Denoising Based Multi-Scale Feature Fusion for Remote Sensing Image Captioning

Wei Huang, Qi Wang, *Senior Member, IEEE*, and Xuelong Li, *Fellow, IEEE*

***Abstract*—Benefiting from deep learning technology, it becomes achievable to generate captions for remote sensing images and great progress has been made in the recent years. However, the large scale variation of remote sensing images, which would lead to errors or omissions in feature extraction, still limits the further improvement of caption quality. To address this problem, we propose a denoising based multi-scale feature fusion (DMSFF) mechanism for remote sensing image captioning in this paper. The proposed DMSFF mechanism aggregates multi-scale features with the denoising operation at the stage of visual feature extraction. It can help the encoder-decoder framework, which is widely used in image captioning, to obtain the denoising multi-scale feature representation. In experiments, we apply the proposed DMSFF in the encoder-decoder framework and perform the comparative experiments on two public remote sensing image captioning data sets including UCM-captions and Sydney-captions. The experimental results demonstrate the effectiveness of our method.**

***Index Terms*—remote sensing, image captioning, deep learning, multi-scale, feature fusion, encoder-decoder**

I. INTRODUCTION

WITH the rapid development of remote sensing equipments and technologies, many applications based on remote sensing images have developed greatly, including remote sensing scene classification [1], [2], geographical image retrieval [3] and geographic semantic segmentation [4]. As well as we know, these tasks mostly concentrate on studying the visual attributes such as class labels and object loactions, ignoring their semantic relationship. As a novel and interesting application, remote sensing image captioning [5]–[7] has been explored recently. It aims at generating a comprehensive sentence for a given remote sensing image at the semantic level, and has the promising potential in the cross-modality tasks of remote sensing such as image indexing.

Different from the absolutely visiual tasks, image captioning considers not only the visual attributes but also their text relationship. In order to achieve this goal, it is necessary to make full use of both image understanding techniques and natural language processing techniques. In the past few years, researchers have proposed many methods for natural image captioning [8]–[13]. Mostly, these methods obey a general framework: encoder-decoder architecture. As the name suggests, this architecture can be divided into two sub-models: encoder model and decoder model. Encoder model is used to extract visual feature at the semantic level, while decoder model is used to generate a well-formed sentence based on the extracted feature.

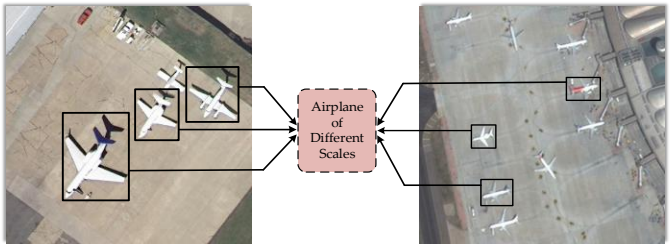

Fig. 1. These two remote sensing images can be described with the same sentence "Some airplanes are parking at an airport". Nevertheless, scale of the most important object $airplane$ varies widely.

Inspired by natural image captioning, researchers have studied the caption generation for remote sensing images. Qu $et\ al.$ [6] firstly transfer the Encoder-Decoder architecture (CNN + RNN/LSTM) from natural image captioning to remote sensing image captioning. Shi $et\ al.$ [14] propose a multilevel convolutional framework for remote sensing image understanding, focusing on improving the accuracy of object recognition. Lu $et\ al.$ [5] utilize the multimodal feature based methods and the deep feature based methods at image encoding stage. Besides, Lu $et\ al.$ also attempt the soft and hard attention model to improve the accuracy of captions. Furthermore, Zhang $et\ al.$ [7] introduce a multi-scale image cropping and training mechanism to achieve data augmentation. In remote sensing image captioning, the classical encoder model includes AlexNet [15], VGG [16], ResNet [17] and others, while the decoder model mainly consists of RNN and LSTM. The whole encoder-decoder framework for remote sensing images is their cross combination.

Although the exciting progress has been made in remote sensing image captioning, the large scale variation of remote sensing images still limits the further improvement of caption quality. As shown in Fig. 1, remote sensing images are collected from aerospace equipments, so they usually cover a large area and contain many types of features and objects, which are at quite different scales. Due to the various scales, some features and objects would be ignored or mis-recognized.

In order to improve the ability of multi-scale feature representation, we propose the denoising based multi-scale feature fusion (DMSFF) mechanism for remote sensing image captioning in this paper. It takes effect by aggregating multiple outputs at different stages of convolutional network, which

W. Huang, Q. Wang and X. Li are with the School of Computer Science, and with the Center for Optical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China (e-mail: hw2hwei@gmail.com, crabwq@nwpu.edu.cn, li@nwpu.edu.cn).This work was supported by the National Natural Science Foundation of China under Grant U1864204, 61773316, U1801262, 61871470, and 61761130079.
Q. Wang is the corresponding author.

contains the features of different scales. As well as we know, the shallower the feature is, the more noise it contains. To reduce the impact of the noise, the denoising operation is specially designed in consideration of spatial location and feature channel. It is worth mentioning that the proposed DMSFF can be embedded into various CNN architectures, which play the role of encoder model in the whole encoder-decoder framework.

In conclusion, our contributions can be summarized as the following three aspects:

(1) To deal with the problem of large scale variation of remote sensing images, we propose a kind of multi-scale feature fusion strategy for remote sensing image captioning in this paper.
(2) In order to reduce the influence of the noise of features from different layers, we design a denoising operation considering not only the spatial location but also the feature channel at different scales.
(3) In order to verify the effectivensess of the proposed method, comparative experiments are conducted on two public remote sensing image captioning data sets. The results demonstrate that our method can help the encoder-decoder framework to get better caption quality.

II. METHOD

In this section, based on encoder model, the denoising based multi-scale feature fusion (DMSFF) mechanism is presented to extract comprehensive visual feature. Following that, the decoder model is introduced to generate a well-formed sentence.

A. Denoising Based Multi-Scale Feature Representation

Image understanding is a process of extracting features and objects from images, which is implemented by the encoder model of encoder-decoder framework. Because of the powerful feature extraion capability, CNNs has become the mainstream of encoder model. It can encode a remote sensing image into a multi-layer feature map.

Usually, only the final feature map is used in the next decoder stage. However, the scale of remote sensing images varies widely. The fixed receptive field of the final feature map cannot deal with the large scale variation. To solve this problem, as shown in Fig 2, we propose a DMSFF mechanism to concatenate multiple denoising features of different scales of CNNs. DMSFF can be roughly diveded into four parts: feature map selection, spatial-wise denoising, channel-wise denoising and multi-scale feature fusion.

*1) **feature map selection**:* CNN is a stacking structure of multiple convolutional blocks which can extract features from shallow to deep. There are many multi-layer feature maps of different scales at different stages of CNN. It is the prerequisite of multi-scale feature fusion to select suitable scales. In this paper, three pieces of multi-layer feature maps are selected from CNN to be fused as the final multi-scale feature representation, which are formulated as:

$$F_1 = conv\_block1(I), \tag{1}$$

$$F_2 = conv\_block2(F_1), \tag{2}$$

$$F_3 = conv\_block3(F_2), \tag{3}$$

here $conv\_block1$, $conv\_block2$ and $conv\_block3$ the three ordered convolutional blocks of the CNN, and their ordered combination is the whole feature extractor of a CNN model. $F_1 \in \mathbb{R}^{H_1 \times W_1 \times C_1}$, $F_2 \in \mathbb{R}^{H_2 \times W_2 \times C_2}$ and $F_3 \in \mathbb{R}^{H_3 \times W_3 \times C_3}$ are the three selected multi-layer feature maps of different scales, where $H \times W$ is the spatial size and C is the number of channels. The detailed settings of the feature maps based on different CNN architectures are provided in the Sec. III.

It is well known that the multi-layer feature map from the shallow layer would contain more noise. It is necessary to denoise the selected feature map. The feature map can be considered from spatial location ($H \times W$) and feature channel (C). The noise is suppressed by the learnable weighting strategy.

*2) **spatial-wise denoising**:* We first reduce the noise of feature map from spatial locaiton. For each multi-layer feature map F, the spatial-wise weighting matrix $\mathcal{W}_s \in \mathbb{R}^{H \times W}$ is calculated by:

$$\mathcal{W}_s(i,j) = f_{s_2}(f_{s_1}(F(i,j))), \quad i \in H, j \in W \tag{4}$$

here f_s is the combination of one fully-connected (FC) layer and one non-linear activation function. Refering to [18], f_{s_1} reduces the dimension of F from C to $C/16$ followed by the function of ReLU, and f_{s_2} shrunk $C/16$ to 1 followed by the function of Sigmoid. These two FC layers with non-linear activation functions can learn the importance of the feature vector at the position of (i,j).

The learned weighting matrix of $\mathcal{W}_s$ is used to weight the feature map as:

$$F'(i,j,k) = \mathcal{W}_s(i,j) * F(i,j,k), \quad i \in H, j \in W, k \in C \tag{5}$$

all of the elements of F' are weighted according to the importance degree at the position of (i,j).

*3) **channel-wise denoising**:* Then the feature map is further denoised along the feature channel. Since the feature at different positions in the same channel represents the same semantic information, the weighting operation works in uints of channel. Therefore, the channele-wise weighting matrix is actually a vector of 1 x 1 x C, denoted as $\mathcal{W}_c \in \mathbb{R}^C$. Motivated by [18], the k-th element of $\mathcal{W}_c$ is calulated by:

$$\mathcal{W}_c(k) = \frac{1}{H \times W} \sum_{i=0}^{H} \sum_{j=0}^{W} F_c(i,j). \tag{6}$$

To explore the cross-channel dependence, we use two FC layers f_{c_1} and f_{c_2} to transform the weighting vector $\mathcal{W}_c$. f_{c_1}, followed by a non-linear activation function of RELU, reduces the dimension of $\mathcal{W}_c$ from C to $C/16$, while f_{c_2} restores the dimension from $C/16$ to C. The transformed weighting matrix is denoted as $\bar{\mathcal{W}}_c \in \mathbb{R}^C$. The relationship $\bar{\mathcal{W}}_c$ and $\mathcal{W}_c$ is represented as:

$$\bar{\mathcal{W}}_c = f_{c_2}(f_{c_1}(\mathcal{W}_c)). \tag{7}$$

.

The feature map F' is weighted again by dot product operation along the channel, which is calculated by:

$$F''(i,j,k) = \bar{\mathcal{W}}_c(k) * F'(i,j,k), \quad i \in H, j \in W, k \in C. \quad (8)$$

*4) **multi-scale feature fusion**:* After Eqn. (4)-(8), we can obtain the denoised feature map of F''. In this paper, only the feature vector is used for the image generation. Therefore F'' is intergrated into a global feature vector $I \in \mathbb{R}^C$ by the operation of global average pooling. Element in the k-th channel of I is calculated by:

$$I(k) = \frac{1}{H \times W} \sum_{i=0}^{H} \sum_{j=0}^{W} F''(i,j), \quad i \in H, j \in W, k \in C \quad (9)$$

For F_1, F_2 and F_3, there are the corresponding global feature vectors of I_1, I_2 and I_3, which represents the features of different scales. They are fused as:

$$I_{cat} = concat(I_1, I_2, I_3), \quad (10)$$

$$I_f = f_{cat}(I_{cat}), \quad (11)$$

here $I_{cat} \in \mathbb{R}^{(C_1+C_2+C_3)}$ is the concatenation of I_1, I_2 and I_3. For fair comparison, I_{cat} is transformed into $I_f \in \mathbb{R}^{C_3}$ by a FC layer of f_{cat}, which is actually a dimension reduction operation. The comparative experiments in this paper are conducted between I_3 and I_f which have the same dimension.

B. LSTM-based Caption Gengration

The demand of caption generation is to generate a well-formed sentence based on the feature vector of I_f/I_3. The sentence is the summary of features, objects and their relationship in the given remote sensing image. It is a sequential problem, so the most popular sequential neural network of Long Short-Term Memory network (LSTM) [19] is selected to achieve this goal in this paper. In this sub-section, we introduce the LSTM-based caption generation and the corresponding loss function used to optimize the whole encoder-decoder framework.

*1) **LSTM**:* LSTM is a chain of repeating the same module of neural network. The core of LSTM is three gates and a memory cell. The three gates of forget gate, input gate and output gate control the LSTM cell state. Each gate can optionally let information through. The input gate determinds whether to pass the new input to memory cell, and the forget gate decides if to forget the current value of cell memory, and the output gate decides the current output of LSTM according to the current input and cell state. The update of three gates and cell memory at timestep t are as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i), \quad (12)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f), \quad (13)$$

$$\tilde{c_t} = tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_i), \quad (14)$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c_t}, \quad (15)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o), \quad (16)$$

$$h_t = o_t * tanh(c_t), \quad (17)$$

where i_t, f_t, c_t, o_t and h_t represent the values of input gate, forget gate, cell memory, output gate and hidden stage at time t, respectively. And there are two types of non-linear activation function of sigmoid σ and hyperbolic tangent $tanh$. All the weighting matrices W and bias b are the trainable parameters. At time t, x_t is the LSTM's input while the hidden state of h_t is used as the output.

*2) **LSTM-based caption generation**:* The LSTM takes the concatenation of feature vector (the feature vector is I_f when applying DMSFF, and is I_3 when using CNN baselines) and the previous word w_{t-1} as the input x_t. And it generate the word at the tims step t of y_t. They are formulated as follows:

$$x_t = concat(y_{t-1}, I), \quad (18)$$

$$h_t = LSTM(x_t), \quad (19)$$

$$y_t = f_{yw}(h_t), \quad (20)$$

$$y = \{y_1, y_2, ..., y_N\}, y_i \in \mathbb{R}^K \quad (21)$$

$$r = \{r_1, r_2, ..., r_N\}, r_i \in \mathbb{R}^K \quad (22)$$

here f_{yw} is a FC layer which embeds h_t into the word space. N is the length of the caption and K is the vocabulary size. In this paper, N is no bigger than 25. y_t and r_t are the predicted word and reference word at time t, respectively. r is the ground truth of a remote sensing image.

The initialization of LSTM is as follows:

$$h_0 = f_{init\_h}(I), \quad (23)$$

$$c_0 = f_{init\_c}(I), \quad (24)$$

here f_{init_h} and f_{init_c} are both the single FC layer. They untilize the feature vector I_f/I_3 to initialize the hidden state and cell memory of LSTM.

The loss of image captioning is the sum of negative log likeihood between the prediction caption of y and the ground truth caption of r at each time step:

$$L(y) = -\sum_{t=1}^{N} CrossEntroy(y_t, r_t), \quad (25)$$

here $CrossEntroy$ is cross entroy loss, which is widely used in multi-classification task. Such a loss converts the image captioning into a serialized sclassifcation-like task.

III. EXPERIMENTS

In this section, the image captioning data sets and evaluation metrics used in this paper are introduced firstly. Then the experimental settings are presented in detail. Finally we report the comparative results and analyse the influence of the proposed DMSFF mechanism when it is applied in remote sensing image captioning.

A. Data Sets and Evaluation Metrics

Two public remote sensing image captioning data sets are used in this paper: Sydney-captions [6] and UCM-captions [6]. For each image, there are five sentences describing it. The split of these two data sets follows the original literature by the ratio of 80%/10%/10% on training/validation/test.

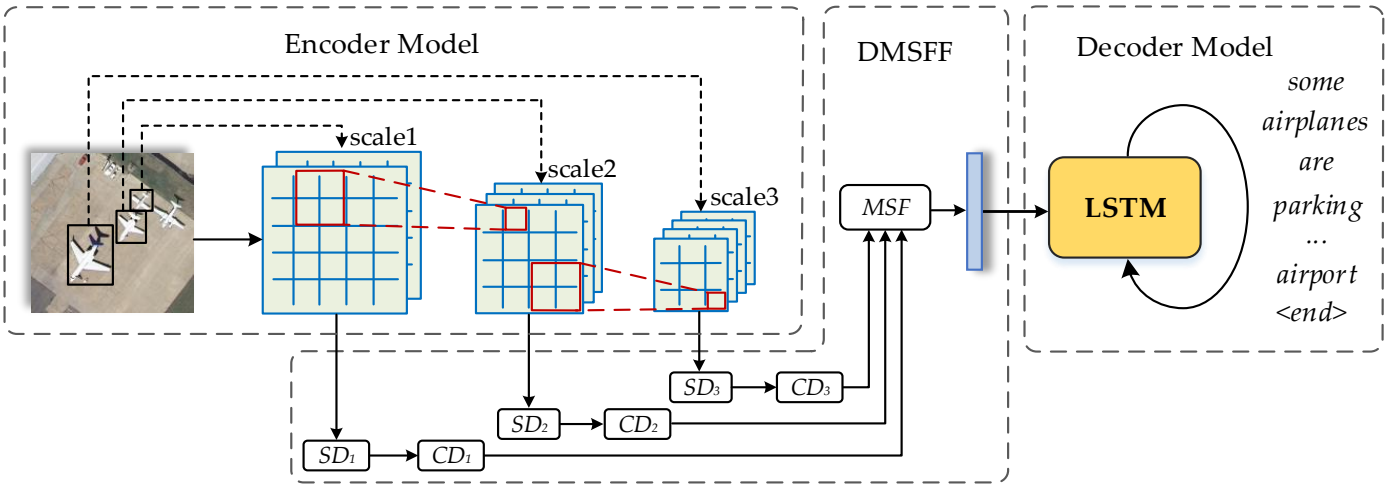

Fig. 2. The encoder-decoder architecture with the proposed DMSFF. **SD**: spatial-wise denoising, **CD**: channel-wise denoising, **MSF**: multi-scale feature feusion.

It is an import issue to evaluate the quality of the generated captions. There are many evaluation metrics used in image captioning. In this paper, the following evaluation metrics are used to comprehensively evaluate the generated captions of remote sensing images: BLEU1, BLEU2, BLEU3, BLEU4 [20], CIDEr [21] and ROUGE-L [22]. The higher scores of these metrics mean the better caption quality.

B. Experimental Settings

For the encoder model, comparative experiments are conducted on two widely used CNN architectures of VGG16 [16] and ResNet18 [17]. All of them are pre-trained on ImageNet in advance. In VGG16 and Resnet18, f_1, f_2 and f_3 are behind the 3rd, 4th and 5th max pooling layer, respectively.

In experiments, the images are resized to 224 x 224 and horizontally flipped with 50% probability. For the decoder model, the dimension size of word embedding vector is 256, and the input size of LSTM is the dimension sum of the word embedding vector and the visual feature vector, and the hidden state dimension size of LSTM is 256. Adam is selected to optimize the whole encoder-decoder framework with/without DMSFF with the learning rate set to 0.0001. All the models are trained for 50 epochs with the size of mini-batch set to 64. Our experiments are realized by Pytorch 1.3.1, and conducted on a computing equipment with 1× NVIDIA GeForce GTX 1080Ti GPU and 64G RAM CPU.

C. Ablation Results and Analysis

To verify the effectiveness of DMSFF, the ablation experiments based on VGG16 and ResNet18 are conducted to explore the influence of spatial-wise denoising (SD), channel-wise denoising (CD) and multi-scale fusion (MSF). The results are provided in Table I and II. According to the results, it could be found that MSF can help CNN-based encoder to get better visual features across data sets and feature extractors. On the basis of the extractor multi-scale features, SD and CD can further improve their quality with the weighting operation to varying degrees. Overall, the best results are achieved by the combination of MSF, SD and CD (DMSFF). Compared with the CNN baselines, for the most stable metric of BLEU, their DMSFF version obtains an average gain of 2.01 in Sydney-captions and 2.26 in UCM-captions. And for the other metrics, the score gain is also significant. The results of ablation experiments demonstrate that the proposed DMSFF can help CNN feature extractor acquire high-quality multi-scale feature, some of which may be ignored by the single-scale feature. There are some caption samples shown in Fig. 3.

(a) There is a white airplane parked on the airport with some airport buildings beside.

(b) An industrial area with many white buildings while some roads go through this area.

(c) There are some runways with white marking lines on while a river beside.

Fig. 3. Some caption results of Sydney-captions.

We further compare our model with some state-of-the-art methods. Their results are also listed in I and II. For Sydney-captions data set, our VGG16_MSF$_{SD+CD}$ and ResNet18_MSF$_{SD+CD}$, *i.e.* the DMSFF version of VGG16 and ResNet18, outperform the others with the significant advantage. In the experiments of this data set, we find that the great results benefit from the optimizer of Adam. For UCM-captions data set, our model outperforms not only the handcrafted feature of FV-LSTM and VLAD-LSTM [5], but also single-scale CNN feature of VGG16 and VGG19 [6]. Our ResNet18_MSF$_{SD+CD}$ just slightly fall behinds the GoogleNet_attention [5].

In general, the proposed DSMFF is beneficial for improving the quality of visual feature and the generated captions across data sets and CNN architectures.

TABLE I
COMPARATIVE RESULTS ON SYDNEY-CAPTIONS. '_MSF' REPRESENTS THE MULTI-SCALE FEATURE FUSION (CONCATENATION), 'SD' REPRESENTS THE SPATIAL-WISE DENOISING OPERATION AND 'CD' REPRESENTS THE CHANNEL-WISE DENOISING OPERATION.

Models	BLEU1	BLEU2	BLEU3	BLEU4	CIDEr	ROUGE_L
VGG16	80.78	71.99	62.07	55.00	3.0360	0.6831
VGG16_MSF	82.60	72.80	63.23	55.94	3.0586	0.6907
VGG16_MSF$_{SD}$	82.50	73.73	65.75	58.46	**3.1970**	**0.7194**
VGG16_MSF$_{CD}$	82.31	73.93	65.22	58.61	3.1702	0.7078
VGG16_MSF$_{SD+CD}$	**83.00**	**74.22**	**66.48**	**59.30**	3.1817	0.7068
ResNet18	81.45	72.23	62.06	54.17	2.9895	0.6921
ResNet18_MSF	82.52	73.41	63.67	57.28	3.0461	0.7112
ResNet18_MSF$_{SD}$	82.69	73.24	62.53	55.04	2.9223	0.6981
ResNet18_MSF$_{SD}$	82.73	73.44	64.74	57.66	3.1366	0.7145
ResNet18_MSF$_{CD+CD}$	**83.24**	**74.89**	**65.91**	**58.51**	**3.1898**	**0.7218**
VGG16 [6]	54.60	39.50	22.30	21.20	—	—
VGG19 [6]	54.80	39.80	22.80	21.50	—	—
FV-LSTM [5]	63.31	53.33	47.35	43.03	1.4761	0.5794
VLAD-LSTM [5]	49.12	34.72	27.60	23.14	0.9164	0.4201
GoogleNet_attention [5]	76.89	66.13	58.40	51.70	1.9863	0.6842
VAA [23]	74.31	66.46	60.29	54.95	2.4073	0.6999
Sound-f-a [24]	71.55	63.23	54.69	46.60	1.8027	0.6035

TABLE II
COMPARATIVE RESULTS ON UCM-CAPTIONS.

Models	BLEU1	BLEU2	BLEU3	BLEU4	CIDEr	ROUGE_L
VGG16	79.38	71.01	64.30	56.79	3.0143	0.6823
VGG16_MSF	80.01	71.21	64.57	56.42	3.0674	0.6929
VGG16_MSF$_{SD}$	80.51	71.62	64.73	56.54	3.0725	0.6934
VGG16_MSF$_{CD}$	**81.91**	71.83	65.78	58.63	3.1022	0.6981
VGG16_MSF$_{SD+CD}$	81.39	**73.23**	**66.65**	**58.82**	**3.1464**	**0.7112**
ResNet18	80.56	72.51	65.27	57.16	3.1759	0.7013
ResNet18_MSF	81.81	74.60	68.07	59.97	3.2496	0.7198
ResNet18_MSF$_{SD}$	82.12	75.69	69.45	61.01	**3.3357**	0.7237
ResNet18_MSF$_{CD}$	82.33	75.76	69.34	60.45	3.2962	0.7225
ResNet18_MSF$_{SD+CD}$	**83.06**	**75.98**	**69.72**	**63.45**	3.2956	**0.7318**
VGG16 [6]	63.50	53.20	37.50	21.30	—	—
VGG19 [6]	63.80	53.60	37.70	21.90	—	—
FV-LSTM [5]	58.97	46.68	40.80	36.83	1.8438	0.5595
VLAD-LSTM [5]	70.16	60.85	54.96	50.30	2.3131	0.6520
GoogleNet_attention [5]	**83.75**	**76.21**	**70.42**	**65.62**	3.2001	**0.7962**
VAA [23]	81.92	75.11	69.27	63.87	**3.3946**	0.7824
Sound-f-a [24]	78.28	72.76	67.59	63.33	3.2906	0.6864

IV. CONCLUSION

In this paper, we consider the remote sensing image captioning from the view of multi-scale feature extraction and fusion with the denoising operaiton. Correspondingly, the denoising based multi-scale feature fusion (DMSFF) mechanism is presented. The proposed DMSFF can be easily embedded into various CNN architectures and an end-to-end trainable encoder-decoder framework is further constructed. The proposed DMSFF can help the encoder-decoder framework to obtain multi-scale feature representation from remote sensing image and further improve the captioning performance. The comparative experiments on two public remote sensing image captioning data sets demomstrate the effectiveness and robustness of the proposed DMSFF.

REFERENCES

[1] Q. Wang, X. He, and X. Li, "Locality and structure regularized low rank representation for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing (T-GRS)*, vol. 57, no. 2, pp. 911–923, 2019.

[2] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing (T-GRS)*, vol. 57, no. 2, pp. 1155–1167, 2019.

[3] X. Lu, X. Zheng, and X. Li, "Latent semantic minimal hashing for image retrieval," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 355–368, 2017.

[4] J. Yuan, D. Wang, and R. Li, "Remote sensing image segmentation by combining spectral and texture features," *IEEE Transactions on geoscience and remote sensing*, vol. 52, no. 1, pp. 16–24, 2014.

[5] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, 2018.

[6] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *2016 International Conference on Computer, Information and Telecommunication Systems (CITS)*. IEEE, 2016, pp. 1–5.

[7] X. Zhang, Q. Wang, S. Chen, and X. Li, "Multi-scale cropping mechanism for remote sensing image captioning," in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, 2019.

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[9] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[10] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651–4659.

[11] Y. Pu, M. R. Min, Z. Gan, and L. Carin, "Adaptive feature abstraction for translating video to text," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[12] W. Jiang, L. Ma, X. Chen, H. Zhang, and W. Liu, "Learning to guide decoding for image captioning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[13] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5659–5667.

[14] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?" *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3623–3634, 2017.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

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

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

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

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[21] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[22] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, 2004.

[23] 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. 137 355–137 364, 2019.

[24] X. Lu, B. Wang, and X. Zheng, "Sound active attention framework for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2019.