Cascaded Revision Network for Novel Object Captioning

Qianyu Feng, Yu Wu, Hehe Fan, Chenggang Yan, Mingliang Xu, and Yi Yang

***Abstract*—Image captioning, a challenging task where the machine automatically describes an image with natural language, has drawn significant attention in recent years. Despite the remarkable improvements of recent approaches, however, these methods are built upon a large set of training image-sentence pairs. The expensive labor efforts hence limit the captioning model to describe the wider world. In this paper, we present a novel network structure, Cascaded Revision Network, which aims at relieving the problem by equipping the model with out-of-domain knowledge. CRN first tries its best to describe an image using the existing vocabulary from in-domain knowledge. Due to the lack of out-of-domain knowledge, the caption may be inaccurate or include ambiguous words for the image with unknown (novel) objects. We propose to re-edit the primary captioning sentence by a series of cascaded operations. We introduce a perplexity predictor to find out which words are most likely to be inaccurate given the input image. Thereafter, we utilize external knowledge from a pretrained object detection model and select more accurate words from detection results by the visual matching module. In the last step, we design a semantic matching module to ensure that the novel object is fit in the right position. By this novel cascaded captioning-revising mechanism, CRN can accurately describe images with unseen objects. We validate the proposed method with state-of-the-art performance on the held-out MSCOCO dataset as well as scale to ImageNet, demonstrating the effectiveness of our method.**

***Index Terms*—Captioning, novel object, visual matching, semantic matching.**

I. Introduction

IMAGE captioning has become a promising direction in the research for computer vision and language [1], [2], [3], [4], [5], [6], [7], [8]. This task aims to automatically generate a natural and concrete description of an image. Recent approaches based on the encoder-decoder structure have achieved encouraging performances on the image captioning task. However, most of the existing methods could only describe the objects shown in the training image-caption pairs, which hinders the generalization of the trained models in real-world scenarios. How to describe images with unseen objects is still a challenge for image captioning [9], [10], [11].

In this paper, we aim to alleviate this problem by equipping the image captioning model with out-of-domain knowledge. Naturally, when seeing an unknown object, human search it in the memory and describe it with the most similar object. For example, when seeing a "zebra", human tends to project the visual features and its environment and deduce that: "It is something like a horse." If an additional knowledge database is available, *e.g.*, picture flashcards or an internet search engine, human could look up the unseen object and select a better "word" to describe it. With the out-of-domain knowledge, it is possible to learn the similarity and discrepancy between a "horse" and a "zebra" and describe the unseen "zebra" with its correct name.

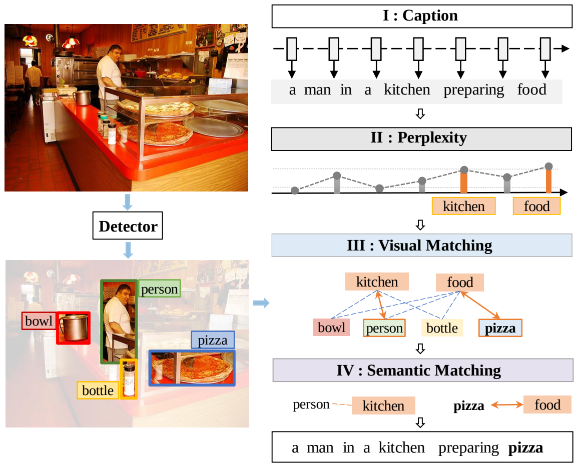


Fig. 1: An example of novel object captioning by Cascaded Revision Network (CRN). We use an object detector to provide out-of-domain information in the form of object-word pairs. CRN cascades a primary image captioner, a perplexity predictor, a visual matching module and a semantic matching module.

In this paper, we introduce a novel framework, Cascaded Revision Network (CRN), for novel object captioning. When describing an image with novel objects, the image captioner, as an agent, is first asked to try its best to characterize the image using existing in-domain knowledge. To this end, the agent could choose synonyms or the most similar words in its vocabulary to describe unknown objects. These synonyms or similar words can be ambiguous and even inaccurate due to the lack of out-of-domain knowledge. We define a sentence generated by the image captioner with only in-domain vocabulary as a primary caption.

Imitating how human-style describes an image with unseen objects, we design three cascaded operations to better revise the primary caption: 1) estimating the uncertainty of each output of the captioner; 2) searching the external knowledge database for a better description word; 3) embedding the

Qianyu Feng, Yu Wu, Hehe Fan and Yi Yang are with Centre for Artificial Intelligence, University of Technology Sydney, NSW 2007, Australia. (e-mail: qianyu.feng@student.uts.edu.au; Yu.Wu-3@student.uts.edu.au; Hehe.Fan@student.uts.edu.au; Yi.Yang@uts.edu.au)

Chenggang Yan is with the Department of Automation, Hangzhou Dianzi University, Hangzhou, 310018, China. (e-mail: cgyan@hdu.edu.cn)

Mingliang Xu is with School of Information Engineering, Zhengzhou University, Zhengzhou, 450001, China. (e-mail: iexumingliang@zzu.edu.cn).

out-of-domain object into the caption without breaking the grammar. In our approach, the above sub-tasks are executed by the perplexity predictor, the visual matching module and the semantic matching module, respectively. The perplexity predictor is designed to predict the perplexity of each output of the captioner in the primary caption. Thereafter, the agent needs to ask for help from external knowledge to generate more accurate words to revise the inaccurate ones in the primary caption. Besides, there are also cases when the agent is capable of captioning the image based on its own knowledge. The agent can handle these cases with a low perplexity of the primary caption. Next, we leverage the external knowledge to find more accurate words for the outputs with high perplexity. In CRN, a pretrained object detector is used to obtain objects with their names in the image. We then design the visual matching module to match the inaccurate outputs with the detected objects. The key-value memory mechanism is adopted to construct the communication between the captioning agent and the object detector. Specifically, the agent uses the candidates to query the memory according to the visual features of objects. To this end, the visual matching module exploits an external object detector [12] as out-of-domain knowledge. In this way, the corresponding name of the selected object becomes a candidate to revise the inaccurate outputs in the primary caption. However, the object detector is not always reliable to retrieve all the objects in the image accurately. In this case, the visual matching module would generate a wrong matching proposal. Therefore, the semantic matching module is designed to eliminate such incorrect matching proposals. Specifically, it measures the similarity between the ambiguous word and the object name with an out-of-domain pretrained word embedding. The incorrect visual matching pair will get a low semantic matching score and thus be ignored. By this cascaded captioning-revising mechanism, novel objects will be described accurately in the final caption sentence. An example of the novel object captioning by CRN is illustrated in Figure 1.

Our proposed method turns out with competitive results compared to the current state-of-the-art performance on the held-out MSCOCO dataset. We also scale the proposed CRN to a larger dataset: ImageNet [13]. With additional analysis, it is revealed that our approach not only improves captioning with novel objects as well as images without novel objects. Finally, the main contributions of this paper are summarized as follows:

- We propose a novel cascaded framework for novel object captioning by imitating how we humans describe an image with unseen objects. At first, the model tries its best to generate a primary caption based on in-domain knowledge. We then gradually revise the primary captioning sentence with a series of cascaded operations.
- In this cascaded network, we develop a perplexity predictor, a visual matching module and a semantic matching module to revise the primary captioning.
- To our knowledge, we are the first to embed the out-of-domain knowledge both visually and semantically in the captioning model to better describe the novel objects.

II. Related Work

Deep Image Captioning. Given an image, the goal of image captioning is to generate a natural and accurate sentence to describe the image. Early approaches [14], [15] composed image captions via slot filling which separate the object recognition and the language template generation. These approaches may generate natural sentences but are less related to the visual contents. Deep Learning has elevated the performance of captioning models with images and videos. Most of related works [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28] follow a multi-modal framework which combines CNN [29], [30] and RNN like Long Short-Term Memory (LSTM) [31] and Gated Recurrent Unit (GRU) [32]. Visual features in high-level with semantic information are first extracted by the CNN encoder, while the RNN decoder predicts the description word by word according to visual features. However, these methods do not consider the situation where a large number of unseen objects exist in the images.

Zero-shot Learning. With the booming development of techniques in computer vision, lack of well-labeled data becomes the bottleneck of performance. Image-paired sentences are in large scarcity and the label tagging of captioning costs much more than other tasks. Zero-shot learning is a good solution to resolve the out-of-domain adaptation for models with limited knowledge. There has been a recent surge on the zero-shot tasks [33], [34], [35] which aims to recognize objects unseen during the training stage. [36], [37], [38] focus on the unseen categories in the target domain to help align the divergence across domains. [39] proposes to first capture the attributes of the unseen objects, then infer the class label with the most similar set of features.

Novel Object Captioning. The novel object captioning task attracts increasing attention recently. The problem exists in how to leverage the unpaired image and semantic data [40] to better describe the unseen objects. The Deep Compositional Captioner (DCC) is proposed by [9], a pilot work to put forward the task of novel objects captioning. DCC [9] combines visual groundings of lexical units to generate descriptions about objects which are not present in caption corpora. Novel Object Captioner (NOC) [41] is introduced as an end-to-end framework training the object classification, language model, and the captioning jointly. The detection model is integrated with the language sequence model by copying detection results into the prediction of the RNN-based decoder model in [42]. An approach is proposed in [10] to generate a language template along with slots and the corresponding region in the image at first. Then objects are fit into the slots by recognizing the region with a detection model. But they have to manually define the category of the novel object with an existing one when captioning. A placeholder is used in [11] to take the place of the novel objects, which generalizes the concept of the novel object but still loses the information of the matched object. These methods rely too much on visual detection. They neglect the original lexical context information. To the best of our knowledge, our model is the first captioning model with self-awareness and a two-way revision mechanism.

Summary. The proposed method CRN focuses on generating accurate captions of images with novel objects. With the cascaded revision mechanism, CRN exploits and embeds the out-of-domain knowledge with the in-domain captions which can generate better captions for images with novel objects. Both of the visual matching and semantic matching ensure the accuracy of the revised caption for out-of-domain objects.

III. The Proposed Approach

The Cascaded Revision Network(CRN) is designed to better embed the out-of-domain knowledge into the in-domain captions. In this section, we first introduce the traditional image captioning model in Section III-A. We then show how CRN describes images with novel objects in Section III-B. The full framework of CRN is illustrated in Figure 2.

A. Image Captioning Model

The main task of an image captioning model is to generate a natural language sentence to describe an image and also maintain the fluency of the sentence. Given an image I and the ground truth caption $w = \{w_1, w_2, ..., w_T\}$, the objective of the captioning model is to minimize

$$\begin{aligned} L &= -\log\ p(y|I) = -\log\ p(w_1, w_2, ..., w_T|I) \\ &= -\sum_{t=1}^{T} \log\ p(w_t|w_1, w_2, ..., w_{t-1}, I). \end{aligned} \tag{1}$$

Eq. 1 aims to maximize the likelihood of each word in the ground-truth caption. Usually, the term $p(w_t|w_1, ..., w_{t-1}, I)$ is modeled by a Long Short-Term Memory network (LSTM) [31]:

$$y_t, h_t = LSTM(x_t, h_{t-1}), \tag{2}$$

$$p_t(\cdot|y_t) = softmax(W_v y_t), \tag{3}$$

where x_t is the embedding of the current word w_t. The beginning of the sentence w_0 is defined as <START>. The hidden state is initialized with the extracted representation of image I. The distribution $p_t(\cdot|y_t)$ is a parametric function of the output y_t. LSTM first generates the current hidden state h_t and then emits the distribution by a fully-connected layer according to h_t. For simplicity, we use $\pi(\cdot|y_t)$ to denote this distribution:

$$\pi(\cdot|y_t) = p(\cdot|w_1, w_2, ..., w_{t-1}, h_0). \tag{4}$$

The current word is generated by

$$w_t = \arg\max_{w}\ \pi(\cdot|y_t). \tag{5}$$

During training, the previous ground-truth words are given. When conducting the evaluation, the previous ground-truth words are unavailable and are generated by maximum likelihood estimation (MLE).

B. Cascaded Revision Network

CRN aims to alleviate the problem of novel object captioning by equipping the model with out-of-domain knowledge. To exploit the out-of-domain knowledge, CRN adopts a cascaded captioning-revising mechanism. Following [10], [11], in this paper, we use the out-of-domain knowledge provided by an object detector and a pretrained word embedding lookup table. CRN contains four cascaded modules: a primary image captioner, a perplexity predictor, a visual matching module, and a semantic matching module. With the setting of *pseudo objects*, CRN learns to distinguish the ambiguous words which are inconsistent with the images.

1) Image Captioner: Since the captioning model never sees the novel objects before, it will generate an output with its existing vocabulary. Ambiguous or even inaccurate words may be used when describing the unknown objects. We denote words in the vocabulary of the image-paired captions as V_c. The objects neither in the images nor the captions are novel objects denoted as O_u. Several words in V_c are selected as novel objects to simulate the existence of novel objects, which are denoted as O_i. Objects $\in O_i$ act as the role of novel objects during training which do not exist in the vocabulary of the model. We replace the selected words $\in O_i$ with *pseudo objects*. With the open-source pretrained embeddings, each object $\in O_i$ is paired with its most similar word $\in V_c$, which acts as a *pseudo object*. The *pseudo object* and its corresponding object $\in O_i$ form a negative pair as the inaccurate description of an object. Furthermore, in order to inform the captioner about the existence of *pseudo object*, we design an additional novel label $\hat{n}$ of each word w to indicate whether it is a novel object or not:

$$\hat{n} = \begin{cases} 1, w \in O_i \\ 0, otherwise. \end{cases} \tag{6}$$

Another embedding function ϕ_n is adopted to embed the novel label into the inputs of the captioner. At time step t, the input vector of the captioner x_t is the concatenation of the embedding of w_{t-1} and its novel label $\hat{n}_{t-1}$:

$$\begin{aligned} x_t &= [\phi_e(w_{t-1}), \phi_n(\hat{n}_{t-1})] \\ &= \left[W_e I^w_{t-1}, W_n I^n_{t-1}\right], \end{aligned} \tag{7}$$

where $W_e \in \mathbb{R}^{N_v \times D_e}$ is the word embedding matrice of the vocabulary V_c. N_v is the number of the vocabulary. D_e denotes the dimension of embedding. $W_n \in \mathbb{R}^{2 \times D_e}$ denotes learnable weight matrice of the novel label $\hat{n}_t$. I^w_{t-1} and I^n_{t-1} are the corresponding one-hot encoding of w_{t-1} and $\hat{n}_{t-1}$. With the input vector x_t, the output hidden state of captioner is:

$$h_t = w_h^T \tanh{(W_s x_t + W_z h_{t-1})}, \tag{8}$$

where w_h^T, W_s, W_z are weights to be learned. At each time step, the distribution of the conditional probabilities over all possible words $\in V_c$ is:

$$p_t = softmax(W_p h_t + b_p), \tag{9}$$

where W_p, b_p are learned weights and biases.

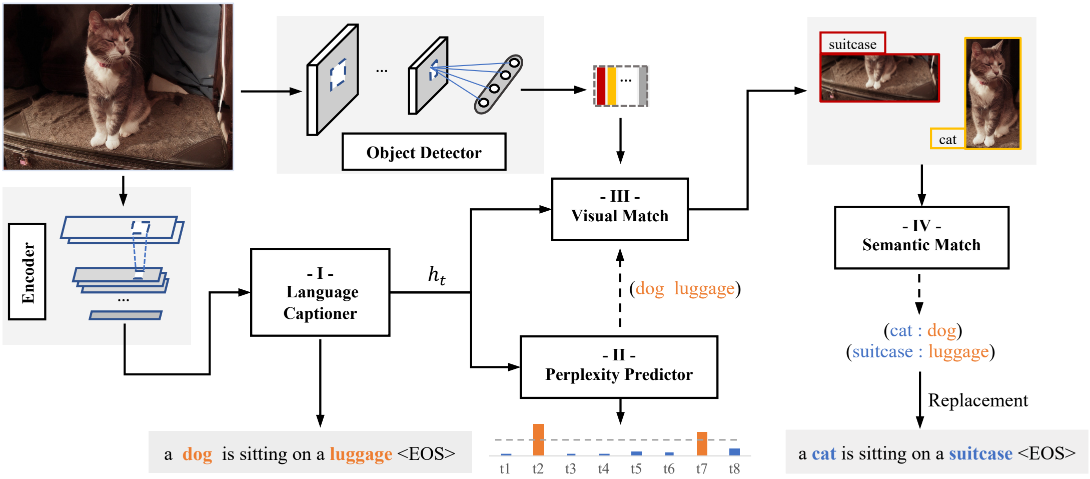


Fig. 2: The overview of the evaluation stage of the proposed framework. “Cat” and “suitcase” are novel objects to the captioning model which never appear in training. At stage I, the image captioner generates a sentence based on the existing vocabulary. We propose to revise this sentence to get a more accurate description by a series of cascaded operations. At stage II, the perplexity predictor is applied to find ambiguous words, *e.g.*, “dog” and “luggage”. At stage III and IV, we match and replace these words with detected objects based on both visual and semantic similarity. Finally, detected objects “cat” and “suitcase” are fit in the right positions to revise the primary caption.

2) Perplexity Predictor: The intuition behind the proposed method is to enable the captioner to justify whether the word is semantically consistent with the image or not. To revise the primary caption, the perplexity predictor is designed to figure out the ambiguous words in it. At the very first step of captioning, the extracted feature of the image is fed into the captioner as the initialization of the hidden state. The captioner embeds both the visual feature of the image and the linguistic feature of the sentence. Then the perplexity predictor learns from the features encoded by the captioner to decide the perplexity of each output semantically. Thus, it is aware of the ambiguity of the outputs in the primary caption. We here define the level of ambiguity as semantic perplexity. In information theory, perplexity is a measurement of how well a model predicts a sample. The perplexity of the current output of the captioner is calculated as:

$$m_t = \sigma(W_m h_t + b_m), \tag{10}$$

where W_m, b_m are learned weights and biases for this layer. σ is the sigmoid activation of confidence probability. All outputs with high perplexity scores will be regarded as inaccurate candidates and will probably be replaced by a matched object in the next revision step. A threshold τ_p is defined here to justify whether the current word is accurate or not. If m_t surpasses τ_p, it indicates that the current prediction will be taken as an inaccurate candidate.

During the training, only novel labels of the selected novel objects are positive samples for the perplexity predictor. These objects are all noun words. And other words, *e.g.*, verbs and adjectives, are negative samples in training. Therefore, it is a sub-goal for the perplexity predictor to learn that only noun words are the inaccurate targets to be revised. Thus, no additional post-processing is used to identify the types of predicted words. To further ensure the quality of revision, we design two modules to match detected objects with inaccurate candidates. During the inference, the captioner takes its output at the last step as the previous word to predict the current word. Similarly, the prediction of the perplexity predictor at the last step is also used as the novel label of the previous output. With the image captioner and the perplexity predictor introduced above, the corresponding objective cross-entropy loss function is:

$$\begin{aligned} &L_{cap}(w_{1:t-1}, I; \theta) = \\ &-\frac{1}{T}\Big(\sum_{t=1}^{T} \log p(w_t|w_{1:t-1}) + \sum_{t=1}^{T} \log p(m_t|w_{1:t-1})\Big). \end{aligned} \tag{11}$$

3) Visual Matching Module: The visual matching module is responsible for acquiring objects in the image with the knowledge of the detector and generate revision proposals. To introduce novel objects out-of-domain into the image captioner, we employ an available trained object detector. Thus, we can take advantage of the detector to detect objects in the image, which are further used to revise the inaccurate words in the primary caption. The extracted visual features $V_d \in \mathbb{R}^{N_o \times D_v}$. N_o is the number of detected objects. D_v is the dimension of the visual feature. The predicted class labels can also be obtained from the predictions of the detector. N_d is the number of target classes of objects. We extract the visual features of objects from the ROI pooling layer of the object detector following [2]. With the hidden state h_t of captioner at time step t, the visual similarities between the current encoded features of all detected objects are calculated as:

$$S_t = V_d h_t. \tag{12}$$

Then we address the probabilities over all classes of the detector at time t:

$$O_t = S_t O_d. \tag{13}$$

Each inaccurate word will be matched with a detected object, which is regarded as a candidate to revise the primary caption. For the matching between the output of captioner and features of the detected objects, the objective for training this module is defined as:

$$L_{det}(h_t;\theta) = -\frac{1}{N_d}\sum_{i=1}^{N_d}\hat{n}_t \log p(o_t|h_t), \tag{14}$$

where N_d is the number of detected objects at time step t, n_t is used as the mask of the current ground truth word which is defined in Eq (6). These three modules of CRN are jointly trained during the training of CRN.

4) Semantic Matching Module: Simply replacing the inaccurate words with the visually matched objects may break the semantic structure of the sentences. Besides, due to the limitation of the compressed features, objects with salient features tend to be matched with a high frequency. It is observed that many ambiguous words are matched with the same detected object, while some are not relative semantically. Therefore, we elevate the quality of revision by employing the semantic matching as the last step. With the selected objects from the detection model, the word similarity is calculated with the pretrained word embedding look-up. The cosine similarity is used to measure the distance between the novel objects and the caption words. The word with the largest word similarity is replaced by the detected object.

Finally, the full framework of CRN is proposed to deal with the captioning of images with novel objects. With the different modules cascaded in the model, each module is optimized with a sub-goal. The gap between the novel object and the existing knowledge is estimated by the perplexity of the captioning outputs and bridged by the visual matching and the semantic matching modules.

IV. Experiments

We start by describing the setups of this task and our experiments. Then, the results of our methods and the state-of-the-art methods in history are compared on the held-out MSCOCO dataset. Furthermore, several ablation studies are carried out with competitive results to prove the effectiveness and reliability of our proposed method.

A. Experimental Settings

MSCOCO is a widely used benchmark for many tasks, including image captioning [40]. The held-out subset of the MSCOCO dataset following [9], [41], [42] are used as the training set in our experiments. In [9], eight classes of MSCOCO objects are chosen. None of the eight classes is included in the captioning in the training split set, but all of them are in the evaluation split set. We follow the same training, validation, and test split in [9] in order to generate comparable captioning results.

Pseudo Object Processing. To select the *pseudo object* of each novel object $\in O_i$ in the train set, we employ the open-source pretrained embedding weights of Glove [44] following [9], [42]. The dimension of word vector is 300. We stress that we have not used any other semantic data or description for these objects; neither do we manually change any word. First, the word embedding vector of each word in the training vocabulary is retrieved from the pretrained embedding weights of Glove. Second, the cosine similarity of each selected novel object is calculated with the rest words in training vocabulary. Finally, the pseudo object is selected with the highest similarity score for each selected novel object. The examples of the *pseudo objects* is shown in Figure 3. All classes except the eight held-out classes in MSCOCO are chosen as novel objects O_i in the train set which are replaced with *pseudo objects* in the in-domain vocabulary. For example, “umbrella”→“parasol”, “zebra”→“horse”, “sandwich”→“burger”, etc. It comes out that the plural format of the word tends to be the most similar word to itself, *e.g.*, “sandwiches” to “sandwich”. It is meaningless if we use the word “sandwiches” to take the place of “sandwich”, as they refer to the same object. It is also noticed that some words are composed of multiple words cannot be found in the pretrained word embedding, *e.g.*, “hot dog”, “hair drier”, etc. In this case, to prevent manual intervention, we simply average the embeddings of the two words.

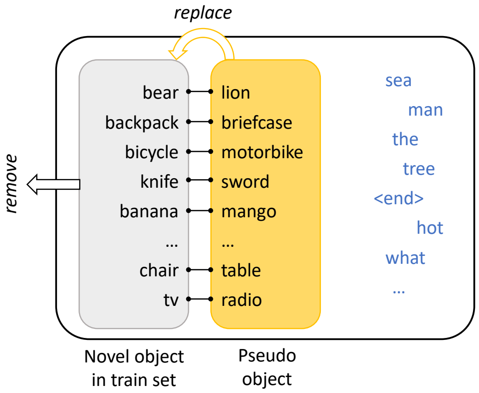
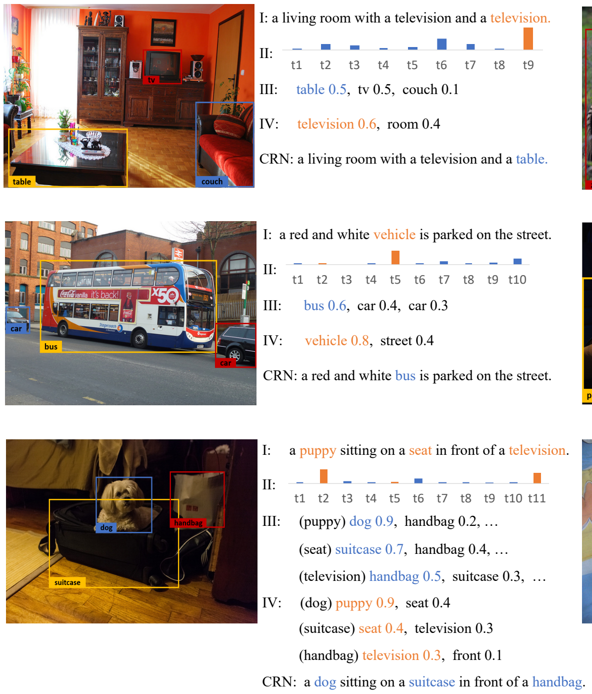


Fig. 3: Example of words in the in-domain vocabulary. The selected novel objects in the training captioning sentences are replaced with pseudo objects. Each pseudo object is matched by comparing the semantic similarity with the selected novel object, *e.g.*, “banana” is replaced by “mango” in our training. We use the pseudo objects to train the matching and replacement mechanism in our framework.

Experiment Details. We apply a 16-layer VGG pretrained on ImageNet following [9], [41], [42] as the image encoder in our model. Parameters of the encoder are frozen during the training. The outputs by layer fc7 are used as the representation of the image and fed into the language decoder. The dimension of the image feature is 4,096. In order to introduce the novel objects into the final captions, a popular open-source pretrained Faster-RCNN model [12] is adopted to detect and crop the objects in an image following [43], [10], [11]. Then, we reuse the VGG Net mentioned above to extract the visual features of the detected objects. The pretrained detection model is released by [45], which is trained on all the 80 classes of objects in the MSCOCO detection dataset. We adopt the LSTM as the captioner with one layer and its dimension is 1,024. We use Adam optimizer with an initial learning rate of 1×10-3 and anneal the learning rate by a factor of 0.9 every 2 epochs. We train the model up to 50 epochs with early stopping. Note that we do not finetune the CNN network, which extracts the feature of images during training. We set the batch size to be 256.

Method	METEOR	F_{bottle}	F_{bus}	F_{couch}	$F_{microwave}$	F_{pizza}	F_{racket}	$F_{suitcase}$	F_{zebra}	$F_{average}$
DCC [9]	21	4.63	29.79	45.87	28.09	64.59	52.24	13.16	79.88	39.78
NOC* [41]	21.32	17.78	68.79	25.55	24.72	69.33	55.31	39.86	48.79	48.79
LSTM-C [42]	22	29.07	64.38	26.01	26.04	75.57	66.54	55.54	**92.03**	54.40
LSTM-C*	23	29.68	74.42	38.77	27.81	68.17	**70.27**	44.76	91.40	55.66
Base+T4† [43]	**23.6**	16.3	67.8	48.2	29.7	77.2	57.1	49.9	85.7	54.0
NBT+G [10]	22.8	7.1	73.7	34.4	**61.9**	59.9	20.2	42.3	88.5	48.5
DNOC [11]	21.57	33.04	76.87	53.97	46.57	75.82	32.98	**59.48**	84.58	57.92
CRN (ours)	21.31	**38.05**	**78.40**	**55.93**	53.76	**81.43**	62.02	57.69	85.38	**64.08**

TABLE I: Comparison with the state-of-the-art methods on F1 score and METEOR score. All results are generated with image feature extracted by VGG-16 [29] and without beam search. † is method with Resnet-based CNN and beam search. F1-score values are reported in format of percentage (%). * indicates training with pretrained Glove word embedding weights.



I: a living room with a television and a television.
II:
III: table 0.5, tv 0.5, couch 0.1
IV: television 0.6, room 0.4
CRN: a living room with a television and a table.

I: a giraffe standing next to another.
II:
III: zebra 0.9, zebra 0.7
IV: giraffe 0.6, standing 0.1
CRN: a zebra standing next to another.

I: a red and white vehicle is parked on the street.
II:
III: bus 0.6, car 0.4, car 0.3
IV: vehicle 0.8, street 0.4
CRN: a red and white bus is parked on the street.

I: a woman is looking at a railway station.
II:
III: bus 0.7, person 0.4
IV: railway 0.6, station 0.4
CRN: a woman is looking at a bus station.

I: a puppy sitting on a seat in front of a television.
II:
III: (puppy) dog 0.9, handbag 0.2, …
(seat) suitcase 0.7, handbag 0.4, …
(television) handbag 0.5, suitcase 0.3, …
IV: (dog) puppy 0.9, seat 0.4
(suitcase) seat 0.4, television 0.3
(handbag) television 0.3, front 0.1
CRN: a dog sitting on a suitcase in front of a handbag.

I: a kitten laying on top of a notebook.
II:
III: (kitten) cat 0.9, book 0.2
(notebook) book 0.9, cat 0.2
IV: (cat) kitten 0.9, laying 0.5
(book) notebook 0.7, top 0.5
CRN: a cat laying on top of a book.

Fig. 4: Captions generated by CRN on the held-out MSCOCO where images contain unseen objects. Boxes with colors are object candidates proposed by a pretrained object detector. Sentences with tag “I” are generated by CRN-I only with the in-domain vocabulary. At step “II”, the perplexity predictor outputs the perplexity of each word. Each ambiguous word with high perplexity will be matched with a detected object at step “III”. At the last step, the matched objects will be fit into the primary caption by semantic matching.

Compared Approaches. To evaluate on the held-out MSCOCO, results of our proposed method are compared with DCC [9], NOC [41], LSTM-C [42], Base+T4 [43], NTB+G [10] and DNOC [11] to demonstrate the competitiveness. During the methods, NTB+G and DNOC do not use the additional semantic data. We follow the same zero-shot setting in our experiments. Furthermore, the results of several ablation versions of the proposed model are compared and discussed. In order to prove the advantage of CRN not only exists in the novel object captioning, we also evaluate F1 scores of other known objects $\in W_{paired}$ in Table II.

B. Compared to the state-of-the-art methods

Captions are being evaluated with the widely-used COCO caption evaluation tool. For the task of novel object captioning, only the METEOR [47] metric is not enough for the evaluation. Sentences with good grammar can obtain high scores even without mentioning the novel objects. The caption is deemed accurate only if the correct novel object appears at least once in the sentence. The results of our proposed model with the F1 scores to measure the performance on novel objects and METEOR are presented in Table I along with all state-of-the-art methods on the held-out MSCOCO dataset. The F1-scores of all novel objects surpass the best state-of-the-art result, while the average F1 score achieves 64.08% (6.16% higher than 57.92%). It is observed that methods [41], [42] with external text data, including the novel objects perform better on several objects than the proposed CRN. However,

Method	F_{bear}	F_{cat}	F_{dog}	$F_{elephant}$	F_{horse}	$F_{motorcycle}$	$F_{average}$
LRCN [46]	66.23	75.73	53.62	65.49	55.20	71.45	64.62
DNOC [11]	62.86	87.28	71.57	77.46	71.20	77.59	74.66
CRN	60.38	86.74	74.04	81.41	75.36	78.39	**76.05**

TABLE II: Comparison on F1 scores of pseudo novel objects from subset 1 with baseline LRCN and DNOC.

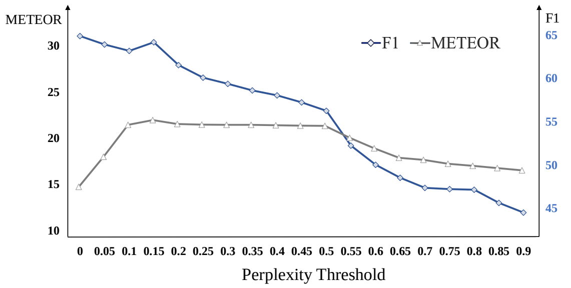


Fig. 5: The effect of perplexity threshold on the performance of CRN on the held-out MSCOCO. The left y-axis is the scale of METEOR and the right y-axis measures the average F1 score.

Method	METEOR	F_{average}
LRCN [46]	19.33	0
CRN I	18.24	0
CRN I + II	19.65	45.30
CRN w/o II	19.26	53.31
CRN w/o IV	20.85	56.32
CRN w/o III	21.01	62.08

TABLE III: Ablation studies on each component of CRN.

Method	Novel	F_{average}	Acc
NOC [41]	69.08	15.63	10.04
LSTM-C [42]	72.08	16.39	11.83
CRN	**77.92**	**19.5**	**16.34**

TABLE IV: Comparison with state-of-the-art methods on ImageNet.

there may always be objects novel to the captioner that it never learned from the text nor the image. Our METEOR score is lower than LSTM-C with GloVe [42] a little bit. Nevertheless, our experiments are carried out based on the zero-shot setting. What is more, all of the eighty classes of objects are novel to our model. It is explainable that the captioning model can better describe the context of the known objects than objects never seen before.

C. Ablation Studies

The ablation studies are conducted on the held-out MSCOCO dataset with the same setting mentioned above. We compare different ablation versions of CRN to prove the effectiveness of the sub-modules: the perplexity prediction and the revision of objects. Results are listed in Table III.

CRN I is CRN only with the captioner, which knows nothing about the novel objects as LRCN. Thus, the F1 score is 0. The existence of the *pseudo novel object* leads to the drop of the METEOR score. **CRN I+II** (perplexity predictor) adds the second task: predicting the perplexity of each word. If the perplexity goes beyond the threshold τ_p, the word will be replaced by a detected object randomly selected from the results of the detection model. It brings a significant rise in the F1 score. The METEOR also increases from 18.12 to 19.65. The threshold τ_p is set as 0.15 in our experiments learned by the model. **CRN w/o II** is CRN without the perplexity predictor. As the average number of words above the perplexity threshold in the training stage is 1.7 per sentence, we choose two positions in the sentence to replace the detected object matched with the two-way matching of visual similarity and word similarity. It shows that *CRN I+II* is better on METEOR than *CRN w/o II*, which indicates the value of the perplexity predictor. The average F1 score of *CRN w/o II* is 53.31%, 8.01% higher than *CRN I+II*. **CRN w/o IV** (semantic matching) is CRN without the matching of word similarity. Objects are matched only with the features from the language decoder and visual features of objects detected. The F1 average score increases from 45.30% to 56.32%. **CRN w/o III** (visual matching) objects are matched only with word similarity, which outperforms *CRN w/o IV* by 5.76% on F1 score. With full stages, our model can better capture features of the unknown objects on visual out-looking and semantic context, which composes more accurate captions about the image. Furthermore, in order to show the advantage of the proposed model not only exist in the novel object captioning, we also evaluate F1 scores of other words $\in W_s$. Our model is also able to generate accurate descriptions of known objects. F1 scores on a different group of known objects are listed in Table II. It turns out that the performance of these objects is also quite qualitative. Figure 4 shows some examples of image captioning results with novel objects. Some failure cases are also observed which are shown in Fig. 6. It is observed that the revision sometimes does not take advantage of the spatial location of objects, e.g., in the three pictures in the left column, the word used to describe an object tends to be replaced with another word referring to an object with more distinctive representation. On the other hand, the matched objects sometimes are already described in the other positions in the sentence. After the revision, the same object will appear twice, e.g., pictures in the right column.

Threshold of Perplexity. We present the performance of F1 score and METEOR along with the change of threshold of perplexity in Figure 5. When the threshold is 0, the F1 score achieves quite high but with a low meteor. It indicates that the objects detected by the detection model in the image are replaced into the caption while it destroys the grammar and structure of the sentence. When the threshold is between (0, 0.15), METEOR both get higher, while F1 score drops slightly. It indicates that the threshold limits the ambiguous words area while the number of objects replaced into the sentence decreased. After that, the F1 score goes down when the threshold is larger than 0.15. METEOR score also decreases and drops more when the threshold goes beyond 0.5.

Scale to Larger Dataset. The proposed CRN takes advantage of an expertised detector to introduce the novel objects. Considering the out-of-MSCOCO objects, a detector with a larger vocabulary should be adopted. Hence, we report the performance of using the detector pretrained on Visual Genome [48] and scaling CRN to ImageNet. Results of additional experiments are reported in Table IV.

V. Conclusion

In this paper, we present a novel cascaded framework CRN to deal with captioning with novel objects. To overcome the

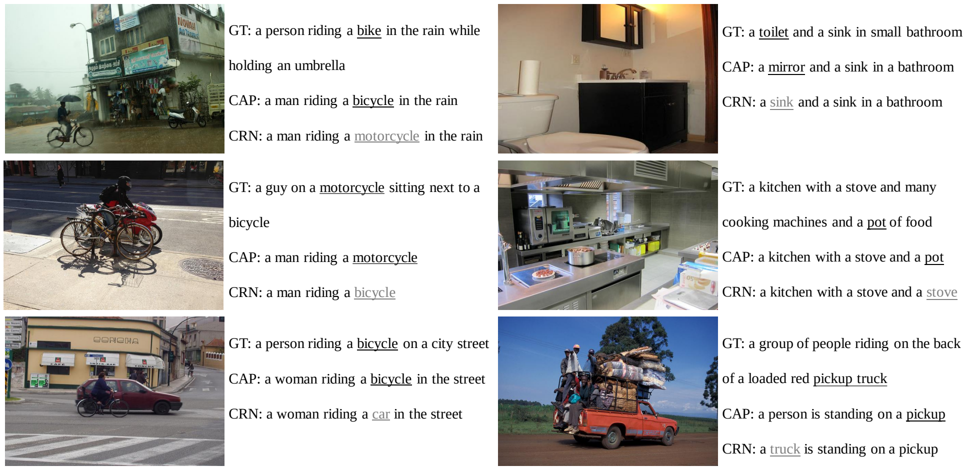

GT: a person riding a bike in the rain while holding an umbrella
CAP: a man riding a bicycle in the rain
CRN: a man riding a motorcycle in the rain
GT: a toilet and a sink in small bathroom
CAP: a mirror and a sink in a bathroom
CRN: a sink and a sink in a bathroom
GT: a guy on a motorcycle sitting next to a bicycle
CAP: a man riding a motorcycle
CRN: a man riding a bicycle
GT: a kitchen with a stove and many cooking machines and a pot of food
CAP: a kitchen with a stove and a pot
CRN: a kitchen with a stove and a stove
GT: a person riding a bicycle on a city street
CAP: a woman riding a bicycle in the street
CRN: a woman riding a car in the street
GT: a group of people riding on the back of a loaded red pickup truck
CAP: a person is standing on a pickup
CRN: a truck is standing on a pickup

Fig. 6: Failure cases in the evaluation on MSCOCO dataset using CRN.

gap between existing knowledge and objects out-of-domain, the captioner in CRN is aware of what is ambiguous or unknown to itself. Furthermore, with a two-way matching mechanism, the unknown object can be better matched and fit in the caption. At a higher level, our proposed method decouples the captioning of novel objects to two sub-tasks: what is the novel object and where to put the novel object. By applying the two-way matching, CRN better integrates the out-of-domain knowledge both visually and semantically.

References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015.

[2] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *CVPR*, 2016.

[3] H. Chen, G. Ding, Z. Lin, S. Zhao, and J. Han, "Show, observe and tell: Attribute-driven attention model for image captioning," in *IJCAI*, 2018.

[4] Y. Mao, C. Zhou, X. Wang, and R. Li, "Show and tell more: Topic-oriented multi-sentence image captioning," in *IJCAI*, 2018.

[5] 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 VQA," in *CVPR*, 2018.

[6] P. V. K. Borges, N. Conci, and A. Cavallaro, "Video-based human behavior understanding: A survey," *IEEE transactions on circuits and systems for video technology*, vol. 23, no. 11, pp. 1993–2008, 2013.

[7] R. V. H. M. Colque, C. Caetano, M. T. L. de Andrade, and W. R. Schwartz, "Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 673–682, 2016.

[8] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang *et al.*, "T-cnn: Tubelets with convolutional neural networks for object detection from videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2896–2907, 2017.

[9] L. Anne Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, "Deep compositional captioning: Describing novel object categories without paired training data," in *CVPR*, 2016.

[10] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *CVPR*, 2018.

[11] Y. Wu, L. Zhu, L. Jiang, and Y. Yang, "Decoupled novel object captioner," in *ACM MM*, 2018.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR*, 2009.

[14] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu, "I2t: Image parsing to text description," *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1485–1508, 2010.

[15] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *NIPS*, 2011.

[16] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," in *ICLR*, 2015.

[17] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, "Uncovering the temporal context for video question answering," *International Journal of Computer Vision*, vol. 124, no. 3, pp. 409–421, Sep 2017.

[18] A. Karpathy and F. Li, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015.

[19] J. Donahue, L. A. Hendricks, S. Guadarrama *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015.

[20] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.

[21] C. Deng, E. Yang, T. Liu, W. Liu, J. Li, and D. Tao, "Unsupervised semantic-preserving adversarial hashing for image search," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 4032–4044, Aug 2019.

[22] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille, "Learning like a child: Fast novel visual concept learning from sentence descriptions of images," in *ICCV*, 2015.

[23] L. Li, S. Tang *et al.*, "Image caption with global-local attention," in *AAAI*, 2017.

[24] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, and D. Xu, "Image classification by cross-media active learning with privileged information," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2494–2502, 2016.

[25] X. Wang, J. Wu, D. Zhang, Y. Su, and W. Y. Wang, "Learning to compose topic-aware mixture of experts for zero-shot video captioning," in *AAAI*, 2019.

[26] S. Coşar, G. Donatiello, V. Bogorny, C. Garate, L. O. Alvares, and F. Brémond, "Toward abnormal trajectory and event detection in video surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 683–695, 2016.

[27] J. Wu and H. Hu, "Cascade recurrent neural network for image caption generation," *Electronics Letters*, vol. 53, no. 25, pp. 1642–1643, 2017.

[28] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt *et al.*, "From captions to visual concepts and back," in *CVPR*, 2015, pp. 1473–1482.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[32] K. Cho, B. van Merrienboer *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *EMNLP*, 2014.

[33] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2013.

[34] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, “Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly,” in *CVPR*, 2017.

[35] Z. Ding, M. Shao, and Y. Fu, “Low-rank embedded ensemble semantic dictionary for zero-shot learning,” in *CVPR*, 2017.

[36] P. P. Busto and J. Gall, “Open set domain adaptation,” in *ICCV*, 2017.

[37] K. Saito, S. Yamamoto, Y. Ushiku, and T. Harada, “Open set domain adaptation by backpropagation,” in *ECCV*, 2018.

[38] Q. Feng, G. Kang, H. Fan, and Y. Yang, “Attract or distract: Exploit the margin of open set,” in *ICCV*, 2019.

[39] A. Frome, G. S. Corrado, J. Shlens *et al.*, “Devise: A deep visual-semantic embedding model,” in *NIPS*, 2013.

[40] T. Lin, M. Maire *et al.*, “Microsoft COCO: common objects in context,” in *CVPR*, 2014.

[41] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. J. Mooney, T. Darrell, and K. Saenko, “Captioning images with diverse objects,” in *CVPR*, 2017.

[42] T. Yao, Y. Pan, Y. Li, and T. Mei, “Incorporating copying mechanism in image captioning for learning novel objects,” in *CVPR*, 2017.

[43] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “Guided open vocabulary image captioning with constrained beam search,” in *EMNLP*, 2017.

[44] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *EMNLP*, 2014, pp. 1532–1543.

[45] J. Huang, V. Rathod, C. Sun *et al.*, “Speed/accuracy trade-offs for modern convolutional object detectors,” in *CVPR*, 2016.

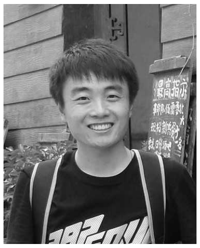
[46] J. Donahue, L. A. Hendricks, S. Guadarrama *et al.*, “Long-term recurrent convolutional networks for visual recognition and description,” in *CVPR*, 2015.

[47] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in *ACL Workshop*, 2005, pp. 65–72.

[48] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma *et al.*, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

Qianyu Feng received the M.S. degree in Shanghai Jiao Tong University, China, in 2018. She is currently a Ph.D. student in University of Technology Sydney, Australia. Her research interests include video understanding, multimodal learning and domain adaptation.

Yu Wu received the B.S. degree in Shanghai Jiao Tong University, China, in 2015. He is currently a Ph.D. student in University of Technology Sydney, Australia. His research interests include vision language and egocentric video understanding.



Hehe Fan received the M.S. degree in Huazhong University of Science and Technology, China, in 2015. He is currently a Ph.D. student in University of Technology Sydney, Australia. His research interests include video classification.

Chenggang Yan received the Ph.D. degree in Computer Science from the Institute of Computing Technology, Chinese Academy of Sciences in 2013. Now he is the Director of Intelligent Information Processing Lab in Hangzhou Dianzi Univeristy. His research interests include intelligent information processing, machine learning, image processing, computational biology and computational photography.

Mingliang Xu received the Ph.D. degree from the State Key Laboratory of CAD & CG, Zhejiang University, China, in 2012. He is currently a professor with the School of Information Engineering, Zhengzhou University. His research interests include computer graphics and computer vision.

Yi Yang received the Ph.D. degree in computer science from Zhejiang University, China, in 2010. He is currently a professor with University of Technology Sydney, Australia. His current research interest includes machine learning and its applications to multimedia content analysis and computer vision.