Multi-Scale Methods for Optical Remote Sensing Image Captioning

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

***Abstract*—Recently, optical remote sensing image captioning task has gradually become a research hotspot because of its application prospects in military and civil fields. Many different methods along with datasets have been proposed. Among them, models following the encoder-decoder framework have better performance in many aspects like generating more accurate and flexible sentences. But almost all these methods are of a single fixed receptive field and couldn't put enough attention on grabbing multi-scale information, which leads to incomplete image representation. In this paper, we deal with the multi-scale proplem and propose two multi-scale methods named Multi-Scale Attention (MSA) method and Multi-Feats Attention (MFA) method, to obtain better representations for captioning task in remote sensing field. Respectively, the MSA method extracts features from different layers and uses multi-head attention mechanism to obtain the context feature. The MFA method combines target-level features and scene-level features by using target detection task as auxiliary task to enrich the context feature. 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, Multi-scale, Auxiliary task, Attention.**

I. INTRODUCTION

OPTICAL remote sensing image captioning is a technology to generate one or more sentences which 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 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 template-based methods, Z. Shi et al. [1] proposed an Fully Convolutional Networks (FCN)model to do captioning which mainly focuses on the multi-level semantics and semantic ambiguity problems. They use detection method to get information of the objects from different levels and use them to fill sentence templates which are pre-designed with multiple forms. For encoder-decoder-based methods, X. Lu et al. [2] proposed several different kinds of models using different Convolutional Neural Networks (CNNs) to extract image features with which to generate sentences using Recurrent Neural Network (RNN) or Long Short-Term Memory (LSTM). They also published a public dataset in this paper named Remote Sensing Image Caption Dataset (RSICD) and did sufficient experiments on it. From their work we can conclude that the methods used in natural image captioning field can be transferred to remote sensing image captioning field but they can only obtain acceptable descriptions. Other works like the method proposed in [3] which tries to obtain 5 sentences at the same time and can get more accurate and diverse results, method named Visual Aligning Attention (VAA) proposed in [4] tries to improve the attention masks' ability to focus on regions of interest in input images. These methods are also important works but are not very relevant to the contents of our article, so we will not elaborate them.

Recently, X. Zhang et al. in [5] proposed a multi-scale cropping mechanism for training remote sensing captioning models, which can extract advanced semantic features. 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, multi-scale 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 images caused by the different distance between camera and imaging objects and the scale difference between images objects. For remote sensing images, due to the large imaging range of the satellite cameras, the scale difference between objects like plane and airport, boat and harbor, people and beach is huge. Besides, for captioning task, to increase the diversity and hierarchy of description sentences, images of the same resolution will be scaled differently. Therefore, multi-scale methods are need to help achieve better captioning models in remote sensing field.

It is known that pyramid method is a widely-used strategy to solve the scale diversity problem. Pyramid based methods can be divided into two categories, image pyramid method [6] and feature pyramid method [7]. The image pyramid method is compute and memory intensive and is avoided in recent related research. The feature pyramid method using feature

The work was supported by the National Key R&D Program of China under the Grant 2017YFC1405605, the National Natural Science Foundation of China under the Grant 61671037, the Beijing Natural Science Foundation under the Grant 4192034 and the National Defense Science and Technology Innovation Special Zone Project. (Corresponding author: Zhenwei Shi)

Xiaofeng Ma(e-mail:max15@buaa.edu.cn), Rui Zhao(e-mail:ruizhao ipc@buaa.edu.cn) and Zhenwei Shi (Corresponding Author, e-mail: shizhenwei@buaa.edu.cn) are with Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and also with State Key Laboratory of Virtual Reality Technology and Systems, School of Astronautics, Beihang University, Beijing 100191, China.

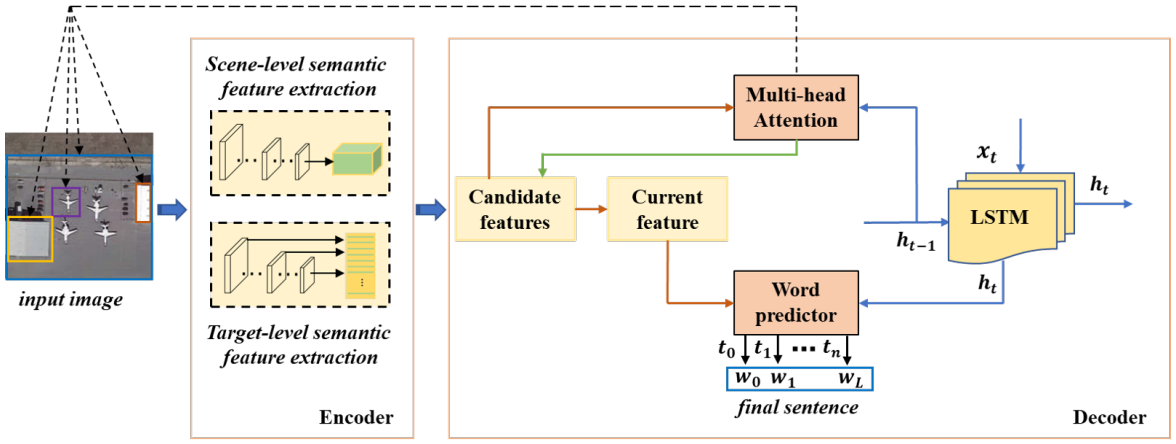


Fig. 1. The model structure of the proposed methods.

fusion is efficient and appears in many models which have best performance in the corresponding field. In this paper, we propose two multi-scale methods with regards to the scale diversity problem based on pyramid skills. Our work mainly has the following two contributions:

- Two multi-scale methods are proposed to achieve better representations of the input image and can alleviate the scale diversity problem to some extent.
- An improved attention mechanism, multi-head attention, is proposed, which can adaptively cascade 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 Multi-scale Method

The overall structure of our proposed method is shown in Fig.1. The structure mainly involves two parts: 1) Encoder, which consists of two modules called scene-level feature extraction module and target-level feature extraction module respectively. 2) Decoder, which consists of three modules named multi-head attention module, LSTM module and word predictor respectively. Among them, scene-level feature extraction module, LSTM module, word predictor are consistent with the structure of the classical method in [8] and we will just give a brief introduction to them afterwards in this subsection. As for target-level feature extraction module and multi-head attention module, which are new components appear 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 image features block by block. The LSTM module consists of 2 hidden layers and word predictor is designed with 2 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 in order to get more fine-grained semantic representations for the input optical remote sensing images. As can be seen from 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, feature vectors of the target-level feature extraction module are sparsely distributed and they will be formed as 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 which is proposed in our previous paper [11]. The parameters of the module will be obtained by taking optical remote sensing image object detection task as an auxiliary task. The output feature vectors of target-level feature extraction module have 21 dimensions, which are the logits of each detection block.

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

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

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

C. Multi-head Attention Module

Multi-head attention module take different feature lists from multiple scales as input, 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 figure 3.

At each time step t, based on the feature list F_i from multi-scale image features F_{all} of encoder output, and previous hidden state h_{t-1} of decoder output, attention network learns to generate attention weight:

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

where $\alpha_{t,i,j}$ is the attention weight corresponding to the feature j in F_i. Feature list is defined as features ($H_{scene,i} \times W_{scene,i} \times$

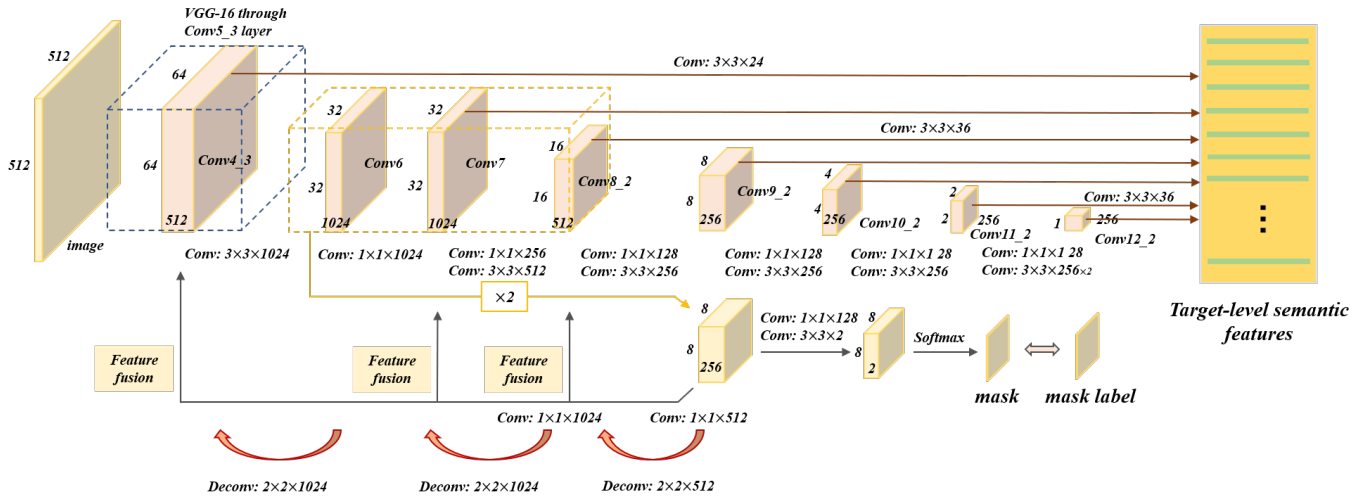


Fig. 2. The structure and prameter settings of the target-level semantic feature extraction module.

$D_{scene,i}$) from the same layer in the scene-level feature extraction backbone or the features ($N_{target} \times D_{target}$) 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 multi-layer perceptron (MLP), N_i is the number of elements in $F_\mathbf{i}$ and $\alpha_{t,i}$ is the weights corresponding to each element.

Base on the Multi-head attention module, the context vector at time step t can be formulated as:

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

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

D. MSA and MFA

As for our proposed two methods MSA and MFA, the difference between them is the selection of feature lists. Method MSA only use 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 optimized spatial attention mechanism. Method MFA use 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 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:

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

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

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

The loss is denoted as

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

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

III. EXPERIMENTS

A. Dataset and Metrics

1) Dataset: We use RSICD as the main dataset for experiments, which is constructed by Lu et al. [2]. It contains a total of 10921 remote sensing images, of which the training set contains 8004 images, the validation set and the test set contains 2187 images. These images are fixed into 224×224 pixels and the resolution of them are various. Each image are labeled with $1 \sim 5$ sentences and there are 24333 different label sentences in the label file altogether including 3323 words. Besides, UCM-captions and Sydney-captions are used to evaluate our methods too.

2) Metrics: Most of the existing evaluation methods for image sentence description generation tasks are from the field of machine translation. The evaluation methods used in our paper 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 co-occurrences of n-grams between the generated and the reference captions, where n-gram is a set of n ordered words. The n taken from 1 to 4, corresponding from BLEU-1 to BLEU-4. The higher the evaluation index score, the better the quality of the generated sentence.

B. Experiment Setting and Training details

The deep learning framework Tensorflow is used to implement our networks. The subnetwork of scene-level feature extraction module in encoder uses the pre-training parameters

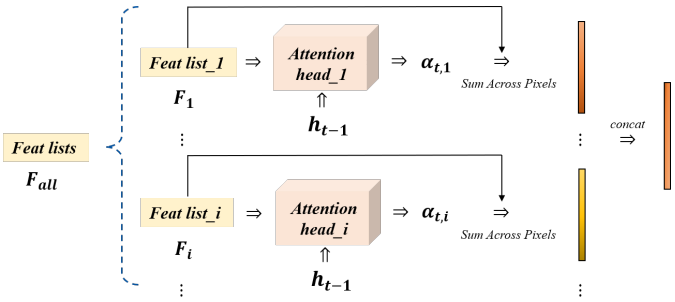


Fig. 3. The schematic of Multi-head Attention Mechanism.

of ResNet_50 [9] on the ImageNet for initialization, and other parameters are initialized randomly. We use the Adam Optimiser with a learning rate of 1e-4 and is decreased by factor of 0.9 after 1e+5 steps. The number of training epochs is set as 80.

The parameters in target-level feature extraction module are trained using auxiliary training dataset named DIOR which is proposed in [16]. DIOR consists 23463 optimal remote sensing images and 192472 object instances, covered by 20 common object categories. The size of images in the dataset is 800×800 pixels and the spatial resolutions range from 0.5m to 30m. We use DIOR as our auxiliary training dataset because the images in it have a large range of object size variations and the num of categories is largest up to now in remote sensing field for target detection. So the target-level feature extraction module can obtain target-level semantic features of different scales which helps achieve better representations of the input image. The training settings of it is the same as [11].

C. Comparison experiments

We compared our methods with the benchmark method which uses the 50-layers deep residual network (ResNet-50) without the last fully connected layer as encoder and LSTM as 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 MSA method and MFA method. Results of these three methods are shown in Table I and II. For the multi-scale cropping method proposed in [5], the authors didn't use attention mechanism in their method and the results in their paper indicate that the method performs not very well even on small datasets. So we won't compare our methods with their method their are also an effective way to solve the multi-sacle problem.

D. Results analysis

According to Table I and II, it can be seen that method MSA achieves significant improvement over the benchmark method in all the evaluations and method MFA achieves improvement over the benchmark in B-1, B-2, METEOR metrics but obtains no improvement in B-3, B-4, 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

TABLE I
RESULTS ON UCM-CAPTINS 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**

TABLE II
RESULTS ON 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

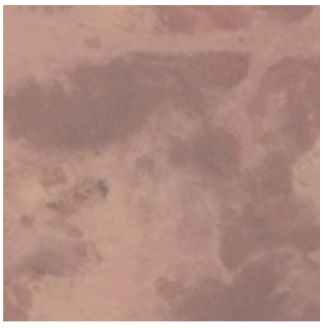
Fig.4. Each row beginning with (a) below the subfigures are the results of the benchmark method and (b)/(c) are the results of MSA and MFA methods. The red words are wrong descriptions of the input images, cyan words are less accurate descriptions of the input images and green words are more accurate 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 recognize 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 that "road", "river" in subfigue 6 and 8. But there are still some less accurate decriptions like "road" in subfigure 2 where it should be "beach" and "resort" in subfigure 5 where it should be "park".

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

IV. CONCLUSION

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

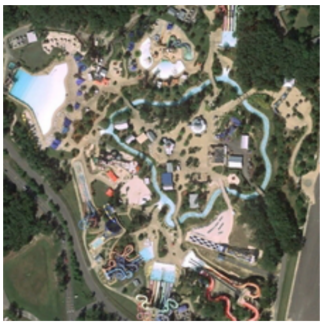
(a) many buildings are in an industrial area.
(b) many planes are parked near a terminal in an airport.
(c) many buildings are in an airport.



(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 benchmark method (a), MSA method (b) and MFA method (c) from different catogries. Red indicates wrong descriptions, cyan indicates less accurate descriptions and green indicates more accurate descriptions of the input images.

References

[1] 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, June 2017.

[2] 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, April 2018.

[3] B. Wang, X. Lu, X. Zheng, and X. Li, "Semantic descriptions of high-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 8, pp. 1274–1278, 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. 137 355–137 364, 2019.

[5] X. Zhang, Q. Wang, S. Chen, and X. Li, "Multi-scale cropping mechanism for remote sensing image captioning," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, July 2019, pp. 10 039–10 042.

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

[7] T. Y. Lin, P. Dollr, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2016.

[8] 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," *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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "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: Springer International Publishing, 2018, pp. 373–383.

[12] Papineni, Kishore, Roukos, Salim, Ward, Todd, Zhu, and Wei-Jing, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL 02. USA: Association for Computational Linguistics, 2002, p. 311318. [Online]. Available: https://doi.org/10.3115/1073083.1073135

[13] Lin and C. Yew, "Rouge: A package for automatic evaluation of summaries," in *In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, 2004.

[14] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*, ser. StatMT 07. USA: Association for Computational Linguistics, 2007, p. 228231.

[15] Vedantam, Ramakrishna, Zitnick, C. Lawrence, Parikh, and Devi, "Cider: Consensus-based image description evaluation."

[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," 2019.