Remote-Sensing Image Captioning Based on Multilayer Aggregated Transformer

Chenyang Liu, Rui Zhao, and Zhenwei Shi, *Member, IEEE*

***Abstract*— Remote-sensing image (RSI) captioning aims to automatically generate sentences describing the content of RSIs. The multiscale information of RSIs contains attributes and complex relationships of objects of different sizes. However, current methods still have some weaknesses in efficiently utilizing multiscale information to generate accurate and detailed sentences. In this letter, we propose a new model based on the "encoder–decoder" framework to address the problem. In the encoder, we fuse the features of different layers in ResNet-50 to extract multiscale information. In the decoder, we propose multilayer aggregated transformer (MLAT) to utilize the extracted information to generate sentences sufficiently. Specially, as the transformer encoding layer goes deeper, the extracted features will be more similar. To sufficiently utilize the features from different transformer encoding layers, compress redundant information, and extract important information, long short-term memory (LSTM) in MLAT aggregates the features to obtain better feature representations. The self-attention mechanism and the aggregation strategy enable MLAT to utilize the features sufficiently. The experimental results show that MLAT as the decoder can help the model address the multiscale problem, significantly improve the model performance on sentence accuracy and diversity, and show that our proposed method performs better than other current methods. Our code is available at https://github.com/Chen-Yang-Liu/MLAT.**

***Index Terms*— Remote-sensing image (RSI) captioning, sentence accuracy, sentence diversity, transformer.**

I. Introduction

REMOTE-sensing image (RSI) captioning is an emerging task in the field of RSI processing. This requires not only visually observing the images, but also describing the images in natural language. It is different from other remote-sensing tasks, such as object detection and image segmentation. It pays attention to the semantic understanding of RSIs from the perspective of human beings. It can be widely applied in image teaching, disaster warning, and image retrieval [1].

At present, there are mainly three kinds of methods in the field: retrieval-based methods, syntax-template-based methods, and encoder–decoder-based methods. Regarding the retrieval-based methods, the models retrieve similar images and use corresponding annotated sentences to generate predicted sentences. To consider five annotated sentences of one image together, Wang *et al.* [2] used five sentences to obtain topic words and proposed a retrieval topic recurrent memory network to utilize the topic words. Besides, Wang *et al.* [3] proposed the collective semantic metering learning framework (CSMLF), in which the image representation and collective sentence representation are embedded into a common semantic space during training. The sentence closest to the image representation is used to generate sentences. Regarding syntax-template-based methods, the models detect the objects in the images, then determine the candidate words, and fill them into the syntax templates. Shi and Zou [4] addressed the multiscale problem by detecting ground elements of three levels with a fully convolutional network (FCN) and then filled them into predesigned templates for captioning. However, the sentences generated by the two kinds of methods are relatively limited and rigid.

The current deep-network methods are based on the "encoder–decoder" framework. These methods can learn RSI representation and grammar and then automatically generate more flexible sentences. These methods generally use convolutional neural networks (CNNs) as the encoder to extract image features and recurrent neural networks (RNNs) as the decoder to generate descriptive sentences. Qu *et al.* [5] combined different CNNs with RNNs and published two datasets for RSI captioning. Lu *et al.* [6] compared the effects of features extracted by traditional methods and CNN-based methods and published a large dataset named RSICD. Ma *et al.* [7] proposed multiscale attention (MSA) and multifeats attention (MFA) to grab multiscale information in RSIs. MSA used multihead attention, and MFA chose object detection as an auxiliary task. To exploit structured spatial relations of semantic contents, Zhao *et al.* [8] proposed structured attention and exploited pixel-level regional image segmentation information. Lu *et al.* [9] proposed a sound-guided dataset and a sound active attention framework, in which the gate recurrent units (GRUs) capture sound information and guide sentence generation. Wang *et al.* [10] proposed a two-stage word-sentence framework, in which the word extractor extracts the valuable words and the sentence generator forms these words into a grammatically correct sentence. These methods promote the development of RSI captioning from different aspects. However, there are still defects in addressing the multiscale problem of RSI well.

The multiscale information of RSIs contains attributes and complex relationships of objects of different sizes [4]. The multiscale problem refers that it is challenging to sufficiently

Manuscript received December 27, 2021; accepted February 8, 2022. Date of publication February 10, 2022; date of current version March 2, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFC1510905, in part by the National Natural Science Foundation of China under Grant 62125102, and in part by the Beijing Natural Science Foundation under Grant 4192034. *(Corresponding author: Zhenwei Shi.)*

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

Digital Object Identifier 10.1109/LGRS.2022.3150957

extract and utilize the multiscale information in RSIs for captioning. There is so much information in RSIs that it is challenging to describe objects of different sizes. Some methods have addressed the multiscale problem, such as [4] and [7] introduced above. However, these methods focus more on extracting multiscale object information in the RSIs and focus less on sufficiently utilizing the extracted multiscale features and complex object relationships. Recently, transformer-based architectures represent state of the art in sequence modeling tasks [11]. In this letter, we introduce transformer and improve it to address the multiscale problem. Transformer can establish efficient relationships among the features with the self-attention, which is helpful to utilize multiscale information. To further sufficiently utilize the extracted features from different transformer encoding layers, we propose multilayer aggregated transformer (MLAT) as the decoder. Besides, the feature representation ability of the model is very significant [12], [13]. In the encoder, we fuse the features of different layers in ResNet-50 to extract multiscale features.

On a related line, Cornia *et al.* [11] presented a meshed transformer. The model can learn the multilevel representation of the relationships between image regions. The mesh-like connectivity enables the model to exploit low-level and high-level features at the decoding stage. Unlike that, our MLAT chooses long short-term memory (LSTM) as the connectivity between the transformer encoder and transformer decoder to obtain better feature representations containing the valid information from each transformer encoding layer. It enables the model to utilize the multiscale information sufficiently. The experimental results demonstrate that our method performs better than other methods.

Our contributions are summarized as the following.

1) We address multiscale problems from two perspectives: the sufficient extraction and utilization of the multiscale information. We fuse the features of different layers in ResNet-50 to extract multiscale features. We propose MLAT to utilize the extracted multiscale features for captioning sufficiently.
2) LSTM as the connectivity aggregates the features from each transformer encoding layer. The aggregated features are sent to each transformer decoding layer. Compared with the original Transformer, the strategy enables the model to sufficiently utilize the features from all transformer encoding layers.
3) We evaluate the sentence accuracy objectively and evaluate the sentence diversity from two perspectives.

II. Methodology

A. Overall Structure

Our model is based on the "encoder–decoder" framework. The overall structure is shown in Fig. 1. In the encoder, we fuse the features of different layers in ResNet-50 [14] to extract multiscale features. In the decoder, MLAT sufficiently utilizes the extracted multiscale features. MLAT can establish better relationships among objects in RSIs and accurately describe images. The features from deep transformer encoding layers

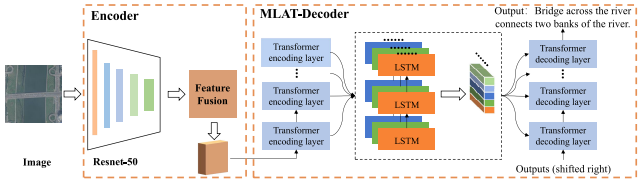

Fig. 1. Structure of our model. In the encoder, we fuse the features of different layers in ResNet-50 [14] to extract multiscale features (see Fig. 2). In the decoder, MLAT is proposed to sufficiently utilize the extracted multiscale information (see Fig. 3). Different transformer encoding layers understand the input information from different levels. LSTM aggregates the features from each encoding layer. Then the aggregated features are sent to each transformer decoding layer for captioning.

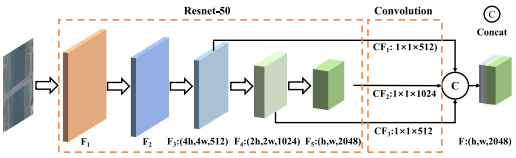

Fig. 2. Feature fusion strategy. $F_1 \sim F_5$ are, respectively, the feature maps of conv1–conv5 in ResNet-50 [14]. CF_1, CF_2, and CF_3 denote different 1×1 convolution kernels, and their output dimensions are, respectively, 512, 512, and 1024. They are used to resize F_3, F_4, and F_5 for subsequent feature fusion.

are similar, and the features from the shallow layers may contain valid information for captioning. Therefore, to sufficiently utilize the features from different encoding layers, we propose LSTM as an aggregator to compress redundant features and extract important features from different encoding layers. This enables the model to utilize the extracted multiscale information in RSIs sufficiently.

B. Feature Fusion Strategy

Different convolutional layers in ResNet-50 can extract features of different scales. In this letter, unlike the previous methods, we focus more on improving the utilization capability rather than the extraction capability of multiscale information. Therefore, we choose a simple fusion strategy to extract the multiscale features in Fig. 2. The lower F_1 and F_2 contain too low-level information, while image captioning requires higher-level semantic information. So, we only perform the feature fusion on F_3, F_4, and F_5.

C. Multilayer Aggregated Transformer

The original transformer is composed of an encoder and the decoder, both of which are stacked by multiple sublayers [15]. As the transformer encoding layer goes deeper, the extracted features will be more similar as each feature will collect information from each other [16], [17]. If we only utilize the features from the top encoding layer, it is actually redundant to send so many similar features to each decoding layer [17]. Besides, the features from the shallow encoding layers contain some valid information for RSI captioning [18]. Therefore, it motivates us to aggregate the features from different encoding layers and compress redundant information to obtain better feature representations. In this letter, we propose a novel multilayer aggregation strategy.

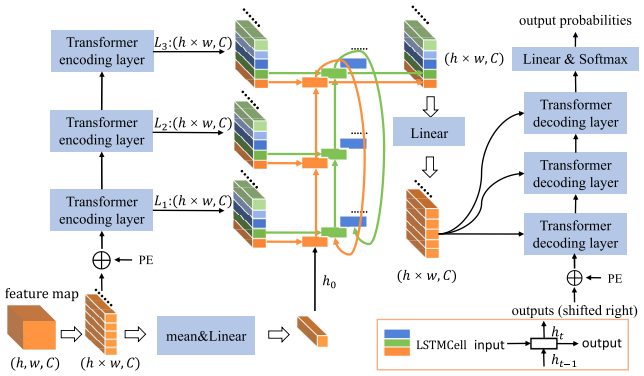

Fig. 3. Structure of MLAT. The encoding layer and decoding layer are the same as in [15]. h, w, and C, respectively, represent the width, height, and channel number of the feature map from the encoder. The feature map is flattened and sent into the transformer encoder after positional encoding (PE) of sine and cosine functions as in [15]. The flattened feature map is transformed to initialize LSTM. Assume that the outputs of N encoding layers are $L_1, L_2, \ldots, L_N$, all tokens of L_i are sent into different LSTMCells to obtain the aggregated features. All LSTMCells of different colors are the same. The aggregated features pass through a linear layer and then are sent to each decoding layer to obtain the probability of each word in the vocabulary. The "outputs (shifted right)" denotes the words that have been output before this moment, and it refers to "⟨start⟩" at the initial moment.

The features extracted by each transformer encoding layer are sequence features. Considering that the gate mechanism of LSTM enables it to learn to forget or remember information when processing sequence data [19]. We propose LSTM as an aggregator to compress redundant features and extract important features from all transformer encoding layers to obtain better feature representations. The structure of MLAT is shown in Fig. 3. MLAT as the decoder can establish relationships among objects with the self-attention mechanism. The features from all transformer encoding layers are sent into LSTM to perform aggregation. That enables MLAT to sufficiently utilize the multiscale features from the encoder for accurately captioning.

Our total loss function contains two components named CrossEntropy loss (L_{Cro}) and multihead attention loss (L_{Att}) in each transformer decoding layer in the training phase. The total loss (L_{Total}) function is as follows:

$$L_{\text{Total}} = L_{\text{Cro}} + L_{\text{Att}} \tag{1}$$

$$L_{\text{Cro}} = -\sum_{l=1}^{L} \log\left(\sum_{k=1}^{K} \widetilde{y}_l^{(k)} p_l^{(k)}\right) \tag{2}$$

where L is the length of the generated sentence. K is the number of words in the vocabulary. $\widetilde{\boldsymbol{y}}_l = [\widetilde{y}_l^{(1)}, \widetilde{y}_l^{(2)}, \ldots, \widetilde{y}_l^{(K)}]$ is the one-hot vector representation of the lth word in the reference sentence. $\boldsymbol{p}_l = [p_l^{(1)}, p_l^{(2)}, \ldots, p_l^{(K)}]$ is the probability vector of the predicted lth word

$$L_{\text{Att}} = \frac{\lambda}{N_{\text{layer}} \times N_{\text{head}}} \sum_{n_1=1}^{N_{\text{layer}}} \sum_{n_2=1}^{N_{\text{head}}} \sum_{i=1}^{h \times w} \left(1 - \sum_{l=1}^{L} \alpha_{l,i,n_2,n_1}\right)^2 \tag{3}$$

where N_{layer} and N_{head} are, respectively, the number of decoding layers and the head number of every layer in MLAT. h and w are as shown in Fig. 3. α_{l,i,n_2,n_1} is the attention state for the n_2th head of the n_1th transformer decoding layer. λ is the weight coefficient of L_{Att} and is set to be 1.0. L_{Att} is added to encourage the model to give equal attention to each token of the aggregated result when decoding.

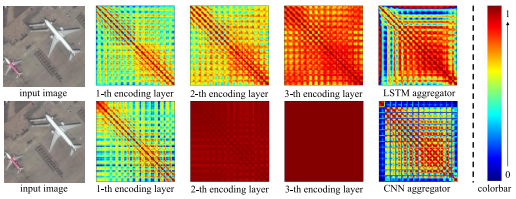

Fig. 4. Visualization of the cosine similarity between each feature from each encoding layer and sequence features from two aggregators. Red denotes higher similarity values, and blue denotes lower values.

III. Experiments

A. Dataset and Metrics

1) Dataset: There are mainly three datasets in the field of RSI captioning: UCM, Sydney, and RSICD. The first two datasets are smaller and contain 2100 and 617 images, respectively [5]. In the experiments, we used the largest dataset in this field: RSICD, which was published by Lu *et al.* [6]. RSICD contains 10 921 RSIs, including 30 common scenes such as airports, mountains, and farmland. Every image is 224×224 pixels and annotated with five sentences. There are 54 605 sentences in total, including 24 333 nonrepeated sentences.

2) Metrics: The model performance evaluation is based on the similarity between candidate sentences and reference sentences. The current automatic evaluation metrics mainly include BLEU, METEOR, ROUGE_L, and CIDEr. These metrics measure the similarity between candidate sentences and reference sentences from different perspectives. The higher the metric scores, the higher the accuracy of the sentences generated by the model.

B. Experimental Setting

The initial learning rate is 0.0001 and decays as training steps increase. We set the maximum epoch to 100, the batch size to 32, and the dimension of the word embedding to 512. Regarding the transformer encoding and decoding layers in MLAT, the layer number is 3, and the head number of multihead attention is 8. We choose the Beam Search strategy instead of the Greedy Search strategy to generate sentences in the experiments. If the beam size is too large, many candidate sentences will be generated for an image. It requires a large amount of computation. If it is too small, the sentence accuracy will decrease. We set the beam size as 5.

C. Experiments and Results

In Table I, all methods are based on Resent-50 as the backbone network for feature extraction. In the baseline, LSTM is chosen as the decoder. Besides, we have tried to choose CNN as an aggregator instead of LSTM. We rearrange the sequence features from each encoding layer into feature maps, concat them together in the channel dimension, and

TABLE I

RESULTS ON SENTENCE ACCURACY OF DIFFERENT METHODS ON RSICD, WHERE THE BOLDED RESULTS ARE THE BEST, TR DENOTES TRANSFORMER, AND FUS DENOTES FEATURE FUSION STRATEGY, AND AGGR-CNN DENOTES CNN-BASED AGGREGATOR

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
Baseline	63.47	48.13	38.47	31.48	25.83	49.23	87.40
Tr	65.08	49.23	39.16	32.00	26.50	49.55	91.00
Fus + Tr	66.24	50.25	39.99	32.93	26.79	49.78	89.81
Aggr-CNN	65.56	49.56	39.55	32.68	26.82	49.48	89.47
MLAT	66.51	50.53	40.68	33.93	27.22	**50.80**	**94.73**
Fus + MLAT	**66.90**	**51.13**	**41.14**	**34.21**	**27.31**	50.57	94.27

TABLE II

COMPARISONS WITH OTHER CURRENT METHODS ON RSICD

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
VLAD + RNN [6]	49.38	30.91	22.09	16.77	19.96	42.42	103.92
VLAD + LSTM [6]	50.04	31.95	23.19	17.78	20.46	43.34	118.01
mRNN [5]	45.58	28.25	18.09	12.13	15.69	31.26	19.15
mLSTM [5]	50.57	32.42	23.19	17.46	17.84	35.02	31.61
mGRU [20]	42.56	29.99	22.91	17.98	19.41	37.97	124.82
CSMLF [3]	57.59	38.59	28.32	22.17	21.28	44.55	52.97
sound-f-a [9]	59.35	45.11	35.29	28.08	26.11	49.57	**132.35**
Soft-attention [22]	65.13	49.04	39.00	32.20	26.39	49.69	90.58
Our model	**66.90**	**51.13**	**41.14**	**34.21**	**27.31**	**50.57**	94.27

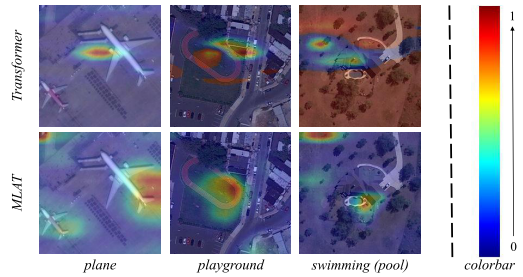

Fig. 5. Visualization of the multihead attention states when the transformer and MLAT are, respectively, chosen as decoders to generate the words corresponding to the multiscale objects. Red denotes higher similarity values, and blue denotes lower values. MLAT can focus on the big and small planes when describing the planes in the first image. Besides, MLAT performs better than transformer when describing the big playground and the small swimming pool regarding the second and third images. These examples show that MLAT can ease the multiscale problem.

send them to a CNN-based aggregator, which is composed of a one-layer CNN. In Table II, we compare our method and other methods regarding sentence accuracy. mRNN [5], mLSTM [5], and mGRU [20] are based on VGG-16 [21] as the encoder and, respectively, choose RNN, LSTM, and GRU as the decoders. We implement the Soft-attention method [22] based on ResNet-50 and LSTM. Furthermore, we objectively evaluate the sentence diversity from two perspectives.

1) Sentence Accuracy: According to Table I, the feature fusion strategy, Aggr-CNN, and MLAT can all improve the model performance because they enable the model to sufficiently extract and utilize multiscale information in RSIs. MLAT and Aggr-CNN can perform better than LSTM and transformer, which shows that the aggregation is effective. Besides, MLAT performs better than Aggr-CNN. We argue that the gate mechanism of LSTM enables it to learn to remember or forget information selectively. It is helpful to compress redundant features and extract important features from all encoding layers. Our final model chooses feature fusion strategy and MLAT. According to Table II, our model performs better than other methods on sentence accuracy.

2) Influence of MLAT: Fig. 4 shows that as the transformer encoding layer goes deeper, the similarity between each feature will increase. Besides, the similarity between the aggregated features has decreased compared to the similarity between the features from the top layer. This shows that the redundant features have been reduced. That is the effect we expect. According to Fig. 5, MLAT can focus on more accurate features than the transformer and focus on multiscale objects and scenes in the RSIs when generating corresponding words. Besides, according to Fig. 6, our model can accurately describe many multiscale objects and correct object relationships in the RSIs. Therefore, MLAT is helpful for the model to address the multiscale problem so that the model can generate more accurate and detailed sentences. Besides, we calculate the number of model parameters and the inference speed (seconds per image) on an NVIDIA GTX 1080 graphics card. The transformer and MLAT are, respectively, 105.22 and 149.77 M regarding the parameters, and 0.36 and 1.34 s regarding the speed. Although the efficiency of MLAT is lower, we think it is acceptable considering the accuracy improvement. We will further improve the model operation efficiency in the future.

3) Sentence Diversity: At present, the evaluation of sentence diversity is not as perfect and objective as the accuracy. In this letter, we propose that the diversity could be evaluated objectively from two aspects: the diversity among sentences of different images and the diversity among k sentences of the same image. The former can be understood as how many different sentences a person can express. The latter can be understood as how many perspectives a person can describe an image. Regarding the former, each model generates the best-1 sentence and the best-3 sentences for each image. Then we calculate the ratio of all unique sentences. Regarding the latter, each model generates the best-3 sentences for each image as shown in Fig. 7. Then we combine three sentences in pairs and calculate BLEU, METEOR, ROUGE_L, CEDIr, and then average them. It should be noted that the lower the scores, the greater the difference among sentences and the

TABLE III

SENTENCE DIVERSITY AMONG DIFFERENT IMAGES, WHERE THE HIGHER THE SCORE, THE BETTER THE DIVERSITY

Method	Best-1	Best-3
Baseline	30.02%	80.51%
Our model	**31.76%**	**87.85%**

TABLE IV

SENTENCE DIVERSITY AMONG BEST-K SENTENCES OF THE SAME IMAGE, WHERE THE LOWER THE SCORE, THE BETTER THE DIVERSITY

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
Baseline	75.95	71.61	67.43	63.12	41.45	75.70	456.92
Our model	**74.91**	**70.19**	**65.91**	**61.59**	**40.54**	**73.74**	**428.99**

(a) Yellow beach and black rocks are between green trees and blue ocean.
(b) Yellow beach and green trees are between blue ocean.
(c) Yellow beach and black rocks are between green trees and blue ocean.

(a) Some buildings and many green trees are near a viaduct shaped three circles.
(b) Some buildings and many green trees are near a viaduct.
(c) Some buildings and many green trees are near a viaduct shaped three circles.

(a) Many orderly buildings and some green trees are in a medium residential area.
(b) On both sides of the road are neatly arranged.
(c) Some buildings and green trees are in a medium residential area.

(a) Some white storage tanks are near several buildings and green trees.
(b) Some white storage tanks are near some green trees.
(c) Several storage tanks in different sizes are near several buildings and green trees.

(a) Many buildings and a playground are in a school.
(b) Many buildings and green trees are around a school.
(c) Many buildings and some green trees are in a school. with a playground.

(a) Many green trees are around a large stadium.
(b) Some green trees are around a stadium.
(c) Some green trees are around a stadium with a football field in it.

Fig. 6. Captioning results. (a) One of the five ground-truth sentences. (b) and (c) Generated by the baseline and our method, respectively. Red words are wrong. Blue words are less accurate. Green words are more accurate. Our method can generate more accurate sentences. For example, for the first image, our model can describe "black rocks" and "ocean" and describe accurate object relationships, such as "between. . . and. . ."

(1) The playground **consists of** a red track and a green football field.
(2) There are many houses **around** the playground.
(3) There are many trees **around** the playground.

(1) A **building** is near several **green trees**.
(2) A **building** is near several pieces of **green meadows**.
(3) Several pieces of **farmlands** are near a **viaduct**.

(1) A playground is surrounded by many buildings.
(2) A playground with two **basketball fields** in it is surrounded by many buildings.
(3) Many buildings are around a playground and **some trees**.

Fig. 7. Sentences generated by our model illustrate the sentence diversity among three sentences for the same image. For example, our model can describe the playground and internal things, or describe the playground and surrounding things for the first image.

better the diversity. Tables III and IV show that our model performs better than the baseline on sentence diversity.

IV. Conclusion

We address the multiscale problem from two perspectives. We fuse the features of different layers in ResNet-50 to extract multiscale information in the images. We propose MLAT as the decoder in which LSTM aggregates the features from different encoding layers to sufficiently utilize the extracted multiscale information. The experimental results show that our strategies are helpful to address the multiscale problem and can improve model performance on sentence accuracy and diversity.

References

[1] X. Lu, X. Zheng, and X. Li, "Latent semantic minimal hashing for image retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 355–368, Jan. 2016.

[2] B. Wang, X. Zheng, B. Qu, and X. Lu, "Retrieval topic recurrent memory network for remote sensing image captioning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 256–270, 2020.

[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. 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.

[5] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *Proc. Int. Conf. Comput., Inf. Telecommun. Syst. (CITS)*, Jul. 2016, pp. 1–5.

[6] 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.

[7] X. Ma, R. Zhao, and Z. Shi, "Multiscale methods for optical remote-sensing image captioning," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 11, pp. 2001–2005, Nov. 2021.

[8] R. Zhao, Z. Shi, and Z. Zou, "High-resolution remote sensing image captioning based on structured attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[9] X. Lu, B. Wang, and X. Zheng, "Sound active attention framework for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1985–2000, Mar. 2020.

[10] Q. Wang, W. Huang, X. Zhang, and X. Li, "Word–sentence framework for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10532–10543, Dec. 2021.

[11] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10578–10587.

[12] X. Sun, B. Wang, Z. Wang, H. Li, H. Li, and K. Fu, "Research progress on few-shot learning for remote sensing image interpretation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2387–2402, 2021.

[13] Q. He, X. Sun, Z. Yan, and K. Fu, "DABNet: Deformable contextual and boundary-weighted network for cloud detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

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

[15] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[16] M. E. Peters *et al.*, "Deep contextualized word representations," 2018, *arXiv:1802.05365*.

[17] M. Zheng *et al.*, "End-to-end object detection with adaptive clustering transformer," 2020, *arXiv:2011.09315*.

[18] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.

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

[20] X. Li, A. Yuan, and X. Lu, "Multi-modal gated recurrent units for image description," *Multimedia Tools Appl.*, vol. 77, no. 22, pp. 29847–29869, 2018.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[22] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.