the approaches without or with limited spatial context [5], [31], we directly take 3-D spatial co-occurrence contexts between image parts as conditional information for the HOI recognition as well as the RHIS prediction. Moreover, the previous studies treated the spatial context feature extraction and the recognition separately [3]. As a result, some important image context features could be lost, and could not achieve the joint optimization performance.

Different from the separated strategy, we propose a new deep hierarchical structure that directly treats 3-D spatial contexts as an additional input to facilitate the recognition of HOI and RHIS. Our proposed structure can innovatively extract the joint features of spatial contexts and the image blocks of a human–object pair. The joint features are further used through multiple convolution-pooling layers to capture high-level relational features, which can improve the HOI and RHIS predictions. This task has been proved to be extremely difficult to implement by using the traditional methodologies [26], [31].

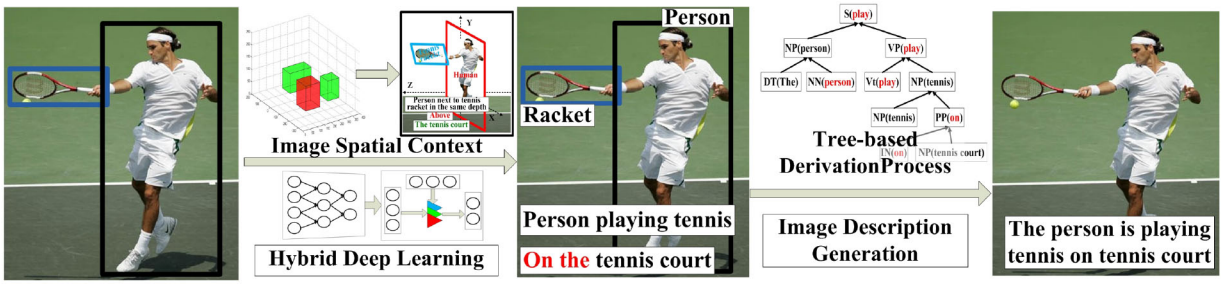

Fig. 1. Architecture of our deep understanding framework. Our framework mainly contains two parts: the hybrid deep-learning model and the description generator. The former focuses on learning the HOI and the relationship between this HOI and a scene, by modeling 3-D spatial context, in a hierarchical stage. The latter discovers the prepositional phrase to describe the relationship between various image patterns and achieve the description generation problem by using a lexicalized tree growth process.

For the image description generation, inspired by the recent advance of the syntactic tree algorithm [20], [31], our tree-based caption generator model combines the advantages of lexicalized PCFGs and the syntactic tree method. One relevant work to ours is the TreeTalk model proposed by Kuznetsova *et al.* [5] to compose expressive image descriptions by selectively combining the extracted (and optionally pruned) tree fragments, but this model cannot generate lexicalized probabilistic context-free grammars. Different from the previous methods without lexicalized tree process, our model innovatively treats the HOI and the RHIS as the key nodes to guide the process of tree composition. These properties enable our model to generate semantically well-formed descriptive sentences and capture the major contents of an image.

III. Proposed Two-Phase Framework

Our proposed framework consists of two phases to generate image captions: 1) a hybrid deep learning and 2) a description generation, as shown in Fig. 1.

In the hybrid deep-learning phase, first, a Faster RCNN is used to learn the features of objects at the low level. It has been proved that Faster RCNN [32]–[34] can produce the invariant features of objects. Then, at the high level, based on the low-level object features, we proposed a novel FTWIM to learn the relational features of the human–object pairs hierarchically with their 3-D spatial configuration, as shown in Fig. 2. These high-level relational features are robust and characterize HOIs well.

Let H, O, V, and R denote the sets of humans, objects, HOIs, and RHISs in an image, respectively. The FTWIM model takes the HOI v and the scene s as inputs and is trained to maximize the likelihood $P(r|v, s)$ of producing a target phrase r that describes the relationship between the HOIs and the scenes. In this way, the recognition problems are transformed into latent structured labeling tasks by modeling spatial context between image parts, which formulate the model learning as a unified max-margin learning problem.

Specifically, to improve the recognition of the HOI and the RHIS, we consider 3-D spatial contexts as additional information to enhance the relationship recognition. We use a 3-D spatial layout detection method, based on our previous work [3], [35], to capture the 3-D spatial contexts of an image. Thus, the HOI $v^* \in V$ can be predicted by maximizing

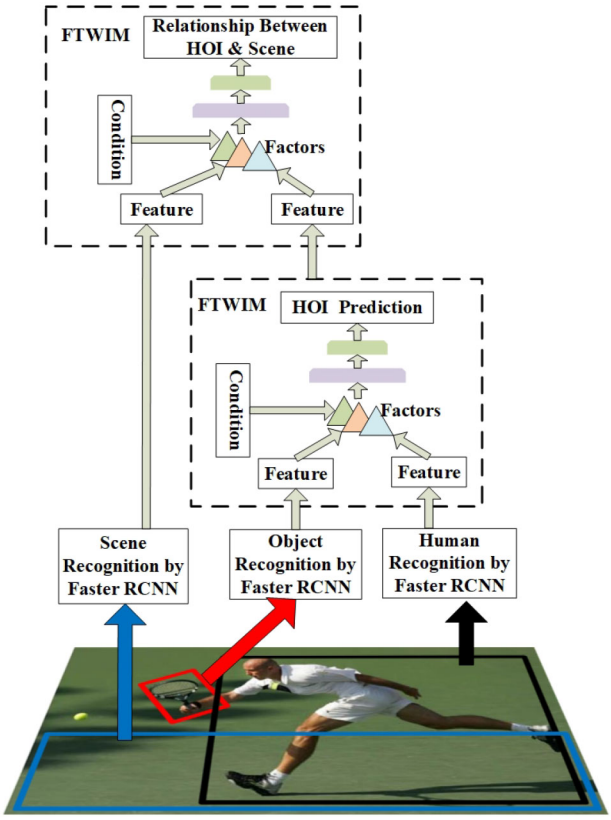

Fig. 2. Graphical illustration of our deep-learning model.

the likelihood of $P(v|h_i, o_j, \boldsymbol{l}(h_i, o_j); \alpha)$ given a human–object pair $h_i \in H$ and $o_j \in O$ and their spatial context $\boldsymbol{l}(h_i, o_j)$ conditioned on parameters α

$$v^* = \arg\max_v P(v|h_i, o_j, \boldsymbol{l}(h_i, o_j); \alpha). \quad (1)$$

The relationship $r^* \in R$ between an HOI and a scene can be learned by maximizing the probability of $P(r|v_i, s_j, \boldsymbol{l}(v_i, s_j); \beta)$ given the HOI $v_i \in V$, scene $s_j \in O$, and their spatial contextual information $\boldsymbol{l}(v_i, s_j)$ parameterized on β

$$r^* = \arg\max_r P(r|v_i, s_j, \boldsymbol{l}(v_i, s_j); \beta) \quad (2)$$

where the variables h_i and o_j represent the ith human and the jth object in the query image, respectively. The relationship r^* depicts the semantic relationship between the HOI and the

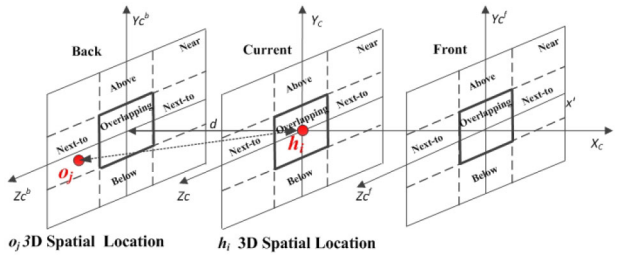

Fig. 3. Illustration of 3-D spatial configuration of a human–object pair (h_i, o_j). The center of cross represents the center of human h_i in the current depth location. We consider the location of object o_j with respect to this new coordinate frame defined by h_i. We divide the space into eight different spatial relationships that discretely detail 3-D spatial layout between the human–object pair. It turns the $\boldsymbol{l}(h_i, o_j)$ into a 8-dimension sparse binary vector. For instance, o_j is next to and back away h_i.

scene (e.g., play tennis on a tennis court). The 3-D spatial context is learned by our previous studies [3], [35], in which we proposed a discriminatively trained model that incorporates multiscale local-image features and a depth coordinate perception system to estimate 3-D depth information from a single monocular image. Thus, $\boldsymbol{l}(h_i, o_j)$ discretizes the 3-D spatial layout between h_i and o_j into some of discrete canonical relations describing as: *above, overlapping, below, near, next-to, front, current*, and *back*, as shown in Fig. 3. Hence, $\boldsymbol{l}(h_i, o_j)$ is a specific sparse binary vector of length K with some 1 for the specific elements when the spatial arrangement satisfying the human–object pair as well as $\boldsymbol{l}(v_i, s_j)$.

As shown in (1) and (2), both $P(v|h_i, o_j, \boldsymbol{l}(h_i, o_j); \alpha)$ and $P(r|v_i, s_j, \boldsymbol{l}(v_i, s_j); \beta)$ share the similar form; thus, we design a unified hybrid deep-learning model to achieve the maximum-likelihood estimation of $P(v|h_i, o_j, \boldsymbol{l}(h_i, o_j); \alpha)$ and $P(r|v_i, s_j, \boldsymbol{l}(v_i, s_j); \beta)$ in a hierarchical strategy, as shown in Fig. 2. This hybrid deep-learning model consists of three levels. At the lowest level, the Faster RCNN is used to detect people, objects, and scenes. Then, we propose a novel hierarchical structure, called FTWIM, to extract the high-level relational features. So at the second level for HOI prediction, an FTWIM model takes a pair of human and object features as inputs, and utilizes 3-D spatial layout as the conditional input. The multiple hidden layers learn the high-level relational features of the human–object pair hierarchically, under the guidance of the 3-D spatial layout, from these varieties of image features. The learned features are used by the following fully connection layers to identify the HOI. At the third level for RHIS prediction, an FTWIM model takes both the HOI and scene regions as inputs and uses a conditional layer to extract the spatial context as condition input. In this way, the second and third levels aim to learn the high-level global features of the query image, with the aid of the spatial context of the image. Finally, based on the extracted features, we propose a lexicalized-tree-driven caption sentence generation model to generate semantically well-formed image caption by connecting various visual patterns (e.g., HOI, scene, and object) syntactically.

In the description generation phase (Fig. 1), inspired by the recent studies in lexicalized probabilistic context-free grammars (lexicalized PCFGs) [5], we used the HOI and the RHIS as key ingredients and innovatively integrated with a syntactic tree scheme to build an image description generator. In this way, the description generation becomes a tree-based lexicalized syntactic derivation process that takes the HOI and RHIS as the anchor nodes to guide the application of tree composition rules. Thus, our model innovatively integrates both tree structure and sequence cohesion into tree growth inference, which guarantees a substantially higher level of linguistic expressiveness, precision, and flexibility than the previous studies.

IV. Hybrid Deep-Learning Model

Recent advances in image captioning have shown that, given a discriminative model, it is possible to learn the correct description of the query image block [3], [36]. It is natural to exploit this idea to recognize the HOI by maximizing the likelihood of the correct HOI phrase, given the features of the human–object pair, which is also applicable to the relationship between HOIs and scenes.

A. Factored Three-Way Interaction Machine

Our FTWIM consists of four layers: 1) the input layers; 2) the hidden layers; 3) the condition layers; and 4) the output layers. The factors are the key ingredient of the proposed FTWIM that looks like a three-way intersection. It modulates the interactions in three different units. Compared with the biased connecting weights between input units and hidden units [37], our model allows the condition units as a special input to directly adjust the interactions, thus truly integrate the conditional information into the deep-learning architecture. The hidden layers and the output layers constitute a deep neural network that contains three convolution-pooling layers and fully connected layers. This hierarchical structure enables the final prediction layer to predict the high-level concept of the HOI, as well as the relationship between the HOI and scene. Directly training the whole networks would lead to model overfitting; therefore, we optimized the HOI recognition and the image captioning under supervision separately.

Different from the previous studies, our proposed method of the factored three-way interactions allow the real-valued context features to control the hidden features learning directly, as shown in Fig. 2. These special structures are powerful to extract relational features of HOI via a joint probability distribution $P(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{c})$ over the input $\boldsymbol{x}$ and hidden $\boldsymbol{y}$, conditional on the context features $\boldsymbol{c}$. The input $\boldsymbol{x}$ represents the features of the human and object, and the hidden $\boldsymbol{y}$ represents the learned relation features of the human–object pair. These variables are all n-dimensional vectors. Thus, we propose a novel energy-based algorithm to describe the configuration of the interesting variables of FTWIM, in which the greater probability $P(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{c})$ indicates the more powerful FTWIM is configured to learn the high-level relational features from inputs

$$P(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{c}) = \frac{e^{-E(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{c})}}{Z} \tag{3}$$

where $Z = \sum_{\boldsymbol{x}, \boldsymbol{y}} e^{-E(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{c})}$ and $E(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{c})$ is the energy function that captures all possible correlations between the components

of FTWIM (input $\boldsymbol{x}$, hidden $\boldsymbol{y}$, and condition $\boldsymbol{c}$)

$$E(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{c}) = -\sum_f \left(\sum_j m_{jf}^x x_j\right)\left(\sum_i m_{if}^y y_i\right)\left(\sum_k m_{kf}^c c_k\right) - \sum_j b_j x_j - \sum_i d_i y_i \tag{4}$$

where f represents the index set of deterministic factors, x_j represents the state of the visible unit j, y_i represents the state of the hidden unit i, c_k represents the state of the condition unit k, and m_{if}^y represents the weight parameter that measures the strength of connection from the hidden unit i to factor f, m_{jf}^x represents the weight parameter that measures the strength of connection from the visible unit j to factor f, and m_{kf}^c represents the weight parameter that measures the strength of connection from the condition unit k to factor f. In (4), $\sum_j b_j x_j$ and $\sum_i d_i y_i$ are the bias terms with respect to the activity of the visible and hidden units.

To learn the model parameters concisely, we estimate the marginal probability distribution of $P(\boldsymbol{x}|\boldsymbol{c})$, rather than the joint conditional probability of $P(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{c})$. To this end, we propose a free-energy function $F_e(\boldsymbol{x}|\boldsymbol{c})$ to accelerate the process of parameters optimization

$$P(\boldsymbol{x}|\boldsymbol{c}) = \frac{\sum_{\boldsymbol{y}} e^{-E(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{c})}}{\sum_{\boldsymbol{x},\boldsymbol{y}} e^{-E(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{c})}} \tag{5}$$

$$F_e(\boldsymbol{x}|\boldsymbol{c}) = -\ln \sum_{\boldsymbol{y}} e^{-E(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{c})}. \tag{6}$$

Because the energy $E(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{c})$ is replaced by the $F_e(\boldsymbol{x}|\boldsymbol{c})$, the $P(\boldsymbol{x}|\boldsymbol{c})$ can be rewritten as $P(\boldsymbol{x}|\boldsymbol{c}) = \exp(-F_e(\boldsymbol{x}|\boldsymbol{c}))/Z$. Instead of computing the $P(\boldsymbol{x}|\boldsymbol{c})$, we estimate the log likelihood of $P(\boldsymbol{x}|\boldsymbol{c})$ with the summation of visible units, as shown in

$$\sum_{\boldsymbol{x}} \ln(P(\boldsymbol{x}|\boldsymbol{c})) = -\sum_{\boldsymbol{x}} F_e(\boldsymbol{x}|\boldsymbol{c}) - \sum_{\boldsymbol{x}} \ln(Z). \tag{7}$$

It is noted that there is a special inverse relationship between $\ln(P(\boldsymbol{x}|\boldsymbol{c}))$ and $F_e(\boldsymbol{x}|\boldsymbol{c})$, which implies that the energy of the FTWIM is smaller, the probability $P(\boldsymbol{x}|\boldsymbol{c})$ is bigger, then the FTWIM is stronger in relational feature extraction. According to the properties of parameter configuration of the energy-based model [37], the energy of a system is smaller, the system is more stable, and the parameter configuration of this system is more optimal. Thus, we capture the optimal estimation of the free-energy function parameters by maximizing the likelihood estimation of $\ln P(\boldsymbol{x}|\boldsymbol{c})$

$$\frac{\partial \ln l(\boldsymbol{x}|\boldsymbol{c},\delta)}{\partial \delta} = \sum_{\boldsymbol{x}} \frac{\partial \ln(P(\boldsymbol{x}|\boldsymbol{c}))}{\partial \delta} = \frac{\partial \ln(P(\boldsymbol{x}|\boldsymbol{c}))}{\partial \delta} \tag{8}$$

$$\begin{aligned}\frac{\partial \ln(P(\boldsymbol{x}|\boldsymbol{c}))}{\partial \delta} &= \frac{\sum_{\boldsymbol{y}} e^{-E(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{c})}\left(-\frac{\partial E(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{c})}{\partial \delta}\right)}{\sum_{\boldsymbol{y}} e^{-E(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{c})}} \\ &\quad - \frac{\sum_{\boldsymbol{x},\boldsymbol{y}} e^{-E(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{c})}\left(-\frac{\partial E(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{c})}{\partial \delta}\right)}{\sum_{\boldsymbol{x},\boldsymbol{y}} e^{-E(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{c})}} \\ &= E_{P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{c})}\left(-\frac{\partial E(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{c})}{\partial \delta}\right) \\ &\quad - E_{P(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{c})}\left(-\frac{\partial E(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{c})}{\partial \delta}\right)\end{aligned} \tag{9}$$

where δ is a collection of model parameters.

B. Deep Neural Networks

The hidden layers of FTWIMs consist of three convolution-pooling modules, and the output layer is a fully connected layers. They jointly mimic the primary cortex to learn the high-level relational visual features. The convolution operation and the followed max-pooling layer can be expressed as

$$\mathbf{y}_c = S\left(\sum_i \mathbf{y}_{\text{in}} * k_i + b\right) \tag{10}$$

$$\mathbf{y}_{\text{out}} = S\left(\beta \sum \mathbf{y}_c^{n\times n} + d\right) \tag{11}$$

where $\mathbf{y}_{\text{in}}$ is the previous maps, k_i is one of the trainable convolution kernels, S is a nonlinear activation function (e.g., a hyperbolic tangent sigmoid), b and d are the biases, and $\mathbf{y}_c$ is the current convolutional map. In each convolution-pooling module, the convolution map learns features from all the maps in the previous layers, and then the max-pooling map has a lower resolution, in which the noise information would have been eliminated. The cascade of convolution-pooling modules thereby extracts higher-level relational visual features as the networks go deeper.

The top fully connected layers model the joint distribution between the last feature layer of deep convolution networks $\boldsymbol{v}$, the hidden layer $\boldsymbol{h}$, and the output layer $\boldsymbol{y}$

$$P(\boldsymbol{v},\boldsymbol{h},\boldsymbol{y}) = \frac{\exp(-E(\boldsymbol{v},\boldsymbol{h},\boldsymbol{y}))}{Z} \tag{12}$$

where $E(\boldsymbol{v},\boldsymbol{h},\boldsymbol{y}) = -\boldsymbol{v}^T\boldsymbol{U}\boldsymbol{h} - \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{y} - \boldsymbol{d}^T\boldsymbol{v} - \boldsymbol{a}^T\boldsymbol{h} - \boldsymbol{b}^T\boldsymbol{y}$ is the energy function of the fully connected structure. Because of the conditional independence properties of this network structure, given the last features $\boldsymbol{v}$ and the top hidden state $\boldsymbol{h}$, the conditional probability of its output $\boldsymbol{y}$ is explicitly expressed as

$$p(y_j|\boldsymbol{v}) = \frac{e^{b_j}\prod_i(1+\exp(\sum_k u_{ik}v_k + w_{ji} + a_i))}{\sum_n e^{b_n}\prod_i(1+\exp(\sum_k u_{ik}v_k + w_{ni} + a_i))} \tag{13}$$

where y_j denotes the jth structure label. The matrices of $\boldsymbol{U}$ and $\boldsymbol{W}$ are the weight parameters that measure the connection strength from the features $\boldsymbol{v}$ to the hidden states $\boldsymbol{h}$ and from the hidden states $\boldsymbol{h}$ to the output states $\boldsymbol{y}$, respectively. The $\boldsymbol{d}^T\boldsymbol{v}$, $\boldsymbol{a}^T\boldsymbol{h}$, and $\boldsymbol{b}^T\boldsymbol{y}$ are the bias terms with respect to the activity of the three types of layer units. We discriminatively train the top fully connected layers by maximizing the likelihood of the target label y_j given the features $\boldsymbol{v}$.

V. Caption Generator

Recent advances in image description generation [9], [38], [39] show that it is possible to achieve state-of-the-art results by directly maximizing the likelihood of the correct description given the features of an image. Hence, the image caption can be learned by maximizing the probability of the correct image caption $P(g|v_i, r_j; \theta)$ given the HOI v_i and the relationship r_j between the HOI and the scene parameterized on θ

$$g^* = \arg\max_g P(g|v_i, r_j; \theta). \tag{14}$$

In the previous study [3], we have proven that the correct image caption g^* can be learned via the tree composition process, given the nouns (objects), the verbs (interactions), and

the scenes. In fact, 92% of image descriptive sentences have no more than three object nouns [40]. Therefore, if the recognized objects in an image are more than three, our model automatically divides them into several sentences that guarantee no more than three object nouns appear in one sentence simultaneously. But how to order the visual patterns in a group and how to connect these visual patterns via proper prepositional phrases are two challenge tasks.

A. Model Definition

To solve these problems and generate accurate and reliable captions, we propose a novel description generation model based on the lexicalized-tree growing strategy. We consider the tree composition as a constraint optimization problem that aims to capture the best combinations between various visual patterns (e.g., object, scene, and HOI).

Our model is defined as a 6-tuple

$$G = (N, \Sigma, S, A, U, q) \tag{15}$$

where N is a finite set of nonterminals in the grammar, Σ is a finite set of lexical items in the grammar, $S \in N$ is a distinguished start symbol, A is a set of parameters that include two types: α_{o_i,p_k} and α_{o_i,o_j,p_k}, which encode the selection and ordering of object nouns, and U is a set of tree composition rules in which each rule takes one of the following three forms.

1) $Y_1(h); Y_2(m) \rightarrow X(h)$, where $X, Y_1, Y_2 \in N; h, m \in \Sigma$.
2) $Y_1(m); Y_2(h) \rightarrow X(h)$, where $X, Y_1, Y_2 \in N; h, m \in \Sigma$.
3) $h \rightarrow X(h)$, where $X \in N; h \in \Sigma$.

In this way, for each rule $u_l \in U$, there is an associated parameter $q(u_l) \in [0, 1]$ that represents the reasonable strength of selecting the rule u_l for the configuration between α_{o_i,p_k} and $\alpha_{o_j,p_{k+1}}$. Therefore, our tree composition for every sentence aims to maximize the following objective function:

$$(u_l^*, o_i^*, o_j^*, p_k^*) = \arg\max_{u,o,p} \sum q(u_l)(\alpha_{o_i,p_k} + \alpha_{o_j,p_{k+1}} + \alpha_{o_i,o_j,p_k}) \tag{16}$$

where $\alpha_{o_i,p_k} \in [0, 1]$ measures the probability of selecting the object noun o_i for the kth position in a sentence; α_{o_i,o_j,p_k} represents the joint probability of the object noun o_i for the kth position and o_j for the $(k + 1)$th position.

However, there is a bottleneck for training the parameters of (16) in a traditional way: the number of sentences grows exponentially alongside the number of object nouns, there might not be enough training data. Fortunately, the recognized HOIs can provide the most important cues to identify the subject, object, and predicate verb for each descriptive sentence. In an HOI, the human, the interaction, and the object could be regarded as the subject, the predicate verb, and the object, respectively. Meanwhile, the last object noun is the adverbial in a sentence, which always is a scene object. Hence, the number of useful sentences is significantly smaller than the number of all possible sentences. In this way, the above-mentioned two challenge tasks can be tackled using the HOI and the RHIS.

One remaining challenge is to find the specific prepositional phrase p which best depicts the RHIS. To this end, we propose a statistical algorithm to capture the best compatibility between p and r in terms of co-occurrence frequency. Fig. 4 illustrates the configuration between HOI, object, scene, and prepositional phrase

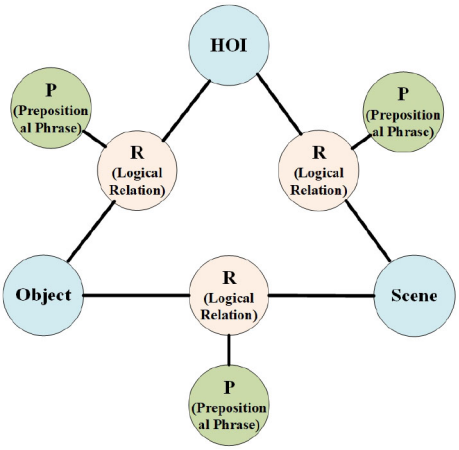

Fig. 4. Illustrating the mapping from the RHIS to the semantical space of the prepositional phrase.

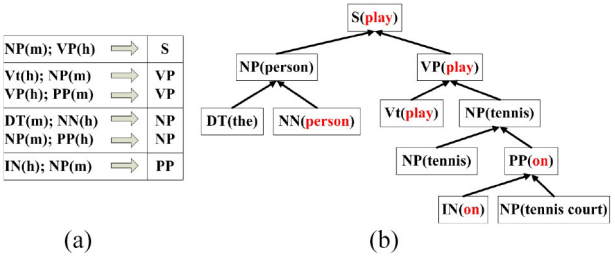

Fig. 5. Illustrating our caption generator. (a) Details rules of the tree composition, where h is the "key," and m is the modifier word. (b) Describes an example of a sentence generation process, where the red words are the "key," and the black words are the modifier word.

$$(p^*, r^*) = \arg\max_{p_g, r_j} \Phi(P, R). \tag{17}$$

The best score between p and r is obtained by searching over all possible configurations between prepositional phrases consistent with the representations of the RHISs

$$\Phi(P, R) = \Sigma_g^P \Sigma_j^R 1_{(R=r_j)} \cdot 1_{(P=p_g)} \cdot \gamma_{jg} \tag{18}$$

where P is the set of prepositional phrases and p_g represents the gth prepositional phrases; similar to R and r_j. $1_{(\cdot)}$ is an indicator function [e.g., $1_{(R=r_j)} = 1$ if R equals r_j; otherwise, 0]. γ_{jg} represents the strength of the co-occurrence interaction between r_j and p_g (the larger γ_{jg} is, the prepositional phrase p_g is more suitable for describing this RHIS). After the parameters γ_{jg} are trained, (18) can identify the most suitable HOI–scene pair and the corresponding prepositional phrase, thus significantly help confirm the nonterminal symbols *VP*, *PP*, and the preposition *IN* in the caption sentence. Fig. 5 illustrates the example of the proposed lexicalized tree-based sentence generation process.

Algorithm 1 Inference of the Best Configuration Between R and P

Input:

The sets of RHIS R and the prepositional phrases P

Initialization: two sets $W = \{w_j = 0, j = 1 \cdots m\}$, $C = \{c_j = \emptyset, j = 1 \cdots m\}$;

Algorithm:

for each $r_j \in R$ **do**
  for each $p_g \in P$ **do**
    if $(w_j < (1_{(R=r_j)} \cdot 1_{(P=p_g)} \cdot \gamma_{jg}))$ **then**
      $w_j \Leftarrow (1_{(R=r_j)} \cdot 1_{(P=p_g)} \cdot \gamma_{jg})$
      $c_j \Leftarrow (r_j, p_g))$
    end if
  end for
end for
return C

B. Model Inference

Given the RHIS set R and the prepositional phrase set P for a real image, we compute the score function

$$(p^*, r^*) = \arg\max_{p_g, r_j} \Sigma_g^P \Sigma_j^R \cdot 1_{(R=r_j)} \cdot 1_{(P=p_g)} \cdot \gamma_{jg}. \quad (19)$$

The aim of inference is to find the best configuration between r_j and p_g. To this end, the straightforward solution is to search for all possible combinations between them. But there is a computational bottleneck: the inner maximization over two sets of R and P. This computational bottleneck is intractable for general pairwise potentials. In this article, we propose an efficient search algorithm to accelerate the inference as detailed in Algorithm 1.

C. Model Learning

Suppose two-tuples of $\{r_j, p_g\}$ collection of the RHIS and the prepositional phrase be given, respectively. We aim to train a model Γ that captures the true combination between the HOI–scene pair and the prepositional phrase, based on a new collections of $\{r, p\}$ in a new image. We formulate these process as a regularized learning task, and use the cutting plane algorithm [20], [41] to implement the training process

$$\arg\min_{\gamma_{jg}, \xi_i} \sum_{\gamma_{jg}}^{m} \frac{1}{2} \|\gamma_{jg}\|^2 + \lambda \sum_{i=1}^{m} \xi_i \quad (20)$$
$$\text{s.t. } \forall k, p_g, \Phi(p_g, r_j) - \Phi(p_k, r_j) \geq L(p_g, p_k) - \xi_i.$$

Based on the constraint from (20), to a training image, the score of the true combination of p_g and r_j is expected to be higher than all other hypothesized combinations of p_k and r_j. λ is a hyperparameter and $\sum_{i=1}^{m} \xi_i$ are a set of slack variables. The loss function L measures the incorrect level between the hypothesis p_k and the true component p_g. Thus, we consider the form of the loss function is $0 - 1$

$$L(p_g, p_k) = \begin{cases} 1, & \text{if } p_g \neq p_k \cap \neg\exists k, \text{ s.t. } ol(p_g, p_k) > 50\% p_g \\ 0, & \text{otherwise} \end{cases} \quad (21)$$

where $ol(p_g, p_k) > 50\% p_g$ means that p_g and p_k are synonyms. Our loss function can correctly penalize predictions that false prediction considered as true positive, when they are not synonym.

VI. Experiments

To extensively evaluate the effectiveness of the proposed model, we conducted the experiments on three different datasets: 1) UIUC phrasal recognition dataset [42]; 2) six-class sport dataset [43], [44]; and 3) Microsoft COCO captions dataset [9], [14]. Section I-A *Data Source and Representation* in the supplementary material provides more details about the three different datasets. sFig. 1 (see supplementary material) shows the examples of the first two datasets. Our experiments focused on the evaluations of the recognition of HOI and the RHIS (on the first two datasets) and the quality of the generated sentences (on the third dataset).

A. Evaluation of the Hybrid Deep-Learning Model

In this experiment section, we compare with state-of-the-art methods: Liyue model [25], HO [36], HO-RCNN [36], DNN-ONLY [26], and Bo model [11]. Section I-B *The State-of-the-Art Methods* in the supplementary material provides more details about these methods.

1) Human–Object Interaction Recognition: One of our goals is to detect the HOI in images. In the experiment setting, following the previous works, we use the average precision (AP) evaluation metric as the major performance metric [44]. We compare with four existing methods: 1) Liyue model [25]; 2) HO [36]; 3) HO-RCNN [36]; and 4) Bo model [11].

Figs. 6 and 7 illustrate the comparison results of the four comparative methods and our model on two different datasets, respectively. In Fig. 6, the types of HOI and the abbreviation are person riding horse (PRH), person next to horse (PNTH), person riding bicycle (PRB), person next to bicycle (PNTB), person next to car (PNTC), person waiting for bus (PWFB), person riding motorbike (PRM), and person next to motorbike (PNTM). The evaluations of these five methods are based on best possible performances using the first two datasets.

It can be seen that our model achieves the best performance on the UIUC phrasal recognition dataset. Our model, Liyue model, HO-RCNN, HO, and Bo model perform overall AP of 76.8%, 68.3%, 71.4%, 62.5%, and 73.5%, respectively, which implies that the proposed model outperforms the Liyue model, HO-RCNN, and HO by 8.5%, 5.4%, 14.3%, and 3.3%, respectively. This justifies that learning relational features of the human–object pair can significantly benefit the HOI recognition. Comparing with these methods without or with a limited human–object-relational features, the proposed FTWIM enables our model to learn the relational features between different image regions at a great length. Furthermore, even comparing with the Liyue model and Bo model where the ground-truth visual pattern labels are used, our model still achieves more than 5% average improvement. Indeed, on the one hand, the proposal FTWIM can learn the high-level relational features of the human–object pair to significantly improve the HOI recognition; on the other hand, the special structure of the FTWIM, that is, the factored three-way interaction mechanism, allows any prior information to be used

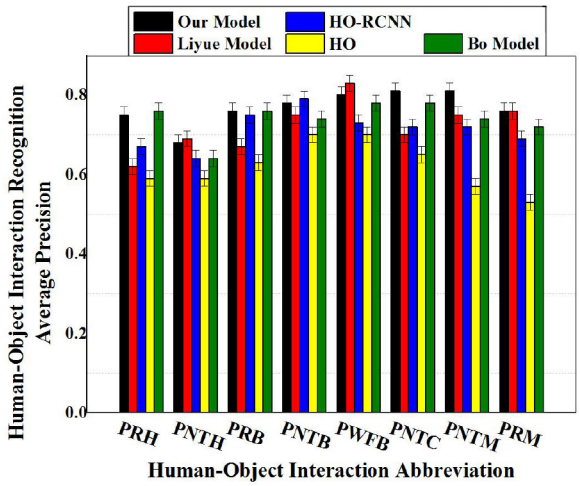

Fig. 6. HOI recognition results on the UIUC phrasal recognition dataset. The *x*-axis labels are the abbreviation of eight classes of HOIs. The *y*-axis label is HOI recognition AP.

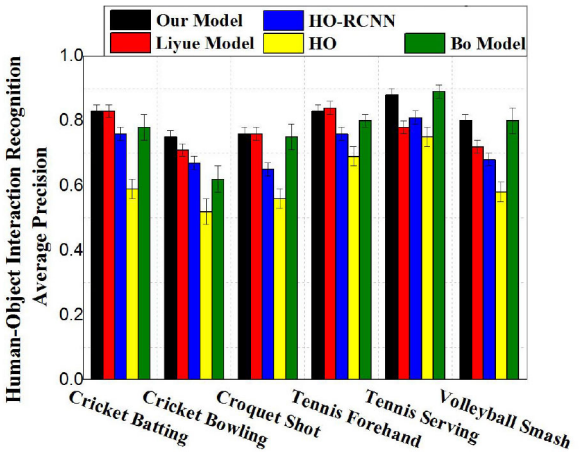

Fig. 7. HOI recognition results on the SCS dataset. The *x*-axis labels are six classes of sport activities. The *y*-axis label is sport recognition AP.

to directly facilitate the relational features to be learned in a unified single network architecture, so that the extracted features are very effective for the goal of recognition task.

On the SCS dataset, the performances of our model, Liyue model, and Bo model are very close in some specific recognition tasks (e.g., tennis forehand). The main reason could be that the scenes of the sports images are very simple and concision where the negative impact of the clutter scene is thus decreased. However, the overall AP result of our model is still better than other models. Moreover, comparing with the Bo model where the similar spatial context is used, the special structure of our model ensures the hidden layers could learn the global relational features of a human–object pair while alleviating the negative impact of the clutter scene in the hierarchy. On the other hand, our model combines the feature extraction and recognition stages in a unified hierarchical architecture so that the loss of useful information is as little as possible.

2) Evaluation of the FTWIM: In order to further evaluate the performance of the proposed FTWIM, we use the

TABLE I
RECALL@K SCORES FOR HOI ACROSS TWO DATASETS. WE COMPARE OUR METHOD RESULTS WITH THE LIYUE MODEL, BO MODEL, HO-RCNN, HO, AND DNN-ONLY

Method	UIUC Dataset		SCS Dataset	
	Recall@50	Recall@100	Recall@50	Recall@100
Liyue Model	67.63	69.88	77.22	79.71
Bo-Model	68.50	69.50	78.33	81.50
HO-RCNN	64.50	66.25	75.23	78.60
HO	53.25	56.25	64.33	68.15
DNN-ONLY	59.25	61.75	67.50	69.75
Our Model	**73.25**	**75.50**	**80.33**	**83.25**

Recall @K, the fraction of ground-truth instances that are correctly recalled in top K predictions, as the supplementary performance metric. In the HOI evaluation, we reported Recall@50 and Recall@100. We set the true positive bounding box if it overlaps with the ground-truth bounding box with an intersection over union greater than 0.5. We compared with five state of the arts: 1) Liyue model; 2) HO; 3) HO-RCNN; 4) DNN-ONLY; and 5) Bo model.

Table 1 illustrates the results of Recall@50 and Recall@100 across the UIUC and SCS datasets. It can be seen that our model achieves the best performance across two evaluation criteria. Comparing with the HO-RCNN model and DNN-ONLY, our model outperforms the Recall@50 and Recall@100 by over 10% and 15% on both datasets. The core modules of the HO-RCNN model and DNN-ONLY are RCNN to extract the invariant features of the input image. Such deep-learning structure is also an important part of our model, but we differ from them is that the proposed FTWIM is used to combine the human features, object features, as well as the spatial context features into a unified deep network architecture in order to learn the relational features of the human–object pair. With such a unique property, our model outperforms these deep-learning methods.

Moreover, the proposed model outperforms the state-of-the-art models, Liyue model and Bo model, as well, by a considerable margin in both evaluation metrics. It is worth noting that both Liyue model and Bo model use region detectors to facilitate the HOI recognition based on the ground-truth visual pattern class labels. Especially, Bo model's deep network structure is similar to ours, but does not adopt the FTWIM. Hence, the proposed FTWIM not only achieves a better recognition performance but also demonstrates that the factored three-way interaction structure of the FTWIM enables the relational feature learning of the human–object pattern to be significantly improved by the context information of the human–object pair, while such a context information is very difficult to be extracted by the traditional deep-learning networks.

Furthermore, the special structure of FTWIM unifies the feature extraction and recognition stages in a hierarchical stage, so that the model parameters can be jointly optimized for the target of the HOI recognition. Our experimental results show that each type of HOI always falls in a very small number of spatial layouts. This means that the spatial layout between humans and objects is more robust across various HOI classes and clutter scenes as shown in Fig. 8. Comparing with the

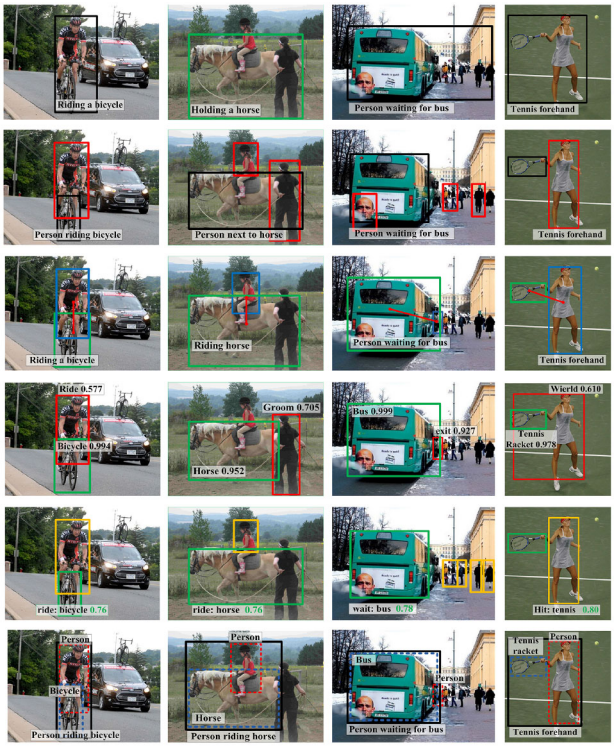

Fig. 8. Rows from 1 to 6 illustrate the recognition results of HO, DNN-ONLY, HO-RCNN, Liyue model, Bo model, and our model, respectively.

above existing methods, only our FTWIM can directly embed the image context into the relational feature extraction and recognition stages. It is no wonder that the proposed model achieves better performance.

3) Relationship Between HOI and Scene: We further evaluate the performance of our model in estimating the relationship between HOI and scene using the Recall@50 and Recall@100 as the major performance metrics. The reason for using recall instead of precision is that the relationship between the HOI and the scene is incomplete, where some true relationship is unlabeled. We compare our model with four state-of-the-art relationship prediction methods: 1) Dai model [6]; 2) GAN model [10]; 3) Li model [45]; and 4) DNN-ONLY. In the experiment, all the training and test images are manually annotated with the ground-truth label of RHISs. For example, *riding bicycle on the road, playing tennis on a tennis court*, etc. We use Faster-RCNN [34] to detect the scene stuffs of every image on the UIUC phrasal recognition dataset and the SCS dataset, respectively.

The evaluation results are summarized in Table II. On both two datasets, it can be seen that our model outperforms the existing models considerably. The GAN model and DNN-ONLY perform poorer than the other models, as it is hard for the traditional deep network to learn the high-level invariant features for both object detection and the relationship prediction. Comparing with the Li model and the Dai model using a limited image context, our model explores very detailed image context information between image regions by using the FTWIM, thus detects the relationship between HOI and scene effectively. Furthermore, our FTWIM unifies the

TABLE II
RECALL@K SCORES FOR MODELING THE RELATIONSHIP BETWEEN AN HOI AND A SCENE ON TWO DATASETS. WE COMPARE OUR METHOD RESULTS WITH THE DAI MODEL, GAN MODEL, LI MODEL, AND DNN-ONLY

Method	UIUC Dataset		SCS Dataset	
	Recall@50	Recall@100	Recall@50	Recall@100
Dai Model	25.85	27.10	41.95	43.63
GAN Model	23.73	24.88	33.22	34.71
Li Model	25.22	26.50	36.72	39.22
DNN-ONLY	18.25	22.50	29.50	31.25
Our Model	**27.72**	**29.28**	**42.69**	**44.28**

Fig. 9. Examples of the recognition of the relationship between HOI and scene. Rows from 1 to 5 describe the results of the Dai model, DNN-ONLY, GAN model, Li model, and our model, respectively.

feature extraction and recognition stage under a deep network architecture, so that the model parameters can be effectively optimized with respect to the target of the recognition task. The experiments demonstrate that the proposed FTWIM can significantly improve the advanced image recognition task by jointly learning the relational features of a variety of image context.

Fig. 9 shows the visualized comparisons on modeling RHIS examples. It can be seen that our model outperforms these baseline models. For example, on the riding bicycle image, PRB on the road, which is next to the grass, only our model can precisely describe the relationships between the HOI and the road. Instead of taking spatial context analysis as bias, our model treats 3-D spatial layout as a special input that directly guides the high-level relational features extraction. The performance demonstrates that this special structure of our model could fully exploit the advantages of the spatial context for the prediction of the HOI and the RHIS.

B. Evaluation of the Image Captioning Model

In this section, we evaluate the effectiveness of our model in image caption generation using two metrics: 1) the automatic

TABLE III
BLEU-N EVALUATION RESULTS OF THE SENTENCE GENERATION TASK ON THE MICROSOFT COCO CAPTIONS DATASET. HIGHER SCORE MEANS BETTER PERFORMANCE. BLEU-N REPRESENTS THE N-GRAM BLEU SCORE

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Our model	69.6 ± 1.0	68.3 ± 0.5	49.5 ± 0.84	36.9 ± 1.0
GAN model	55.1 ± 1.5	36.3 ± 0.5	28.6 ± 0.3	21.1 ± 0.1
WC model	46.7 ± 2.5	31.5 ± 1.5	21.5 ± 0.8	18.9 ± 0.3
LYBP model	58.6 ± 1.0	39.3 ± 0.7	31.8 ± 0.6	24.6 ± 0.4
KJKL model	49.3 ± 1.0	32.5 ± 0.9	26.3 ± 0.7	22.5 ± 0.5
JD model	69.0 ± 1.0	38.5 ± 0.5	33.5 ± 0.84	36.5 ± 1.0

TABLE IV
HUMAN EVALUATION FOR OUR MODEL AND OTHER FIVE METHODS ACROSS FOUR EVALUATION CRITERIA

Model	Grammar	Cognition	Action	Scene
Human descriptions	4.9	4.9	4.9	4.9
GAN model	4.0	3.7	3.2	3.4
WC model	3.8	3.5	2.8	3.0
LYBP model	4.1	3.8	3.1	3.3
KJKL model	4.2	3.6	3.0	3.1
JD model	4.2	3.8	3.6	3.3
Our model	4.2	4.2	4.4	4.1

evaluation and 2) human evaluation. For both of evaluation metrics, we compare our model with five state-of-the-art methods: 1) GAN model [10]; 2) WC model [9]; 3) LYBP model [46]; 4) KJKL model [38]; and 5) JD model [32]. Section I-B *The State-of-the-Art Methods* in the supplementary material provides more details about these methods.

1) Automatic Evaluation: For the automatic evaluation, we use the BLEU score to estimate how similar the generated sentences and the humanity-standard descriptions is. The BLEU score is originally proposed for measuring the performance of automatic machine translation. Now, BLEU is the standard evaluation matric for image description generation [9], [39]. Following previous studies, the evaluation metric of BLEU scores includes four different criteria: BLEU-1, BLEU-2, BLEU-3, and BLEU-4. For our model and five comparative methods, every generated descriptive sentence is compared against the same manually generated sentence. The performances of our model and five other methods are shown in Table III where a higher score represents better performance.

It can be seen that our model achieves the best scores among all four BLEU-n criteria. Especially, in terms of the BLEU-2 and BLEU-3 indexes, our model scores more than twice as much as the other methods do. In particular, the BLEU-2 and BLEU-3 indexes represent the two-character words and the trisyllabic words, which cover most of the subject–verb phrase, subject–verb–object phrase, and the adverbial phrase. This demonstrates that the proposed image caption generator can capture the most important parts of an image accurately (e.g., subject, verb, object, and adverbial), and order these parts in syntactically well-formed sentence precisely. Indeed, comparing with the other methods with or without limited verb prediction, the tree composition strategy used by our model takes the HOI as the anchors to present the subject, verb, and object accurately, which other models cannot generate. Such a strategy significantly improves the efficiency of our model in object ordering and the relationship description between subject and object. Moreover, our model explores the adverbial structure and the corresponding prepositional phrase in detail by jointly optimizing the RHIS, the prepositional phrase, and the tree composition rules. This guarantees that the major relationship between image parts (e.g., subject, object, and adverbial) can be well described within syntactically and semantically well-formed descriptions.

In other BLEU-n scores, our model also outperforms existing methods by a large margin. Different from the existing methods, our model explicitly captures the co-occurrence compatibility between activity and the RHIS, as well as the corresponding linguistic expression by the proposed lexicalized-tree growing strategy. Moreover, our model combines the lexicalized-tree growing strategy and the natural language rules into sequence inference. These advanced inferences enable our model to create more semantic, flexible, and creative image descriptive sentences than those methods which used probability estimates of object co-occurrence [9], [10], [46].

2) Human Evaluation: The above automatic evaluation directly measures the accuracy and grammatical correctness of generated sentences. Furthermore, in this section, we conduct a human judgment on the quality of the generated sentences. Following previous studies [10], [38], we invite ten students as raters to estimate the quality of the generated descriptive sentences produced by our model and these four methods. The performance evaluation metrics include four criteria on a scale from 1 to 5.

1) *Grammar:* Give high scores if the generated sentence does not include obviously grammatical errors.
2) *Cognition:* Give high scores if the generated sentence is rational, and can depict the major caption of the query image.
3) *Action:* Give high scores if the generated sentence correctly describes the interaction of a human–object pair.
4) *Scene:* Give high scores if the generated sentence correctly describes the scene of the query image.

Then, we randomly select 100 images from the Microsoft COCO dataset, and obtain five descriptive sentences for each test image from these six methods (including our model and the five other methods) for human judgment experiments. Table IV illustrates the mean scores of each criterion for our model and the other five methods. The manual standard descriptions elicited judgments around five, which are significantly better than the model-generated sentences on all aspects. All methods produce highly grammar performance, with mean ratings of between 3.8 and 4.2. There are no big differences between our model and five other methods in terms of *Grammar* criterion. This result can be explained by the fact that all of the methods rely on natural language grammar rules. In terms of the criteria—*Cognition, Action*, and *Scene*, it can be seen that our model significantly outperforms the five existing methods.

In terms of the *Action* criterion, our model achieves a remarkable score 4.4, which shows the strength of our model in the description of the SVO phrase and the adverbial structure. The other five methods only earn scores from 2.8 to 3.5,

Fig. 10. Some examples of generated descriptions produced by our model and five other methods, respectively.

even though the semantic relation predictions are used in the Bo model and the LYBP model. This might benefit from that our model has the ability to precisely recognize the interaction relationship between subjects, while the other methods just learn the phrase of $<subject, predicate, object>$. Comparing with the approaches without adverbial generating, we design a special statistical algorithm that explores the best prepositional phrase to depict the semantic relationship between SVO and scene in the lexicalized-tree growing process. These properties of the proposed image caption generator guarantee the generated sentences can detail the action attribute of the query image.

In terms of the *Cognition* criterion, it can be seen that the five existing methods generate similarity scores, with mean ratings from 3.5 to 3.8. The reason might be that all these methods only have limited relation estimation between regions lacking the ability to predict the relation between SVO and scene. In contrast, our syntactic-tree-based model focuses on the recognition of the variety of semantic relationship between image parts, and optimizing the compatibility between the phrase structure, HOI, and RHIS. Therefore, our model achieves the mean score 4.2 by an average improvement from 0.7 to 0.4. For example, in the second picture of Fig. 10, our model precisely describes that the adverbial structure of the scene is *in a grass field*. Thus, the proposed model significantly improves the cognition of the generated sentence.

In terms of the *Scene* criterion, it can be seen that our model significantly outperforms the five existing methods, even showing 0.7 average improvement over the GAN model and the LYBP model where the ground-truth scene labels are used. It is noted that the proposed lexicalized-tree growing strategy focuses on learning the relationship between SVO and scene, then the relationship is further processed through a special statistical algorithm to explore the best compatibility between the prepositional phrase and the scene. These properties of the proposed model enable our model to generate a syntactically and semantically well-detailed description and outperform the-state-of-the-art methods across different human assessments. From the examples of the generated sentence in Fig. 10, we also learn that for the action recognition and the relationship between an HOI and a scene, the visual phrases always fall in a specific small number of spatial layouts. This situation indicates that our model always capture the most suitable conjunctions for the descriptive sentence generation.

VII. Conclusion

In this article, we have proposed a hybrid deep-learning model and an image caption generator in a deep network architecture to generate image descriptive sentences. The former consists of the proposed FTWIMs and the cascade of deep convolutional networks to learn the high-level relational features of the HOI and the RHIS, given the image spatial context, in a hierarchy of stages by progressively integrating various features from lower levels. The experiments demonstrate that the high-level relational features dramatically improve the relationship recognition. The latter treats the image caption generation as a syntactic-tree generation process. It takes the HOI and RHIS as anchors to improve the tree-growth process in generating syntactically well-formed descriptions. The extensive experimental results demonstrate that the proposed model outperforms the existing methods in the prediction of the HOI and the RHIS. Moreover, the proposed description generation model not only discover the most suitable linguistics phrase to describe the semantic relationship between the image parts but also generate semantically image descriptive sentences.

References

[1] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, "Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2321–2334, Dec. 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2642953

[2] L. Bai, K. Li, J. Pei, and S. Jiang, "Main objects interaction activity recognition in real images," *Neural Comput. Appl.*, vol. 27, no. 2, pp. 335–348, 2016.

[3] L. Bai and Q. Chen, "Visual phrase recognition by modeling 3D spatial context of multiple objects," *Neurocomputing*, vol. 253, pp. 183–192, Aug. 2017.

[4] G. Kulkarni *et al.*, "BabyTalk: Understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013. [Online]. Available: https://doi.org/10.1109/TPAMI.2012.162

[5] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi, "TreeTalk: Composition and compression of trees for image descriptions," *Trans. Assoc. Comput. Linguist.*, vol. 2, no. 10, pp. 351–362, 2014. [Online]. Available: https://www.aclweb.org/anthology/Q14-1028

[6] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3076–3086.

[7] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," 2014. [Online]. Available: arXiv:1411.4555.

[8] 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.

[9] Q. Wang and A. B. Chan, "CNN+ CNN: Convolutional decoders for image captioning," 2018. [Online]. Available: arXiv:1805.09019.

[10] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional GAN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2970–2979.

[11] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "Pose-aware multi-level feature network for human object interaction detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 9469–9478.

[12] X. Li, A. Yuan, and X. Lu, "Vision-to-language tasks based on attributes and attention mechanism," *IEEE Trans. Cybern.*, early access, May 17, 2019, doi: 10.1109/TCYB.2019.2914351.

[13] F. Wu, J. Cheng, X. Wang, L. Wang, and D. Tao, "Image hallucination from attribute pairs," *IEEE Trans. Cybern.*, early access, Apr. 7, 2020, [Online]. Available: https://doi.org/10.1109/TCYB.2020.2979258

[14] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," in *Proc. Int. Conf. Learn. Represent.*, San Diego, CA, USA, May 2015.

[15] Y. Bin, Y. Yang, F. Shen, N. Xie, H. T. Shen, and X. Li, "Describing video with attention-based bidirectional LSTM," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2631–2641, Jul. 2019.

[16] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele, "Translating video content to natural language descriptions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2013, pp. 433–440.

[17] 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.

[18] J. Deng, X. Xie, and S. Zhou, "Conversational interaction recognition based on bodily and facial movement," in *Image Analysis and Recognition*, A. Campilho and M. Kamel, Eds. Cham, Switzerland: Springer Int., 2014, pp. 237–245.

[19] T. Zhou, W. Wang, S. Qi, H. Ling, and J. Shen, "Cascaded human-object interaction recognition," 2020. [Online]. Available: arXiv:2003.04262.

[20] L. Bai and K. Li, "Predicting image caption by a unified hierarchical model," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2015, pp. 1–6.

[21] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli, "Learning to detect human-object interactions with knowledge," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 468–476.

[22] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," 2017. [Online]. Available: arXiv:1704.07333.

[23] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 401–417.

[24] Y. Li, W. Ouyang, X. Wang, and X. Tang, "VIP-CNN: Visual phrase guided convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1347–1356.

[25] L. Shen, S. Yeung, J. Hoffman, G. Mori, and L. Fei-Fei, "Scaling human-object interaction recognition through zero-shot learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2018, pp. 1568–1576.

[26] C. Baldassano, D. M. Beck, and L. Fei-Fei, "Human–object interactions are more than the sum of their parts," *Cerebr. Cortex*, vol. 27, no. 3, pp. 2276–2288, 2017.

[27] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4D human-object interactions for event and object recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2013, pp. 3272–3279.

[28] Z. Li, J. Sedlar, J. Carpentier, I. Laptev, N. Mansard, and J. Sivic, "Estimating 3D motion and forces of person-object interactions from monocular video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8640–8649.

[29] W. Choi, Y. W. Chao, C. Pantofaru, and S. Savarese, "Indoor scene understanding with geometric and semantic contexts," *Int. J. Comput. Vis.*, vol. 112, no. 2, pp. 204–220, 2014.

[30] Q. Jin and J. Liang, "Video description generation using audio and visual cues," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2016, pp. 239–242.

[31] X. Xiao, L. Wang, K. Ding, S. Xiang, and C. Pan, "Dense semantic embedding network for image captioning," *Pattern Recognit.*, vol. 90, pp. 285–296, Jun. 2019.

[32] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 684–699.

[33] L. Bai and L. Yang, "A unified deep learning model for protein structure prediction," in *Proc. 3rd IEEE Int. Conf. Cybern. (CYBCONF)*, 2017, pp. 1–6.

[34] X. Sun, P. Wu, and S. C. Hoi, "Face detection using deep learning: An improved faster RCNN approach," *Neurocomputing*, vol. 299, pp. 42–50, Jul. 2018.

[35] H. Xu, K. Li, F. Lv, and J. Pei, "3D depth perception from single monocular images," in *MultiMedia Modeling*. Heidelberg, Germany: Springer, 2015, pp. 510–521.

[36] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2018, pp. 381–389.

[37] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[38] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 317–325.

[39] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 706–715.

[40] M. Mitchell *et al.*, "MIDGE: Generating image descriptions from computer vision detections," in *Proc. 13th Conf. Eur. Assoc. Comput. Linguist. Assoc. Comput. Linguist.*, 2012, pp. 747–756.

[41] K. Li and L. Bai, "Generating image description by modeling spatial context of an image," in *Proc. Int. Joint Conf. Neural Netw.*, 2015, pp. 1–8.

[42] M. A. Sadeghi and A. Farhadi, "Recgnition using visual phrases," in *Proc. CVPR*, 2011, pp. 1745–1752.

[43] A. Gupta and P. Mannem, "From image annotation to image description," in *Neural Information Processing*. Heidelberg, Germany: Springer, 2012, pp. 196–204.

[44] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2010, pp. 17–24.

[45] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1261–1270.

[46] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7219–7228.