Understanding Objects in Video: Object-Oriented Video Captioning via Structured Trajectory and Adversarial Learning

FANGYI ZHU\textsuperscript{1}, (Member, IEEE), JENQ-NENG HWANG\textsuperscript{2}, (Fellow, IEEE), ZHANYU MA\textsuperscript{1}, (Senior Member, IEEE), GUANG CHEN\textsuperscript{1}, AND JUN GUO\textsuperscript{1}

\textsuperscript{1}Beijing University of Posts and Telecommunications, Beijing 100876, China
\textsuperscript{2}University of Washington, Seattle, WA 98105, USA

Corresponding author: Jun Guo (guojun@bupt.edu.cn)

Abstract. Traditional video captioning requests a holistic description of the video, yet the detailed descriptions of the specific objects may not be available. Besides, most methods adopt frame-level intertwined features among objects and ambiguous descriptions during training, which is difficult for learning the vision-language relationships. Without identifying the classes and locations of separate objects, neither associating the transition trajectories among the objects, these data-driven image-based video captioning methods cannot reason the activities with visual features only, not to mention performing well under small-samples. We propose a novel task, namely the object-oriented video captioning, which focuses on understanding the videos in object-level. We re-annotate the object-oriented video captioning dataset (Object-Oriented Captions) with object-sentence pairs to facilitate more effective cross-modal learning. Thereafter, we design the video-based structured trajectory network via adversarial learning (STraNet) to effectively analyze the activities along the time domain and proactively capture the vision-language connections under small datasets. The proposed STraNet consists of four components: the structured trajectory representation, the attribute explorer, the attribute-enhanced caption generator, and the adversarial discriminator. The high-level structured trajectory representation provides useful supplement over previous image-based approaches, allowing to reason the activities from the temporal evolution of visual features and the dynamic movement of spatial locations. The attribute explorer helps to capture discriminative features among different objects, with which the subsequent caption generator can generate more informative and accurate descriptions. Finally, by adding an adversarial discriminator on the caption generation task, we can improve learning the inter-relationships between the visual contents and the corresponding visual words. To demonstrate the effectiveness, we evaluate the proposed method on the new dataset and compare it with the state-of-the-arts for video captioning. From the experimental results, the STraNet exhibits the ability of precisely describing the concurrent objects and their activities in details.

Index Terms. Video understanding, video captioning, object-level analysis, structured trajectory, adversarial learning.

I. INTRODUCTION

With the rapid growth of videos on the Internet, it becomes more important to automatically understand the videos. Video captioning, which calls for a systematic description of the videos, poses an intriguing challenge to learn the connections between vision and natural language [1]–[8], [10], [12]–[18], [28]–[34].

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenhua Guo\textsuperscript{169146}.

With the methods of image captioning gradually getting matured [8]–[11], more and more attentions are shifted to video captioning [12]–[18]. Video captioning has significance to human-robot interaction and visually impaired people [6], [7].

Previously, video captioning offers a holistic description of the entire video with less detailed information associated with objects, and this kind of video captioning cannot provide information of concurrent objects and activities as well. Actually, while human watching videos, instead of focusing on the
under the situation of multi-objects; the other is that it cannot bring two problems: one is it cannot extract effective features using the pre-trained models for action recognition to extract the objects have individual or interactive activities. Hence, most videos have multiple concurrent objects, and all usually only a single object which has activities in the videos datasets for video captioning, the main objects sometimes definition only consist of actions of human. However, in the [16], [18], [23]. Nonetheless, most datasets for action recognition [20]–[22] to extract the spatio-temporal besides the visual features, most works adopt the models for quite limited.

Therefore, the applicability of these data-driven methods are able sufficient training data in the many application scenarios. In reality, there are not always available large amounts of labeled data to learn the vision-language relationships. We construct a new dataset, Object-oriented Captions, with consistent object-sentence pairs. All available datasets for classical video captioning just have contents and the corresponding visual words.

The activities from the temporal evolution of visual features, allowing us to achieve real video understanding. Second, to capture more discriminative features among different objects and explore precise attributes, we further design the attribute explorer and attribute-enhanced caption generator. Finally, we modify the adversarial learning scheme on top of the caption generation task to help bridging the visual contents and the corresponding visual words.

We shift the holistic frame-level video captioning to object-level video captioning. Instead of a coarse and holistic description of the entire video or a sporadic object, we aim at understanding the video in object-level, which is closer to human experience while watching videos. Understanding the activities in object-level leads to a greater understanding of the video contents, which deserves more of our attentions.

A. OBJECT-ORIENTED VIDEO CAPTIONING

We further design an object-level structured trajectory network (STraNet) via adversarial learning, which can achieve promising performance under small-samples and adapt to more application scenarios. The main contributions of our work are three-fold:

A. OBJECT-ORIENTED VIDEO CAPTIONING

We propose the object-oriented video captioning aimed at understanding the videos in object-level, which can describe concurrent objects and activities with more details.

B. VIDEO-BASED STRUCTURED TRAJECTORY NETWORK VIA ADVERSARIAL LEARNING

We design the object-oriented structured trajectory network (STraNet) via adversarial learning to replace previous data-driven image-based methods. Since the actions are continuous in the time domain, we focus more on understanding the rules and the temporal evolution in the trajectories to avoid the need of training under large amounts of data. We first design the structured trajectory representation, with which we can achieve the aim of reasoning activities without supplementary cues from other tasks. The structured trajectory representation improves the model ability of understanding the activities from the temporal evolution of visual features, allowing us to achieve real video understanding. Second, to capture more discriminative features among different objects and explore precise attributes, we further design the attribute explorer and attribute-enhanced caption generator. Finally, we modify the adversarial learning scheme on top of the caption generation task to help bridging the visual contents and the corresponding visual words.

B. VIDEO-BASED STRUCTURED TRAJECTORY NETWORK VIA ADVERSARIAL LEARNING

We design the object-oriented structured trajectory network (STraNet) via adversarial learning to replace previous data-driven image-based methods. Since the actions are continuous in the time domain, we focus more on understanding the rules and the temporal evolution in the trajectories to avoid the need of training under large amounts of data. We first design the structured trajectory representation, with which we can achieve the aim of reasoning activities without supplementary cues from other tasks. The structured trajectory representation improves the model ability of understanding the activities from the temporal evolution of visual features, allowing us to achieve real video understanding. Second, to capture more discriminative features among different objects and explore precise attributes, we further design the attribute explorer and attribute-enhanced caption generator. Finally, we modify the adversarial learning scheme on top of the caption generation task to help bridging the visual contents and the corresponding visual words.

C. OBJECT-ORIENTED VIDEO CAPTIONING DATASET

We perform experiments on the new dataset and compare with the state-of-the-arts for video captioning, capture the action information of the objects which are not human. Besides, the C3D features and the visual features are usually independent, which means that these two kinds of features fed into a recurrent neural network (RNN) at the same time step always represent the information of different frames, resulting in confusion in the training of neural networks.

To deepen the research, we propose a novel task, named object-oriented video captioning, to transform the frame-level video captioning to object-level video captioning. We further design an object-level structured trajectory network (STraNet) via adversarial learning, which can achieve promising performance under small-samples and adapt to more application scenarios. The main contributions of our work are three-fold:

A. OBJECT-ORIENTED VIDEO CAPTIONING

We propose the object-oriented video captioning aimed at understanding the videos in object-level, which can describe concurrent objects and activities with more details.

B. VIDEO-BASED STRUCTURED TRAJECTORY NETWORK VIA ADVERSARIAL LEARNING

We design the object-oriented structured trajectory network (STraNet) via adversarial learning to replace previous data-driven image-based methods. Since the actions are continuous in the time domain, we focus more on understanding the rules and the temporal evolution in the trajectories to avoid the need of training under large amounts of data. We first design the structured trajectory representation, with which we can achieve the aim of reasoning activities without supplementary cues from other tasks. The structured trajectory representation improves the model ability of understanding the activities from the temporal evolution of visual features, allowing us to achieve real video understanding. Second, to capture more discriminative features among different objects and explore precise attributes, we further design the attribute explorer and attribute-enhanced caption generator. Finally, we modify the adversarial learning scheme on top of the caption generation task to help bridging the visual contents and the corresponding visual words.

C. OBJECT-ORIENTED VIDEO CAPTIONING DATASET

We perform experiments on the new dataset and compare with the state-of-the-arts for video captioning.
The experimental results demonstrate that our proposed approach can achieve the superior performance in terms of BLEU@4, METEOR, CIDEr and ROUGE-L metrics [24]–[27]. More importantly, our proposed method can understand most concurrent activities with more details rather than giving a holistic description.

The paper is organized as follows. Section 2 provides overviews of recent related works. Section 3 describes our STraNet in details. In Section 4, we first introduce the re-annotated dataset. Then, report the experimental results and give the ablation study. Finally, Section 5 presents the conclusions and future works.

II. RELATED WORKS

In this section, we review related works in the context of video captioning, as well as adversarial learning.

A. VIDEO CAPTIONING

Video captioning, which bridges two modalities: vision and language, poses a great challenge for artificial intelligence. Recently, a large number of methods have been proposed for video captioning, where the encoder-decoder architectures have been widely adopted [6]–[8], [10], [12]–[18], [28]–[34]. The decoder learns a high-level representation of the video, then the decoder generates the descriptions word-by-word for the decoder.

B. ADVERSARIAL LEARNING

Generative adversarial learning (GAN) is proposed as an scheme for improving the performance of generative models in [41]. In a GAN, two models are simultaneously trained, i.e., a generative model \( G \) to capture the data distribution and a discriminative model \( D \) for distinguishing that a sample is from the training data or \( G \). The training procedure for \( G \) is to maximize the probability of \( D \) making a mistake and it is exactly similarly to a minimax two-player game. To further control the mode of the generated data, conditional generative adversarial nets (CGANs) are proposed in [42]. After adding the conditional labels in a CGAN, the scheme has been applied to more tasks, e.g., image classification and multi-modal learning. Henceforth, the adversarial learning draws significant attentions and applications on image-to-image translation [43]–[47], text-to-image synthesis [48], domain adaptation [49], medical image generation [50], and so on.

Yang et al. [33] propose LSTM-GAN, which introduces the adversarial learning into video captioning. LSTM-GAN directly adopts the general framework of adversarial learning to distinguish whether the generated sentences are real or fake, without modifying the discriminator based on the caption generation task. Whereas, the grammar of the generated caption is much more difficult to distinguish because of the diversity of natural language. Moreover, the generated natural language is in discrete space, different from the image pixel in continuous space, resulting in difficulties during training the discriminator.

Overall, previous video captioning methods fails to provide sufficient and detailed information for object-level analyses. Although significant improvements have been achieved, most methods are data-driven image-based, which cannot reason the step-by-step object-level activities along time.

III. ARCHITECTURE

As illustrated in Fig. 2, our STraNet consists of four modules: (1) Structured trajectory representing: For each input video, we first obtain a set of object trajectories, then extract...
the global features and local features respectively for each object at every frame where the object occurs. Next, to present each object, we construct a high-level structured trajectory representation along the time domain. (2) **Attribute exploring:** The local features from module (1) are fed to the attribute explorer, where the attribute scores of the objects are utilized for selecting more discriminative features of different objects automatically during the backpropagation. (3) **Attribute-enhanced caption generating:** Based on attribute scores from module (2), we generate the attribute-enhanced masks for reducing the vocabulary space dynamically. Next, with the object-oriented structured trajectory representation from module (1), the final word-by-word descriptions for the objects are generated. (4) **Adversarial learning:** We further design an adversarial discriminator to help in deciding if the generated descriptions are accurate for the given visual contents. In the following subsections, we will address these four modules in details.

**A. STRUCTURED TRAJECTORY REPRESENTING**

In this subsection, we present how to build the object-oriented structured trajectory representation. With the advances of object-level visual analyses, many excellent works have emerged, such as Faster R-CNN, YOLO, Mask R-CNN for object detection and Tracktor, TrackletNet Tracker(TNT) for multi-object tracking [51]–[55]. Recently, Yang et al. propose MaskTrackRCNN, built upon Mask R-CNN, to perform object detection, instance segmentation and object tracking at the same time. MaskTrackRCNN is trained on natural videos from YouTube, which are similar to our training data, whereas most other tracking methods are mainly trained on scenes for video surveillance or autonomous driving. Therefore, adopting the MaskTrackRCNN as a combined detector-tracker in our work has two salient advantages: (1) MaskTrackRCNN can solve the above mentioned three tasks at the same time, we do not need to do each task separately; (2) MaskTrackRCNN is trained on the videos which have more similar scenes to our scenarios. This module consists of three steps: first, the detection and tracking results are obtained by the MaskTrackRCNN. Second, with the detection and tracking results, the local features and global features of corresponding objects at different frames are extracted. Third, the high-level representation via the structured trajectory for each corresponding object in the video is built. In the following, we will introduce the three steps successively.

First, given a $T$-frame video $V = \{v^1, \ldots, v^T\}$, the MaskTrackRCNN is adopted to obtain the trajectories for all objects as

$$O = F_{MTRCNN}(V),$$

$$o = \{v^o, b^o\},$$

where $F_{MTRCNN}$ is the MaskTrackRCNN model. $O = o_1, \ldots, o_n$ is a set of detected objects of the video. For each object $o$, we record the time stamps $v^o$ when it occurs and the corresponding spatial locations where it exists $b^o = \{v^o_1, \ldots, v^o_m\}$ denotes a set of frames where the object $o$ occurs, $m$ is the total number of the frames where the object $o$ exists. $b^o = \{b^o_1, \ldots, b^o_m\}$, where $b^o_1 = (x^o_1, y^o_1, w^o_1, h^o_1)$ denotes the spatial location of $o$ at time $t$.

Second, for each object trajectory, we extract its local and global features as

$$\phi^o_{lo} = \Theta^o(v^o),$$

$$\phi^o_{go} = \Theta^o(v^o) \cdot c^o.$$
where $\Theta$ is a pre-trained neural network for extracting the semantic visual features. We feed each cropped object into $\Theta$ to get its local features. The features from the top layers of a neural network contain more semantic information, yet less detailed appearance information, such as color and texture. However, the color information is important for distinguishing different individuals, therefore, we further combine the color histograms into the local features. Finally, the local feature $\phi^o_t$ of $o$ at $t$ consists of two components: local visual features of objects from the neural network $\Theta(o')$ and the color histogram vectors $c^o_t$ as shown in (3).

To incorporate the interactions of a tracked object with other objects and the stuff, we again adopt $\Theta$ to extract the visual features of the frames where the object occurs as the global features $\phi^{go}_t$ in (4). Our method is quite different from other works which directly adopt frame-level features. In previous works, the activities are mainly learned based on the frame-level features which are inter-tangled features of all objects and background. However, in our work, with the foreground objects and background being separated, the activities are learned by analyzing the local features of the objects along time. The global features serve as a supplement of the interactions and background. Some works, e.g., the Fine-Grained Spatial Temporal Attention Model (FSTA) [6], also use the results from Mask R-CNN [54] to extract the foreground objects, but the background is completely ignored in their efforts. In fact, background makes up the majority of our visual surroundings, e.g., road, sky, grass, beach, building, and is useful to infer the positions and orientations of the objects, object-object interactions and object-background interactions. Therefore, background is crucial for scene understanding. Besides, some works adopt the C3D features as a supplement of local information for learning activities [12], [15], [16], [18], [23]. However, the global features and the C3D features at the same time step, which are jointly fed into an RNN, usually represent the information for different periods of the video. For this reason, the model cannot effectively discover the relationships between the global information and the C3D features. In our work, we extract the features for the frame where the object exists as global features, which are also paired with local features among detected objects, so the network can learn the inherent relationships between local and global features more effectively. To better exploit the connections between objects and the background, we combine the spatial locations of the objects $b^o_t$, with the joint local and global features $\phi^o_t = [\phi^l_t, \phi^{go}_t]$. As shown in Fig. 3, our paired local and global features are more informative for learning the interactions between different objects or between objects and background.

Based on the results obtained in the previous step, we further build the high-level structured trajectory representation $H^o_t$ for each object at time $t$. $H^o_t$ consist of the combined feature $\phi^o_t$ and the spatial location $b^o_t$. FSTA [6] and OA-BTG [19] use the detector to get effective features for objects as well, yet without combining the spatial information. Actually, the temporal evolution of spatial locations associated with the object helps a lot in understanding the activities, e.g., jump, walk, squat. Moreover, combing the spatial transition can help for learning activities under limited data as well.

By adopting the structured trajectory representation, we directly utilize the trajectory-sentence pairs for training. It is more effective and specific for learning the activities from the transition of visual features and the evolution of spatial locations. This structured trajectory representation achieves the objective of reasoning activities based on visual features only, without extra temporal cues from other tasks. In Fig. 4, we show an example of our trajectory-sentence pairs. Actually, we have a trajectory-sentence pair for each object in the video. Thus, we have more trajectory-sentence pairs in more complex videos to process all concurrent activities.

**B. ATTRIBUTE EXPLORING**

Furthermore, to provide more informative and accurate descriptions for different objects and identities, we add an attribute explorer. Actually, the objects which may have activities in most videos can be roughly grouped into three super-classes: human, animal and vehicle. Different from traditional video captioning, we also incorporate some detailed information, such as gender, which is very important in object-level descriptions. Nonetheless, this kind of detailed information cannot be obtained with a standard detector or tracker. To learn more effective and discriminative features as the detailed information of different super-classes, we thus
design the attribute explorer (see Fig. 2).

\[ \gamma_o = F_p \left( W_p \sum_{t_1}^{t_m} \phi_{10}^{\theta} / m + b_p \right). \] (5)

As shown in (5), \( F_p \) is the attribute explorer, which consists of fully connected layers, and \( \gamma_o \) is the obtained attribute score vector. \( W_p \) and \( b_p \) separately denote the parameters and bias to be learned. Each dimension of \( \gamma_o \) represents the probability of \( o \) associated with each predefined class. The categorical cross-entropy is adopted as the loss function \( L_{att} \).

Finally, the attribute scores are concatenated with the high-level structured trajectory representation as the input for the following caption generating module. Thus, for \( o \), the final input of the caption generator at \( t \) is \( \tilde{\phi}_o = [\phi_{10}^{\theta}, d_o, \gamma_o] \).

With the attribute explorer, the network can learn more discriminative features for different super-classes, and the following captioning module can thus learn better the connections between vision and natural language.

C. ATTENTION-ENHANCED CAPTION GENERATING

With the results from previous step, we can train a model with the obtained trajectory-sentence pairs shown in Fig. 4. Given an object-oriented trajectory \( H \), the caption generator is required to understand the activities and automatically generate a sentence \( s = \{s_1, s_2, \ldots, s_K\} \) word-by-word in (6), where \( \theta \) represents the parameters to be learned, and \( K \) is the length of the sentence. \( s_1, s_2, \ldots, s_{K-1} \) denote the generated partial words. During training, the parameters \( \theta^* \) are learned by maximizing the formulation in (7).

\[ P(S(H)) = \prod_{k=1}^{K} P(s_k | s_1, s_2, \ldots, s_{k-1}, H ; \theta). \] (6)

\[ \theta^* = \arg \max_{\theta} \sum_{(H,s)} \log P(S(H); \theta). \] (7)

An RNN has the capability to decode video contents to sentences. Most works adopt LSTMs or GRUs to generate descriptions [31], [38]. In our scheme, we choose the GRU, which is a good alternative for LSTM, since it has fewer parameters and is easier to converge under less data. Instead of using three gates (input gate, forget gate, output gate) as in an LSTM, a GRU only has two gates: update gate and reset gate. Different from an LSTM using the memory cells to transfer the information, a GRU directly uses the hidden states. The reset gate decides how much past information to forget, and the update gate controls what information to throw away and what information to carry over. In brief, the GRU can be updated by (8), where \( \phi_k \) and \( h_{k-1} \) denote the current input and the previous hidden state, respectively. In addition we also adopt the temporal attention mechanism to help decide which frames are the key frames for the current word generation and avoid the negative impact of incorrect tracking results (e.g., ID switches in tracking).

\[ h_k = GRU \left( \phi_k, h_{k-1} \right). \] (8)

Most previous methods end with outputting the generated words created by feeding the outputs of the GRU to the softmax layer. In our STraNet, we further process the outputs from the softmax layer with the attribute-enhanced masks to revise the generated detailed information of the objects. The words in the vocabulary which do not belong to the super-class of the current object will be masked as shown in Fig. 2. In the example, the current object is a girl. Thus, the probabilities of generating the words, which can not describe the girl, e.g., boy and man, are masked to zero.

\[ softmax(h_k) \odot M_o, \] (9)

where \( softmax(h_k) \) represents the probabilities of generating all words in the vocabulary, \( M_o \) is the one-hot attribute-enhanced mask for object \( o \) generated according to the above process, and \( \odot \) is the element-wise product. By the backpropagation, the attribute-enhanced masks can benefit the exploration of the informative features among different objects.

Given a target ground truth sequence \( s^* = \{s_1^*, \ldots, s_K^*\} \), we train the caption generator by minimizing the negative log likelihood in (10). The attribute explorer and the caption generator are trained jointly. Totally, the loss of the generating part is calculated by (11), where \( \lambda_{att} \) is used to balance the attribute explorer,

\[ L_{cap}(\theta) = - \sum_{k=1}^{K} \log \left( p_o \left( s_k^* | s_1^*, s_2^*, \ldots, s_{k-1}^* \right) \right), \] (10)

\[ L_G = L_{cap} + \lambda_{att} L_{att}. \] (11)

D. ADVERSARIAL LEARNING

The adversarial learning scheme has facilitated numerous generation tasks [32], [33], [41]-[50]. To further promote learning the vision-language relationships, we build an adversarial discriminator on top of the caption generator. Different from the use of the discriminator in other tasks, we notice that it is difficult to judge if the generated sentences are real or fake for the diversity of language expression. Moreover, in contrary to the image pixels in continuous space, the target of caption generation is to generate natural language in discrete space, which is not conducive for training. Therefore, rather than to decide the generated sentences are real or fake, we add the corresponding visual contents as condition labels, and design the adversarial discriminator to distinguish whether the generated visual words are accurate for the given visual contents, e.g., color descriptions, clothing descriptions, and gender information. The detailed architecture of the discriminator is shown in Fig. 5. We set up four situations for the discriminator to learn vision-language correspondence:

\[ L_D = E_{v,s^*} \left[ \log D(v, s^*) \right] + E_{v} \left[ \log(1 - D(v, G(v))) \right] \]
\[ + E_{v, s} \left[ \log(1 - D(v, s^*)) \right] \]
\[ + E_{v} \left[ \log(1 - D(v, G(v))) \right], \] (12)

where \( v \) denotes the local feature of the object; \( s^* \) is the ground-truth sentence; and \( G(v) \) is the generated descriptions from the caption generator. The first two items in (12) are
TABLE 1. Comparisons with the standard datasets for video description. We group all objects which may have activities into three super object-classes: human, animal and vehicle. Except FSN, which requests to describe human activities only, the rest datasets need to describe all the three object-classes.

| Dataset           | Task               | Data Type       | Length of Video (sec) | Object Classes | Length of Sentence | Verbs per Sentence | Verbs Ratio (%) | Adjectives per Sentence | Adjectives Ratio (%) |
|-------------------|--------------------|-----------------|-----------------------|----------------|-------------------|--------------------|-------------------|-----------------------|----------------------|
| MSR-VTT          | video captioning   | clip-sen        | 20                    | 3              | 9.28              | 1.37               | 14.80             | 0.66                  | 24.83                |
| MSVD             | video captioning   | clip-sen        | 10                    | 3              | 8.67              | 1.33               | 19.60             | 0.25                  | 17.48                |
| ActivityNet      | dense video captioning | clip-sen     | 180                   | 3              | 13.48             | 1.41               | 10.40             | 0.67                  | 21.16                |
| FSN              | fine-grained video captioning | clip-sen | 5                     | 1              | 9.39              | 1.67               | 18.30             | -                     | -                    |
| Ours             | object-oriented video captioning | object-sen | 73                    | 3              | 16.56             | 2.02               | 21.00             | 1.97                  | 11.81                |

FIGURE 5. Architecture of the adversarial learning module. The discriminator requires to distinguish if the input visual contents and the sentences are paired.

common losses of a GAN. Referring to negative sampling, we add the latter two items to further allow the discriminator to learn the correspondence between the visual contents and the visual words. We randomly select the visual features for other objects as $v_{\text{un}}$. Learning the positive and negative samples at the same time, the discriminator is equipped with better ability to distinguish whether the generated sentence can describe the visual contents accurately, the performance of the caption generator can be further improved as well. Briefly, our objective is

$$G^* = \arg \min_G \max_D \mathcal{L}(G, D).$$

IV. EXPERIMENTS

In this section, we first introduce the dataset used in our object-oriented video captioning, followed by the implementation details. Next, our experimental results are reported accompanied by the comparisons with other methods. Finally, the detailed ablation studies are presented to analyze the impact of each component in our STRaNet.

A. OBJECT-ORIENTED CAPTIONS DATASET

The most widely used datasets for video captioning are the MSR-Video to Text (MSR-VTT) dataset [23] and the Microsoft Video Description (MSVD) dataset [55]. The MSR-VTT dataset contains 10K short video clips and 200K video-sentence pairs, and the MSVD dataset provides 1970 YouTube clips. Moreover, the ActivityNet Captions dataset [18], which contains 20K videos from 200 activity classes (e.g., drinking, dancing, playing games), is the most popular benchmark for dense video captioning. We summarize all these datasets in Table 1 in terms of type of data, length of sentences, verbs per sentence, adjectives per sentence.

Using this kind of data for training cannot effectively learn the functional mappings and the vision-language connections due to the one-to-many nature.

To better learn functional mappings across vision and natural language, we re-annotate a portion of videos from the ActivityNet dataset with explicit object-sentence pairs to construct a new dataset for object-oriented video captioning purpose, named the Object-Oriented Captions dataset. We choose all videos from the class of ‘playing games’ in the ActivityNet Captions dataset. The videos belonged to this class have more diverse activities and scenes than that from other classes. Also each video contains more diverse individuals and interacting activities. Totally, we re-annotate 75 videos with 534 object-sentence pairs. Each video is of length between 10 seconds to 234 seconds. On average, each video contains about 5 objects which have activities, and the average length of each object trajectory is about 248 frames. Most importantly, the identities of the objects are obtained in our object-oriented video captioning dataset. And there is one sentence associated with each object’s activities or motion trajectories in our re-annotation. According to the detailed statistics of our captioning annotations, our sentences contain almost twice many words in each sentence including verbs and adjectives. In MSR-VTT, MSVD and ActivityNet Captions, one description only has less than 1.4 verbs on average, whereas we utilize more than 2 verbs to describe each object in our re-annotation. Fine-grained Sports Narrative (FSN) Dataset is a dataset for fine-grained video captioning of Sports Narrative [39], which has more verbs to describe fine-grained actions. Even compared with FSN, our sentences provide richer information. Similarly, we analyze the adjectives in the annotated
sentences. Each sentence in our dataset has about 2 adjectives, yet, the sentences from all the other datasets only have less than 0.67 adjectives. Nonetheless, the ratio of adjectives and verbs is not significantly higher than that of the other datasets, which is considered as a result of having much fewer videos than the others, and thus, we can adequately describe the objects using these adjectives. All the above statistical differences show our data are more informative to distinguish the distinct objects and corresponding activities. In our experiments, we utilize 55 videos for training and 20 videos for test. Although the size of our videos is smaller than those of the existing datasets, it is sufficient for this work. Because we focus on a particular scene of ‘playing games’ to test the learning capabilities and effectiveness of the proposed model right now, and its generalization to other scenarios can be investigated in the future.

Table 2 further shows the constituents of words in the Object-Oriented Captions dataset. Our sentences contain more detailed information for distinguishing different individuals, for example, color of the clothes (common color, plaid, floral), type of clothes (shirt, sweater, pants, shorts, vest, dress, skirt). Fig. 6 shows two examples of our re-annotated data.

| TABLE 2. Constituents of most words in the Object-Oriented Captions dataset, where the words in boldface type are more common words. |
|-----------------|-----------------|
| **Object**      | **Activity**    |
| boy, girl, man, woman, hopscotch, dog, car, truck, bike, motor, bottle, stone, camera, building, steps, chair, mat, curb, glasses, shorts, pants, sweater, tee, shirt, dress, skirt, vest, water, board, building, box, house, chalk, coat, computer, certificate, crowd, floor, TV, flower, game, hat, hand, sofa, tree, umbrella, curtain, bubble, brick, bowl, bag, cabinet, cap, glove, photos, post, sofa, tennis, vegetable, ... |
| stand, play, look, jump, run, adjust, begin, bend, blow, celebrate, clasp, come, continue, draw, hug, fall, finish, follow, give, help, hold, lie, keep, lay, put, sit, lean, leave, listen, move, park, pass, pat, pick, pull, push, reach, squat, swing, smile, take, throw, try, turn, wait, walk, wave, dance, speak, smooth, ... |

**FIGURE 6. Examples of the Object-Oriented Captions dataset.** We annotate most salient objects in the video. Obviously, our data has more diverse sentences especially in complex scenarios.

**B. IMPLEMENTATION DETAILS**

1) OBJECT TRAJECTORY PROCESSING

For each trajectory, we sample 40 equally-spaced frames. We adopt VGG19 pre-trained on ImageNet as the backbone to extract the semantic visual features from the last pooling layer. While building the object-oriented structured trajectory representation, we feed the cropped objects into the backbone to get their object-level local features. Meanwhile, the entire frames where the object exists are fed into the backbone to get the corresponding global features. Finally, for each channel of RGB, we extract 16-dimensional color histograms, resulting in the final 4144-dim local features and 4096-dim global features.

2) SENTENCE PROCESSING

For the sentences, we remove the punctuations, split them with blank space and convert all words into lower-case. We set the maximum length of each sentence to be 25. The sentences longer than 25 are truncated. We randomly initialize all the word embedding with a fixed-size of 512.

**TABLE 3. Attribute definitions.** All the objects which have activities are grouped into three super-classes: human, vehicle and animal. For human class, we further split it into two sub-classes: male and female.

| Super-Class   | Words                          |
|---------------|-------------------------------|
| 1 Human-Male  | woman, girl, ...              |
| 2 Human-Female| man, boy, ...                 |
| 3 Vehicle     | car, truck, motor, bicycle,... |
| 4 Animal      | dog, cat, ...                 |

3) TRAINING DETAILS

The attribute explorer consists of three fully-connected layers. Table 3 shows our pre-defined super-classes. For the caption generator, the GRU is initialized to have 2 layers with 1024-dimensional hidden units. For the adversarial generator, we first process the visual features and the embedding of the yielded descriptions from the caption generator or the ground-truth separately. Then, concatenate the two modalities and decide whether they are paired via fully-connected layers. We empirically set the hyper-parameter $\lambda_{att}$ in (11) to 0.1, and adopt the adaptive moment estimation (Adam) for optimization, with the initialized learning rate being 0.0001 and a mini-batch of 50 object-sentence pairs.

**C. EXPERIMENTAL RESULTS**

1) COMPARISONS WITH STATE-OF-THE-ART METHODS

We adopt the metrics, BLEU [24], METEOR [27], ROUGE-L [26], CIDEr-D [25], which are widely used in text generation tasks, to quantitatively evaluate the proposed approach.
TABLE 4. Performance comparisons with the state-of-the-arts of video captioning methods on the Object-Oriented Captions dataset. ST denotes the method adopting the structured trajectory representation only. Full is our entire method.

| Model      | BLEU@1 | BLEU@2 | BLEU@3 | BLEU@4 | METEOR | ROUGE-L | CIDEr-D |
|------------|--------|--------|--------|--------|--------|---------|---------|
| MP-LSTM [16] | 43.2   | 29.2   | 21.7   | 16.2   | 18.2   | 40.3    | 38.8    |
| SA-LSTM [15] | 43.7   | 29.8   | 22.0   | 16.3   | 18.3   | 39.7    | 37.3    |
| S2VT [7]    | 40.2   | 26.8   | 19.6   | 14.1   | 16.6   | 37.9    | 31.7    |
| RecNet [14] | 42.7   | 28.5   | 21.0   | 15.6   | 17.5   | 40.3    | 42.1    |
| FSTA [6]    | 41.6   | 27.1   | 19.2   | 13.5   | 16.8   | 39.0    | 34.3    |
| OA-BTG [19] | 39.4   | 25.7   | 19.2   | 14.5   | 16.1   | 38.6    | 37.9    |
| Ours(ST)    | 45.0   | 32.4   | 25.0   | 19.3   | 19.7   | 45.0    | 50.2    |
| Ours(Full)  | 49.2   | 35.5   | 27.2   | 21.2   | 21.0   | 47.5    | 54.3    |

The higher scores represent better performance of the methods. We compare the performance of our method with the state-of-the-art video captioning methods, MP-LSTM [16], SA-LSTM [15], S2VT [7], RecNet [14], FSTA [6], and OA-BTG [19]. All these traditional methods utilize video-sentence pairs during training to generate descriptions for videos. To deal with the proposed object-oriented video captioning tasks, these methods still remain the same general framework, yet directly utilize object-sentence pairs to generate the descriptions for objects.

MP-LSTM is a baseline method relying on the mean pooling to process the frame features. SA-LSTM adopts a temporal attention mechanism to decide the key frames. S2VT adopts LSTM both in the encoder and the decoder. RecNet, FSTA and OA-BTG achieve the state-of-the-art performance of video captioning, however, these data-driven methods cannot perform well under small size of training samples due to its inability of understanding the temporal evolution as shown in Table 4.

FSTA uses the masks generated by Mask R-CNN to filter effective features of objects from the background. However, the features with an explicit boundary are less conducive to training. Therefore, the performance is worse than ours. In addition, it adopts Mask R-CNN as the backbone to extract visual semantic features. However, the visual features extracted from models for object detection pre-trained on COCO have less semantic information compared to the visual features extracted from models for image classification pre-trained on ImageNet. Therefore, FSTA takes more time to converge. For OA-BTG, owing to the errors of tracking operations and the use of large amounts of network parameters, it cannot obtain a satisfactory performance under small datasets as well. The MP-LSTM and SA-LSTM perform better than the other data-driven methods, because they have fewer network parameters and can be trained well under small-samples.

As shown in Table 4, by building the object-oriented structured trajectory representation only, we already obtain better performance than the other methods. With the full structure of our STraNet, we further achieve the highest BLEU@4 score of 21.2. In the proposed STraNet, by representing the moving objects via the structured trajectory representation, we thus can analyze the activities and interactions among objects from the temporal evolution of the visual features and the dynamic movement of the spatial locations, yet not relying on training on large amount of data. Therefore, our STraNet has better ability to understand the activities and attributes under limited data, and effectively learn the cross-modal relationships. From the visualization examples in Fig. 7, it can be seen that the proposed STraNet can generate more accurate descriptions for the activities, like ‘draw’, ‘throw a stone’. Meanwhile, it can describe the interactions between objects and background, and the attributes of objects in details more accurately.

2) ABLATION STUDY

To validate the importance of each component of our proposed method, we perform detailed ablation study. From Table 4, the BLEU@4 score with only building the high-level structured trajectory representation is 19.3, which is already much higher than the other methods. The significant improvement demonstrates that our object-oriented structured trajectory representation has good ability to express the objects including the activities, attributes, object-object interactions, object-background interactions as well. As shown in Table 5, after adding the attribute explorer, the performance has another great improvement to 20.2. It proves that, with the attribute explorer, the network can learn more discriminative features for different super-classes and describe the objects in details accurately. Comparing the third row and the fourth row in Table 5, we can find that the attribute-enhanced caption generator brings more improvement than the adversarial discriminator. Finally, the full scheme achieves the best performance of 21.2 in terms of BLEU@4, which is the most widely used metric in evaluating the text generation task.

Specially, for each object in the object-oriented video captioning dataset, we only have one ground-truth sentence as reference to compare with. For the rich language expressions, it is difficult to generate the sentences which are exactly the same as the ground-truths, since it is very likely two different sentences can express the same meaning. Therefore, it is more difficult to improve the BLEU score while comparing with only one reference than comparing with multiple references. In our work, we have a significant improvement from 18.3 to 21.2, which shows our structure has good ability to understand the objects and bridge the vision-language connections.

To explore the impact of different components of the object-oriented structured trajectory representation, we further do more detailed ablation study in Table 6. The BLEU@4 performance with only global features is only 16.1, which is worse than all the others. However, the performance with
SA-LSTM: The girl in pink vest and blue pants is standing on the concrete ground.
RecNet: The girl in pink vest and blue shorts plays hopscotch on the concrete ground on the lawn.
Ours: The girl in white tee and blue pants is squatting on the concrete ground and drawing the hopscotch.
GT: The woman in yellow tee and blue pants is drawing hopscotch on the sidewalk beside the lawn.

SA-LSTM: The girl in white tee and blue pants is playing hopscotch on the concrete ground.
RecNet: The girl in white tee and blue pants plays hopscotch on the sidewalk and speaking to the camera.
Ours: The girl in pink vest and black pants throws a stone on the ground and begins to play hopscotch.
GT: The little girl in pink vest and black pants runs back and forth in a playground.

GT is the ground-truth sentence for the trajectory.

FIGURE 7. Visualization examples of comparisons with SA-LSTM and RecNet.

G+L: The girl in white tee and blue pants is jumping on the concrete ground.
G+L+S: The boy in white tee and blue shorts is jumping on the right and looking at others playing games.
Ours: The girl in blue tee and blue shorts is jumping on the concrete ground and looking at the girl.
GT: The girl in blue dress plays hopscotch following the woman in brown tee.

FIGURE 8. Visualization examples for ablation study. G, L, S represent the global features, the local features, and the spatial location separately.
Ours is our full framework.

TABLE 5. Ablation study of the proposed STraNet on the Object-Oriented Captions dataset, where ST, AE, AC, and AD represent the structured trajectory representation, the attribute explorer, the attribute-enhanced caption generator, and the adversarial discriminator respectively.

| ST | AE | AC | AD | BLUE@1 | BLUE@2 | BLUE@3 | BLUE@4 | METEOR | ROUGE-L | CIDEER-D |
|----|----|----|----|--------|--------|--------|--------|--------|--------|--------|
| ✓  | ✓  | ✓  | ✓  | 45.0   | 32.4   | 25.0   | 19.3   | 19.7   | 45.0   | 50.2   |
| ✓  | ✓  |    | ✓  | 47.0   | 33.4   | 25.7   | 20.2   | 20.0   | 45.1   | 50.4   |
| ✓  | ✓  | ✓  | ✓  | 48.5   | 34.7   | 26.7   | 20.7   | 20.8   | 47.0   | 54.2   |
| ✓  | ✓  | ✓  | ✓  | 48.2   | 35.1   | 26.6   | 21.0   | 20.8   | 46.9   | 53.3   |
| ✓  | ✓  | ✓  | ✓  | 49.2   | 35.5   | 27.2   | 21.2   | 21.0   | 47.5   | 54.3   |

only local features are 16.6 which accounts for that the explicit features of objects that benefit for learning the relationships between vision and language. After adding the color information, the performance is improved to 18.7. It verifies that our color vectors can effectively solve the problem that the semantic features from top layers of a neural network usually miss most color information. Therefore, combining the color information can facilitate the generating of more accurate descriptions of the attributes. Subsequently, we further combine spatial locations of the objects to represent the moving trajectories more clearly. From the experimental results, it verifies that the temporal evolution of spatial locations works great in understanding activities along time.

Furthermore, we show some visualization results for ablation study in Fig. 8. It can be seen that, after adding the spatial locations, the model can learn more detailed spatial information of the objects, such as generating ‘in front of’,
FIGURE 9. The first is an example of describing an object which is not human. The proposed method can accurately describe the object class ‘car’, the status of the object ‘parking’, the spatial location of the object ‘across the road’ in details. The second example shows our method still can generate accurate description under the inaccurate trajectory owing to the situation of ID switch.

FIGURE 10. Failed visualization examples due to heavy overlap and too frequent ID switch.

‘on the right’. In the second example, the model can further reason the boy ‘then back to the camera’ from the spatial location evolution. In addition, after combining other modules in our framework, the detailed attributes of the objects becomes more accurate.

We show an example of describing unhuman object in Fig. 9. It can be seen that the proposed method can accurately describe the object ‘car’, the status of the car ‘parking’, the location of the car ‘across the road’. In the second example, while tracking ‘the boy in black’, ID switches to ‘the girl in pink’ due to the object occlusion and camera moving. However, the proposed method can process the inaccurate tracking results and generate correct description for ‘the boy in black’. Whereas, if the tracking results are extremely poor, for example, the objects are highly occluded or ID switch occurs too frequently, the model cannot capture accurate visual features, resulting in the overall performance degradation. In Fig. 10, we further show some failed visualization examples. How to improve the detection and tracking algorithm to adapt to natural videos is an important future direction.

Overall, the experimental results indicate that our method can reason the activities more precisely even without temporal cues from action recognition. It demonstrates that our object-oriented structured trajectory representation is
effective for expressing the activities of the objects, the object-object interactions, as well as the object-background interactions. Meanwhile, the attribute explorer, the modified attribute-enhanced caption generator, and the adversarial discriminator can significantly improve descriptions of detailed attributes and activities. Moreover, the object-sentence pairs benefit the cross-modal learning. So far, we focus on the scene of ‘playing games’, however, we believe our framework can be extended to other scenarios. All in all, the proposed method can accurately describe the concurrent objects and their corresponding dynamic attributes in more details.

V. CONCLUSION

In the paper, we propose a novel task of object-oriented video captioning scheme, which transforms the task of traditional video captioning in frame-level to the task of object-level analyses. Unlike most previous image-based methods, we propose the object-oriented video captioning framework for more realistic video understanding which can analyze the concurrent activities in details. We design the object-oriented structured trajectory network via adversarial learning, which consists of four components: the structured trajectory representation, the attribute explorer, the attribute-enhanced caption generator, and the adversarial discriminator. The proposed approach can reason the activities, object-object interactions, and object-background interactions, along the time domain based on the spatio-temporal evolutions of visual features. In addition, we modify the discriminator to make the adversarial learning more adapted to the caption generation task, and reinforce the vision-language relationship learning. To the best of our knowledge, this is the first work which proposes to shift the video captioning from generating a holistic description to detailed descriptions for each object. The proposed framework is the first attempt of object-oriented video captioning, which can understand the activities along time without temporal cues form other tasks, allowing the approach turning into real video understanding. The Object-Oriented Captions dataset is the first dataset which provides explicit object-sentence pairs. With this kind of data, further works can explore the vision-language relationships better. The experimental results demonstrate that our method have significant improvement than the state-of-the-arts for traditional video captioning. The results also indicate that the proposed method can effectively understand the activities and provide richer information of the whole scene and the concurrent objects.

| Global | Local | Color | Spatial |
|--------|-------|-------|---------|
| ✓      | ✓     | ✓     | ✓       |

| BLEU@1 | BLEU@2 | BLEU@3 | BLEU@4 | METEOR | ROUGE-L | CIDEr-D |
|--------|--------|--------|--------|--------|---------|---------|
| 41.3   | 28.4   | 21.3   | 16.1   | 16.9   | 39.3    | 38.5    |
| 42.8   | 29.7   | 22.1   | 16.6   | 18.0   | 42.3    | 45.5    |
| 46.7   | 32.6   | 24.1   | 18.1   | 19.3   | 44.5    | 52.1    |
| 45.8   | 31.4   | 24.7   | 18.7   | 19.5   | 44.3    | 59.9    |
| 45.0   | 32.4   | 25.0   | 19.3   | 19.7   | 45.0    | 50.2    |

REFERENCES

[1] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, “Video paragraph captioning using hierarchical recurrent neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 4584–4593.

[2] Q. Wu, C. Shen, L. Liu, A. Dick, and A. Van Den Hengel, “What value do explicit high level concepts have in vision to language problems?” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 203–212.

[3] M. Luo, X. Chang, Z. Li, L. Nie, A. G. Hauptmann, and Q. Zheng, “Simple to complex cross-modal learning to rank,” Comput. Vis. Image Understand., vol. 163, pp. 67–77, Oct. 2017.

[4] Z. Zeng, Z. Li, D. Cheng, H. Zhang, K. Zhan, and Y. Yang, “Two-stream multirate recurrent neural network for video-based pedestrian reidentification,” IEEE Trans. Ind. Informat., vol. 14, no. 7, pp. 3179–3186, Jul. 2018.

[5] Z. Li, F. Nie, X. Chang, L. Nie, H. Zhang, and Y. Yang, “Rank-constrained spectral clustering with flexible embedding,” IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 12, pp. 6073–6082, Dec. 2018.

[6] A.-A. Liu, Y. Qiu, Y. Yong, Y.-T. Su, and M. Kankanhalli, “A fine-grained spatial-temporal attention model for video captioning,” IEEE Access, vol. 6, pp. 68463–68471, 2018.

[7] B. Wang, L. Ma, W. Zhang, and W. Liu, “Reconstruction network for video captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7622–7631.

[8] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 2048–2057.

[9] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, “Bottom-up and top-down attention for image captioning and visual question answering,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6077–6086.

[10] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, “SCACNN: Spatial and channel-wise attention in convolutional networks for image captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 5659–5667.

[11] C. Liu, J. Mao, F. Sha, and A. Yuille, “Attention correctness in neural image captioning,” in Proc. 31st AAAI Conf. Artif. Intell., 2017.

[12] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, “Attention-based multimodal fusion for video description,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 4193–4202.

[13] J. Johnson, A. Karpathy, and L. Fei-Fei, “DenseCap: Fully convolutional localization networks for dense captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 4565–4574.

[14] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, “Sequence to sequence—Video to text,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2015, pp. 4534–4542.

[15] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, “Describing videos by exploiting temporal structure,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2015, pp. 4507–4515.

[16] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, “Translating videos to natural language using deep recurrent neural networks,” arXiv preprint arXiv:1412.4729, 2014.

[17] J. Donahue, L. A. Hendrickx, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, “Long-term recurrent convolutional networks for visual recognition and description,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 2625–2634.

[18] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, “Dense-captioning events in videos,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 706–715.

[19] J. Zhang and Y. Peng, “Object-aware aggregation with bidirectional temporal graph for video captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2019, pp. 8327–8336.
JENQ-NENG HWANG (Fellow, IEEE) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1981 and 1983, respectively, and the Ph.D. degree from the University of Southern California. In Summer 1989, he joined the Department of Electrical and Computer Engineering (ECE), University of Washington, Seattle, where he has been a Full Professor since 1999. He is currently the Associate Chair of the Global Affairs and International Development, ECE Department. He is also a Founder and the Co-Director with the Information Processing Laboratory. He has written more than 350 journal articles, conference papers, and book chapters in machine learning, multimedia signal processing, computer vision, and multimedia system integration and networking. He has authored a textbook *Multimedia Networking: from Theory to Practice* (Cambridge University Press). He has close working relationship with the industry on artificial intelligence and machine learning. He received several AI city challenges awards. He served as the Associate Chair for Research from 2003 to 2005 and from 2011 to 2015.

ZHANYU MA (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the KTH Royal Institute of Technology, Sweden, in 2011. From 2012 to 2013, he was a Postdoctoral Research Fellow with the School of Electrical Engineering, KTH Royal Institute of Technology. He was an Associate Professor with the Beijing University of Posts and Telecommunications, Beijing, China, from 2014 to 2019, where he has been a Full Professor since 2019. He has also been an Adjunct Associate Professor with Aalborg University, Aalborg, Denmark, since 2015. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in computer vision, multimedia signal processing, and data mining.

GUANG CHEN received the B.E. and Ph.D. degrees from the Beijing University of Posts and Telecommunications (BUPT), China, in 2000 and 2006, respectively. He has been an Associate Professor with BUPT since 2008. His main research interests include natural language processing, including information retrieval and text mining.

JUN GUO received the B.E. and M.E. degrees from the Beijing University of Posts and Telecommunications (BUPT), China, in 1982 and 1985, respectively, and the Ph.D. degree from Kokugakuin University, Japan, in 1993. He is currently the Dean of the Artificial Intelligence Institute, BUPT. He has published over 200 articles on the journals and conferences, including *Science*, *Nature*, *Scientific Reports*, the IEEE TRANSACTION ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, *Pattern Recognition*, AAAI, CVPR, ICCV, SIGIR, and so on. His research interests include pattern recognition theory and application, information retrieval, content-based information security, and bioinformatics.