Multiscale Methods for Optical Remote-Sensing Image Captioning

Xiaofeng Ma, Rui Zhao, and Zhenwei Shi, *Member, IEEE*

***Abstract*— Recently, the optical remote-sensing image-captioning task has gradually become a research hotspot because of its application prospects in the military and civil fields. Many different methods along with data sets have been proposed. Among them, the models following the encoder–decoder framework have better performance in many aspects like generating more accurate and flexible sentences. However, almost all these methods are of a single fixed receptive field and could not put enough attention on grabbing the multiscale information, which leads to incomplete image representation. In this letter, we deal with the multiscale problem and propose two multiscale methods named multiscale attention (MSA) method and multifeat attention (MFA) method, to obtain better representations for the captioning task in the remote-sensing field. The MSA method extracts features from different layers and uses the multihead attention mechanism to obtain the context feature, respectively. The MFA method combines the target-level features and the scene-level features by using the target-detection task as the auxiliary task to enrich the context feature. The experimental results demonstrate that both of them perform better with regard to the metrics like BLEU, METEOR, ROUGE_L, and CIDEr than the benchmark method.**

***Index Terms*— Remote-sensing image captioning, multiscale, auxiliary task, attention.**

I. Introduction

OPTICAL remote-sensing image captioning is a technology to generate one or more sentences that can describe the contents of the given image accurately and concisely. It is not only an exploration of new processing methods of remote-sensing images but also an effective attempt to help satellites see and tell like a real “clairvoyance.” Furthermore, it can also be applied in many practical fields like mass data retrieval, automatic military intelligence generation, and assisted image interpretation. Deep learning methods are used to explore more excellent models, and many achievements have been made in these years.

According to the way that the sentences are generated, the works can be divided into two categories, namely, template-based methods and encoder–decoder-based methods. For the template-based methods, Shi and Zou [1] proposed a fully convolutional network (FCN) model to do captioning, which mainly focuses on the multilevel semantics and semantic ambiguity problems. They use the detection method to get the information of the objects from different levels and use them to fill the sentence templates that are predesigned with multiple forms. For the encoder–decoder-based methods, Lu *et al.* [2] proposed several different kinds of models using different convolutional neural networks (CNNs) to extract the image features to generate sentences using the recurrent neural network (RNN) or the long short-term memory (LSTM). They also published a public data set in this letter named the remote-sensing image captioning data set (RSICD) and did sufficient experiments on it. From their work, we can conclude that the methods used in the natural image-captioning field can be transferred to the remote-sensing image-captioning field, but they can only obtain acceptable descriptions. Other works like the method proposed in [3], which tries to obtain five sentences at the same time and can get more accurate and diverse results, and method named visual aligning attention (VAA) proposed in [4] try to improve the attention masks’ ability to focus on the regions of interest in the input images. These methods are also important works but are not very relevant to the contents of our letter, so we will not elaborate them.

Recently, Zhang *et al.* [5] proposed a multiscale cropping mechanism for training the remote-sensing captioning models, which can extract the advanced semantic features. The cropping mechanism is one of the training tricks popularly used in deep-learning-based image-processing tasks as a data-augmentation method and can help alleviate the overfitting problem. However, the multiscale problem caused by scale diversity is not completely solved and will still limit the performance of the captioning model. Scale diversity is an inherent property of the images caused by the different distances between the camera and the imaging objects and the scale difference between the images objects. For the remote-sensing images, due to the large imaging range of the satellite cameras, the scale difference between the objects like plane and airport, boat and harbor, and people and beach is huge. In addition, for the captioning task, to increase the diversity and hierarchy of the description sentences, the images of the same resolution will be scaled differently. Therefore, the multiscale methods are needed to help achieve better captioning models in the remote-sensing field.

It is known that the pyramid method is a widely used strategy to solve the scale-diversity problem. The pyramid-based methods can be divided into two categories: image pyramid method [6] and feature pyramid method [7]. The

Manuscript received March 15, 2020; revised June 4, 2020; accepted July 11, 2020. Date of publication July 29, 2020; date of current version October 26, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFC1405605, in part by the National Natural Science Foundation of China under Grant 61671037, in part by the Beijing Natural Science Foundation under Grant 4192034, and in part by the National Defense Science and Technology Innovation Special Zone Project. *(Corresponding author: Zhenwei Shi.)*

The authors are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, also with the Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and also with the State Key Laboratory of Virtual Reality Technology and Systems, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: max15@buaa.edu.cn; ruizhaoipc@buaa.edu.cn; shizhenwei@buaa.edu.cn).

Color versions of one or more of the figures in this letter are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2020.3009243

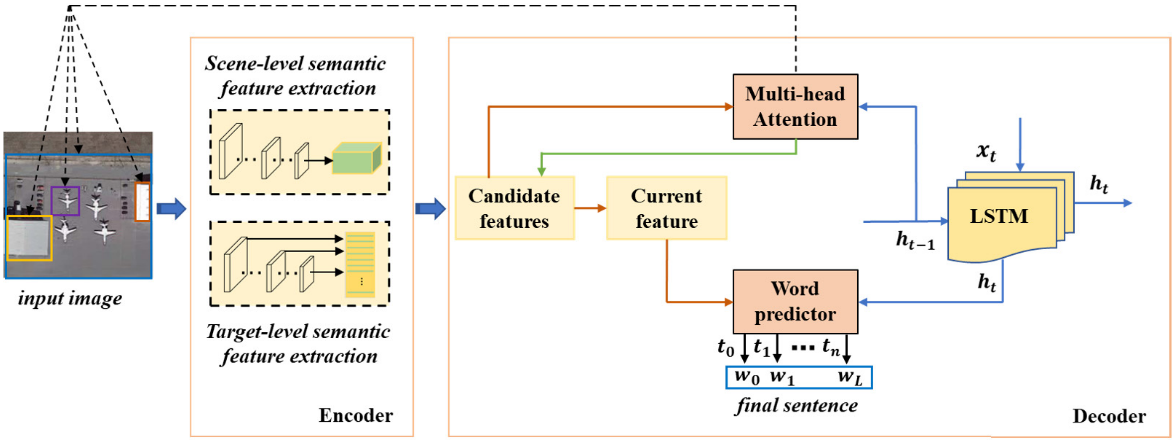


Fig. 1. Model structure of the proposed methods.

image pyramid method is compute and memory-intensive and is avoided in the recent related research. The feature pyramid method using feature fusion is efficient and appears in many models that have the best performance in the corresponding field. In this letter, we propose two multiscale methods with regard to the scale-diversity problem based on the pyramid skills. Our work mainly has the following two contributions.

1) Two multiscale methods are proposed to achieve better representations of the input image and can alleviate the scale-diversity problem to some extent.
2) An improved attention mechanism, multihead attention, is proposed, which can adaptively cascade the features from different levels to get the exact representation of the input image, leading to more accurate captioning results.

II. Methodology

A. Overall Structure of Multiscale Method

The overall structure of our proposed method is shown in Fig. 1. The structure mainly involves two parts.

1) Encoder that consists of two modules called scene-level feature extraction module and target-level feature extraction module, respectively.
2) Decoder that consists of three modules named multihead attention module, LSTM module, and word predictor, respectively. Among them, the scene-level feature-extraction module, the LSTM module, and the word predictor are consistent with the structure of the classical method in [8] and we will just give a brief introduction to them afterward in this section. As for the target-level feature-extraction module and the multihead attention module, which are the new components appearing in the encoder–decoder framework, we will elaborate them in the following subsections.

In our model, the scene-level feature-extraction module uses a deep residual network [9] (ResNet-50) of which the fully connected layers are removed. It is a basic CNN backbone extracting the image features block by block. The LSTM module consists of two hidden layers, and the word predictor is designed with two dense layers and a softmax layer. These two modules are used to abstract contextual information and predict the following word at each moment.

B. Target-Level Feature-Extraction Module

Besides the typical scene-level feature-extraction module, we design a target-level feature-extraction module to get more fine-grained semantic representations for the input optical remote-sensing images. As shown in Fig. 1, the feature vectors output by these two modules will be taken as the input of the decoder together. Unlike the scene-level feature-extraction module of which the output vectors are spatially adjacent, the feature vectors of the target-level feature-extraction module are sparsely distributed and they will be formed as the vector list instead of a feature cube. The design details are illustrated in Fig. 2. We take the SSD-512 [10] framework based on VGG-16 as the basic backbone of the module and add a target-location mask-prediction task that is proposed in our previous paper [11]. The parameters of the module will be obtained by taking the optical remote-sensing image object-detection task as an auxiliary task. The output feature vectors of the target-level feature-extraction module have 21 dimensions that are the logits of each detection block.

In the training phase, there are three components in the loss function, which are named localization loss (loc), mask loss (mask), and confidence loss (conf). The total loss (total) can be formulated as follows:

$$L_{\text{total}} = \frac{1}{N}(L_{\text{conf}} + \alpha L_{\text{loc}}) + \beta L_{\text{mask}} \tag{1}$$

where N is the number of matched default boxes. If $N = 0$, we set the loss to 0. In addition, we set α and β to be 1.0 and 0.25, which is the same as that in [11].

C. Multihead Attention Module

The multihead attention module take different feature lists from multiple scales as input, and each attention unit will do the same work and output the weights of the vectors in each feature list. The module structure is illustrated in Fig. 3.

At each time step t, based on the feature list F_i from the multiscale image features F_{all} of the encoder output and the

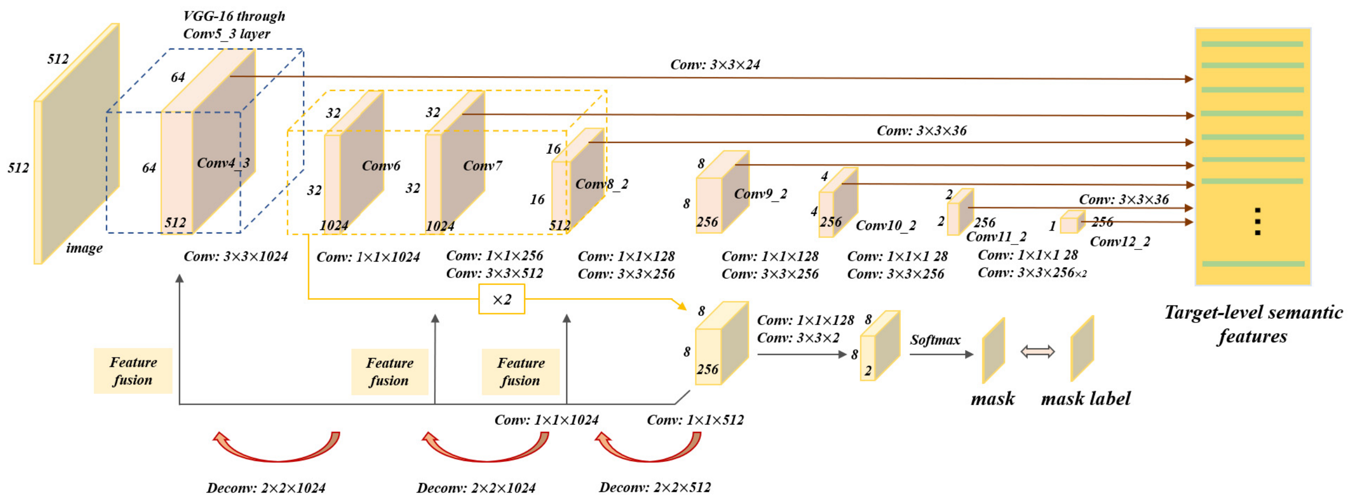


Fig. 2. Structure and parameter settings of the target-level semantic feature-extraction module.

previous hidden state h_{t-1} of the decoder output, the attention network learns to generate the attention weight

$$\boldsymbol{\alpha}_t = [\alpha_{t,i,1}, \ldots, \alpha_{t,i,j}, \ldots, \alpha_{t,k,N}] \tag{2}$$

where $\alpha_{t,i,j}$ is the attention weight corresponding to the feature j in F_i. The feature list is defined as the features $(H_{\text{scene},i} \times W_{\text{scene},i} \times D_{\text{scene},i})$ from the same layer in the scene-level feature-extraction backbone or the features $(N_{\text{target}} \times D_{\text{target}})$ from the target-level feature-extraction module. The calculation process for $\alpha_{t,i,j}$ is as follows:

$$att_{t,i} = f_m(F_{t,i}, h_{t,i}) \tag{3}$$

$$\alpha_{t,i,j} = \frac{e^{att_{t,i,j}}}{\sum_{j=1}^{N_i} e^{att_{t,i,j}}} \tag{4}$$

where f_m is the multilayer perceptron (MLP), N_i is the number of elements in $F_{\mathbf{i}}$, and $\alpha_{t,i}$ is the weights corresponding to each element.

Based on the multihead attention module, the context vector at time step t can be formulated as

$$\boldsymbol{context_t} = \text{concat}(F'_{t,1}, \ldots, F'_{t,i}, \ldots, h_t) \tag{5}$$

where $\boldsymbol{context_t}$ is the context feature used to predict the word at time step t.

D. MSA and MFA

For the methods multiscale attention (MSA) and multifeat attention (MFA) proposed by us, the difference between them is the selection of feature lists. The method MSA only uses the feature lists with different scales output by the scene-level feature-extraction module. The feature lists are spatially consistent, and it can be interpreted as the optimized spatial attention mechanism. The method MFA uses the last-layer feature list output by the scene-level feature-extraction module and the feature list output by the target-level feature-extraction module—the latter can obtain more accurate semantic information by using the optical remote-sensing image target-detection task for auxiliary training. Both of them predict the word w_t at time step t using the word-predictor module. The process can be represented as follows:

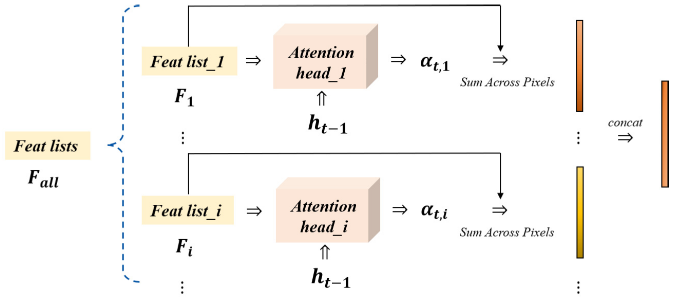


Fig. 3. Schematic of the multihead attention mechanism.

$$\boldsymbol{logits} = W_{d2}(W_{d1}\boldsymbol{context_t} + b_{d1}) + b_{d2} \tag{6}$$

where W_{d1},b_{d1},W_{d2}, and b_{d2} are the parameters in the word-predictor module. The prediction at time step t is

$$P(w_t|I, w_0, w_1, \ldots, w_{t-1}) = \text{Softmax}(\boldsymbol{logits}). \tag{7}$$

The loss is denoted as

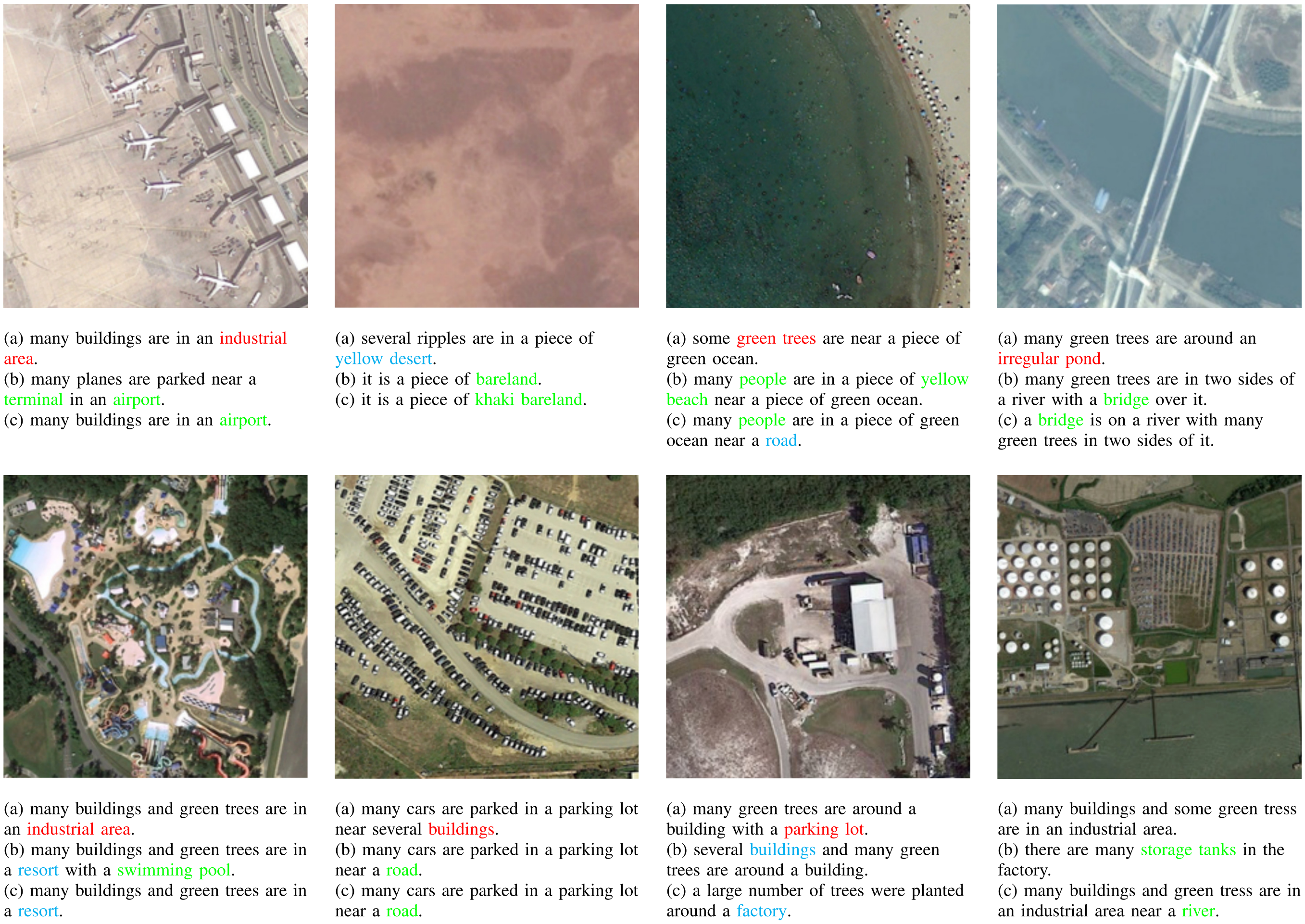
$$\text{Loss} = \frac{1}{L}\sum_{l=0}^{L} log(w_l|I, w_0, w_1, \ldots, w_{l-1}) \tag{8}$$

where I is the input image and L is the length of the sentence.

III. Experiments

A. Data Set and Metrics

1) Data Set: We use RSICD as the main data set for the experiments, which is constructed by Lu *et al.* [2]. It contains a total of 10 921 remote-sensing images, of which the training set contains 8004 images, and the validation set and the test set contain 2187 images. These images are fixed into 224 × 224 pixels, and the resolution of them are various. Each image is labeled with one to five sentences, and there are 24 333 different label sentences in the label file altogether including 3323 words. In addition, the UCM-captions and Sydney-captions are used to evaluate our methods too.



(a) many buildings are in an industrial area.
(b) many planes are parked near a terminal in an airport.
(c) many buildings are in an airport.

(a) several ripples are in a piece of yellow desert.
(b) it is a piece of bareland.
(c) it is a piece of khaki bareland.

(a) some green trees are near a piece of green ocean.
(b) many people are in a piece of yellow beach near a piece of green ocean.
(c) many people are in a piece of green ocean near a road.

(a) many green trees are around an irregular pond.
(b) many green trees are in two sides of a river with a bridge over it.
(c) a bridge is on a river with many green trees in two sides of it.

(a) many buildings and green trees are in an industrial area.
(b) many buildings and green trees are in a resort with a swimming pool.
(c) many buildings and green trees are in a resort.

(a) many cars are parked in a parking lot near several buildings.
(b) many cars are parked in a parking lot near a road.
(c) many cars are parked in a parking lot near a road.

(a) many green trees are around a building with a parking lot.
(b) several buildings and many green trees are around a building.
(c) a large number of trees were planted around a factory.

(a) many buildings and some green tress are in an industrial area.
(b) there are many storage tanks in the factory.
(c) many buildings and green tress are in an industrial area near a river.

Fig. 4. Results output by (a) benchmark method, (b) MSA method, and (c) MFA method from different categories. Red indicates wrong descriptions, cyan indicates less accurate descriptions, and green indicates the more accurate descriptions of the input images.

2) Metrics: Most of the existing evaluation methods for the image sentence-description generation tasks are from the field of machine translation. The evaluation methods used in our letter are BLEU [12], ROUGE [13], METEOR [14], and CIDEr [15]. These evaluation indicators are mainly used to evaluate the similarity between the sentence generated by the model and the labeled sentence. BLEU measures the cooccurrences of n-grams between the generated and reference captions, where n-gram is a set of n ordered words. The n is taken from 1 to 4, corresponding to BLEU-1 to BLEU-4. The higher the evaluation index score, the better the quality of the generated sentence.

TABLE I
RESULTS OF UCM-CAPTIONS AND SYDNEY-CAPTIONS OF MSA, WHERE THE METRICS IN BOLD ARE THE BEST

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE _L	CIDEr
	Results on UCM-captins						
benchmark	0.8321	0.7678	0.7109	0.6602	0.4293	0.7763	3.1478
MSA	**0.8337**	**0.7822**	**0.7406**	**0.7021**	**0.4504**	**0.7918**	**3.2571**
	Results on Sydney-captions						
benchmark	0.7305	0.6437	0.5667	0.5280	0.3650	0.6979	2.1521
MSA	**0.7507**	**0.6800**	**0.6147**	**0.5565**	**0.3674**	**0.7019**	**2.2433**

B. Experimental Setting and Training Details

The deep-learning framework Tensorflow is used to implement our networks. The subnetwork of the scene-level feature-extraction module in the encoder uses the pretraining parameters of ResNet_50 [9] on the ImageNet for initialization, and other parameters are initialized randomly. We use the Adam Optimizer with a learning rate of 1e−4, and it is decreased by a factor of 0.9 after 1e+5 steps. The number of training epochs is set as 80.

The parameters in the target-level feature-extraction module are trained using the auxiliary training data set named DIOR that is proposed in [16]. The DIOR consists 23 463 optimal remote-sensing images and 192 472 object instances, covered by 20 common object categories. The size of the images in the data set is 800 × 800 pixels, and the spatial resolutions range from 0.5 to 30 m. We use the DIOR as our auxiliary training data set, because the images in it have a large range of object size variations and the number of categories is the largest up to now in the remote-sensing field for target detection. Therefore, the target-level feature-extraction module can obtain the target-level semantic features of different scales, which helps achieve better representations of the input image. The training settings of it are the same as that in [11].

TABLE II

RESULTS OF RSICD OF DIFFERENT METHODS, WHERE THE METRICS IN BOLD ARE THE BEST

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
benchmark	0.6644	0.5344	0.4432	0.3737	0.2851	0.5541	1.5753
MSA	**0.6869**	**0.5527**	**0.4600**	**0.3921**	0.3007	**0.5661**	**1.6676**
MFA	0.6802	0.5366	0.4395	0.3684	**0.3028**	0.5516	1.5600

C. Comparison Experiments

We compared our methods with the benchmark method that uses the 50-layer deep residual network (ResNet-50) without the last fully connected layer as the encoder and the LSTM as the decoder based on the attention mechanism proposed in [8]. The features in the benchmark method are extracted by the last convolutional layer. All the experiment settings are the same as the MSA method and the MFA method. The results of these three methods are shown in Tables I and II. For the multiscale cropping method proposed in [5], the authors did not use the attention mechanism in their method and the results in their paper indicate that the method performs not very well even on small data sets. Therefore, we would not compare our methods with their method as they are also an effective way to solve the multiscale problem.

D. Results Analysis

According to Tables I and II, it can be seen that the method MSA achieves significant improvement over the benchmark method in all the evaluations and the method MFA achieves improvement over the benchmark in the B-1, B-2, and METEOR metrics but obtains no improvement in the B-3, B-4, and ROUGE_L, CIDEr metrics. For more direct analysis, we list some representative images from different categories along with the result sentences generated by the benchmark method and our proposed methods in Fig. 4. Each row beginning with Fig. 4(a) below the subfigures is the result of the benchmark method, and Fig. 4(b) and (c) are the results of the MSA and MFA methods. The red words are the wrong descriptions of the input images, the cyan words are the less accurate descriptions of the input images, and the green words are the more accurate descriptions of the input images. From the result sentences, we can see that our methods improve the model's performance mainly in two aspects. One is that the model recognizes the scene category much better than the base model like "airport" in subfigure 1, "bareland" in subfigure 2, and so on. The other is that the model can gain more semantic information than the base model such as "road" and "river" in subfigures 6 and 8. However, there are still some less accurate descriptions like "road" in subfigure 2, where it should be "beach" and "resort" in subfigure 5, where it should be "park,"

Furthermore, we know that the auxiliary data set used in MFA has the distribution bias with the data set RSICD and the target categories labeled in DIOR are far less than the target categories that appear in RSICD. These facts will affect the performance of the MFA method. The data set with target labels, caption labels, and other labels should be conducted in the remote-sensing fields, and that is the work we will do next.

IV. CONCLUSION

In this letter, two methods are proposed against the multiscale problem in the optical remote-sensing image-captioning task. The target-level feature extraction module and the multihead attention module are added to get better representations of the input image. The experiments show that our methods perform much better than the methods that only use features from the last layer of the encoder. Considering the lack of target labels, the auxiliary task and auxiliary data set are used to help achieve satisfactory results. However, the distribution bias between DIOR and RSICD still limits the model's performance. The comprehensive and unified data set needs to be conducted in future, which can lead to efficient solutions and elegant networks.

REFERENCES

[1] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?" *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3623–3634, Jun. 2017.

[2] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.

[3] B. Wang, X. Lu, X. Zheng, and X. Li, "Semantic descriptions of high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1274–1278, Aug. 2019.

[4] Z. Zhang, W. Zhang, W. Diao, M. Yan, X. Gao, and X. Sun, "VAA: Visual aligning attention model for remote sensing image captioning," *IEEE Access*, vol. 7, pp. 137355–137364, 2019.

[5] X. Zhang, Q. Wang, S. Chen, and X. Li, "Multi-scale cropping mechanism for remote sensing image captioning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 10039–10042.

[6] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. ICLR*, 2014, pp. 1–16.

[7] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2016, *arXiv:1612.03144*. [Online]. Available: https://arxiv.org/abs/1612.03144

[8] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," *CoRR*, vol. abs/1502.03044, 2015. [Online]. Available: http://arxiv.org/abs/1502.03044

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[10] W. Liu *et al.*, "SSD: Single shot MultiBox detector," *CoRR*, vol. abs/1512.02325, 2015.[Online]. Available: http://arxiv.org/abs/1512.02325

[11] X. Ma, W. Li, and Z. Shi, "Attention-based convolutional networks for ship detection in high-resolution remote sensing images," in *Pattern Recognition and Computer Vision*, J.-H. Lai, C.-L. Liu, X. Chen, J. Zhou, T. Tan, N. Zheng, and H. Zha, Eds. Cham, Switzerland: Springer, 2018, pp. 373–383.

[12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*. New York, NY, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318, doi: 10.3115/1073083.1073135.

[13] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization Branches Out (WAS)*, 2004, pp. 74–81.

[14] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proc. 2nd Workshop Stat. Mach. Transl. (StatMT)*. New York, NY, USA: Association for Computational Linguistics, Jun. 2007, pp. 228–231.

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

[16] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.