# Deep Semantic Understanding of High Resolution Remote Sensing Image

Bo Qu*†, Xuelong Li*, Dacheng Tao‡, Xiaoqiang Lu*
*Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China.
†University of the Chinese Academy of Sciences, 19A Yuquanlu, Beijing, 100049, P. R. China.
‡Centre for Quantum Computation & Intelligent Systems and the Faculty of Engineering and Information Technology, University of Technology Sydney, 81 Broadway Street, Ultimo, NSW 2007, Australia.

***Abstract*—With the rapid development of remote sensing technology, huge quantities of high resolution remote sensing images are available now. Understanding these images in semantic level is of great significance. Hence, a deep multimodal neural network model for semantic understanding of the high resolution remote sensing images is proposed in this paper, which uses both visual and textual information of the high resolution remote sensing images to generate natural sentences describing the given images. In the proposed model, the convolution neural network is utilized to extract the image feature, which is then combined with the text descriptions of the images by RNN or LSTMs. And in the experiments, two new remote sensing image-captions datasets are built at first. Then different kinds of CNNs with RNN or LSTMs are combined to find which is the best combination for caption generation. The experiments results prove that the proposed method achieves good performances in semantic understanding of high resolution remote sensing images.**

## I. Introduction

In recent years, with the huge progress of remote sensing technology and the reduction of acquisition costs, lots of *high spatial resolution* (HSR) remote sensing images can be easily obtained. In order to make full use of these data, understanding the HSR images automatically and correctly by computers is a very meaningful problem, since semantic understanding of HSR images has many useful applications in both civilian and military fields, such as disaster monitoring, resources survey and so on.

Unfortunately, researches in the field of high resolution remote sensing nowadays mainly focus on object recognition [1],[2], image classification [3],[4], scene classification [5],[6] and image segmentation [7],[8]. These works could only recognize the objects in the images or get the class labels of the images. However, they could not recognize the attributes of the objects and the relation between each object. Namely, the aforementioned works can't understand the HSR images in a semantic level and semantic understanding of HSR images is still blank in remote sensing analysis area currently. Therefore, a framework is proposed in this paper to use both the visual and textual information of HSR images to understand the images in semantic level. Since the visual information of an image could reflect the image contents roughly and the textual information can describe the object attributes in more details, combining the visual information with textual information can build a powerful and effective model for semantic understanding the HSR images.

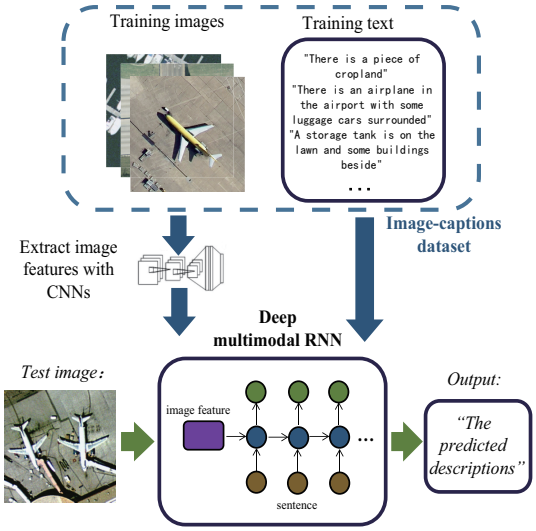


Fig. 1. The whole process of the the proposed method.

Recently, *Convolutional Neural Network* (CNN) has a very powerful ability of extracting expressive features from images which has achieved tremendous successes in the area of computer vision [9],[10],[11]. Thus, the deep CNNs is used in the proposed method to extract the abundant visual features from the HSR images. Furthermore, the *Recurrent Neural Network* (RNN) and *Long-Short Term Memory networks* (LSTMs) have the special ability of calculating output conditioned on the previous inputs. Therefor these networks are efficient in dealing with sequential data, such as text data, speech data and so on. Therefore, RNN and LSTMs are also used in the proposed model to combine the visual and textual information of images to further generate image captions.

Inspired by the success of deep learning in computer vision, in this paper, a deep multimodal recurrent neural network model is proposed, which combining the deep CNNs with RNN or LSTMs for semantic understanding of HSR images. Fig. 1 shows the overview of this method. More specifically, different CNNs are used to extract excellent visual features of HSR images and then the image features and reference

978-1-5090-0690-8/16/$31.00 

sentences of the image are imported to the RNNs or LSTMs to train the deep multimodal neural network model. Then the descriptions of new images can be generated by our trained model. In the experiments, the proposed model combines different CNNs with RNNs or LSTMs to find the best combination for semantic understanding of HSR images. In summary, the main contributions of this paper are as follows:

1) A deep multimodal neural network model for semantic understanding of HSR remote sensing images is proposed in this paper. This model is the first to take advantages of both visual and textual information to generate image captions in the field of high resolution remote sensing.
2) In this paper, different CNNs with RNN or LSTMs are combined into a deep multimodal model to find the best combination for HSR remote sensing image caption generation.
3) Two new HSR image-captions datasets are built based on the HSR remote sensing image datasets: *UC Merced* (UCM) dataset [12] and Sydney dataset [6], which are utilized for all the experiments.

The rest of this paper is organized as follows. In section II, we briefly discuss the related work on image caption generation. Then the deep multimodal neural network model is described in detail in section III. Section IV presents the experiments setup and the results of experiments. Finally, section V concludes this paper.

## II. Related Work

Semantic understanding of an image, which means letting computers automatically describe the content of the image, is a very important task in computer vision. But some basic tasks [4],[6],[13] still play a major role in the field of high resolution remote sensing.

As for the issue of generating natural language descriptions from an image, a growing body of researches have been done in both industrial and academic circles. First, lots of works have been proposed on learning joint image-word embeddings in [14],[15],[16]. They use the pre-trained deep neural network to transform the images into feature vectors. In parallel, the corresponding descriptions are projected to the same embedding space by a pre-trained language model. Then, by computing the similarity of images and sentences, their methods can generate captions for the new images.

In recent years, many researchers [17],[18],[19] propose to use a multimodal neural language model to generate image captions. They utilize the deep CNNs to extract the image features, then generate captions conditioned on the visual features and the words predicted before.

Although the aforementioned methods achieved good results in natural image processing, semantic understanding of the HSR remote sensing images is still blank. Besides, the effectiveness of these methods has not been verified in the field of high resolution remote sensing. Inspired by these facts, a deep multimodal neural network model for semantic understanding of HSR images is proposed in this paper. However, different from the previous methods, different CNNs with RNN or LSTMs are combined in the proposed model to find the best combination for semantic understanding of HSR images.

## III. Proposed model

The ultimate goal of this paper is to generate descriptions of a given HSR image. And the proposed model combines two different kinds of deep neural networks, CNN and RNN or LSTMs, to extract the images features and then combine with textual information to generate sentences. The overview of the proposed method is in Fig. 1. In addition, the proposed model will be introduced in detail in the following parts.

### A. Deep neural network

Recent research [20] shows that the deeper the neural network is, the more expressive the network is. So deep learning achieves notable success in computer vision. And there are some popular deep neural networks, such as deep CNNs, *Deep Belief Networks* (DBNs), RNNs and so on. We can find that CNN is deep in space and RNN is deep in time, for the reason that CNN is basically a standard neural network which is extended across space while RNN is extended through time. Both the CNN and RNN will used in the proposed model since CNN has a very powerful ability to extract expressive features from images and RNN is good at handling sequential data.

*1) Deep in space:* Traditional neural networks normally have an input layer, a hidden layer and an output layer. This kind of network is easy to train and understand, but it cannot deal with complicated problems. Recently, researchers find that neural networks with multiple hidden layers have excellent ability of feature learning and multi-layers can represent complex functions with fewer parameters. So they proposed the deep CNNs. CNNs have a number of convolutional layers and pooling layers after input layer where the convolutional layers could reduce the noise and enhance the signal as well as the pooling layers could down sample the feature maps. And finally the net connects to full-connected layers in order to form a feature vector which represents the input image.

*2) Deep in time:* RNN (Fig. 2) can deal with sequential data properly while the traditional neural network could not do. This is because that the recurrent connections of hidden layer units in RNN allow the memory of previous inputs to keep in the network's hidden layers, and then the memory will influence the output of network. The computations in RNN are in the following recurrent equations:

$$h_t = f(W_{xh}x_t + W_{hh}h_{t-1}), \tag{1}$$

$$y_t = f(W_{hy}h_t), \tag{2}$$

where $x_t, h_t \in R^N, y_t$ represent the input, hidden state and output of RNN at time $t$, respectively. And $f(\cdot)$ is a non-linearity function, such as hyperbolic tangent function or sigmoid function. In addition, all the symbols $W$ are the weight matrices between the layers of RNN.

Although RNN has been successfully used in the tasks of natural language processing, speech recognition and so on, it still could not deal with the problem of "long-term dependencies" of states. This is because of the gradient vanishing and

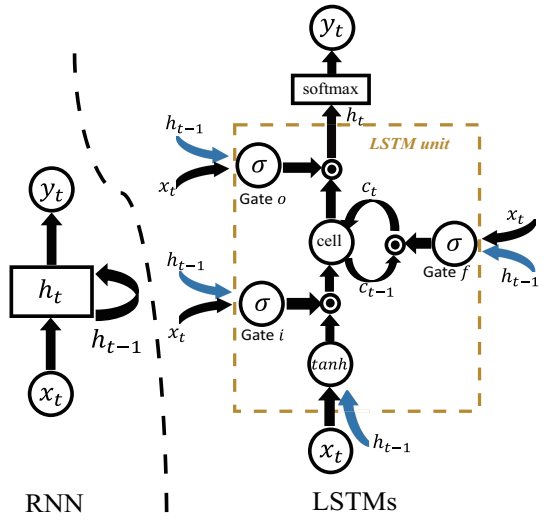


Fig. 2. Left is the recurrent neural network and right is the LSTM networks.

gradient exploding problem in the training stage of RNN, but the LSTMs could solve this problem well.

The cores of LSTMs are three "gates" and a memory cell, as we can see in Fig. 2. Each gate is a way to optionally let information through and is composed of a sigmoid function layer and a multiplication operation. The non-linear sigmoid function squashes the inputs to a [0,1] range, determining how much information could go through. 0 means letting nothing go through the gate and 1 means allowing everything to go through. Then the three gates work together to control the final output of the LSTM unit. More specifically, the input gate controls whether to send the new input to the cell, the forget gate decides if it should forget the current cell value and output gate determines whether to output the cell value. The definition of the gates and the update of cell information at timestep $t$ are as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1}), \quad (3)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1}), \quad (4)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1}), \quad (5)$$
$$g_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1}), \quad (6)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t, \quad (7)$$
$$h_t = o_t \odot c_t, \quad (8)$$

where $i_t, f_t, o_t, c_t$ represent the outputs of input gate $i$ , forget gate $f$ , output gate $o$ and the cell $c$ at time $t$, respectively. And $\sigma(\cdot)$, $\tanh(\cdot)$ are the sigmoid function and hyperbolic tangent function, $\odot$ represents the product with a gate value, and all the matrices $W$ are the parameters to train. The three control gates could deal well with the problems of exploding and vanishing gradients therefore make it possible to train the LSTMs robustly.

### *B. Deep multimodal neural network model*

In order to generate a description for a HSR remote sensing image, it is necessary to consider maximizing the probability

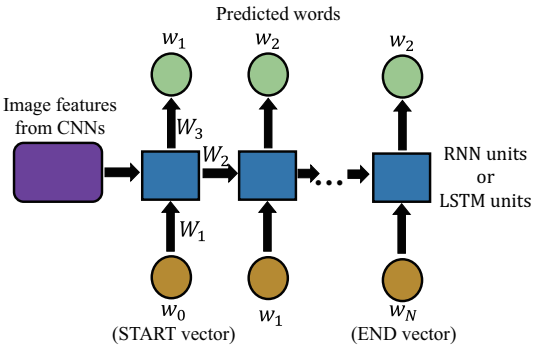


Fig. 3. Training process of the deep multimodal neural network.

of the correct descriptions conditioned on the given image information. Then specific to every single description, we can get the following formulation:

$$\lambda^* = \arg\max_{\lambda} \sum_{(I,S)} \sum_{t=0}^{N} \log p(w_t \,|I, w_0, w_1 ... w_{N-1}; \lambda), \quad (9)$$

where $I$ is a HSR image, $S = \{w_1, w_2 ... w_N\}$ is the correct image caption and $\{w_1, w_2 ... w_N\}$ are the words in the description, $\lambda$ is the parameter of our model. The inner summation is to sum the log probabilities of words in a single sentence and the outer is to sum all image-caption pairs in training data. And different CNNs are utilized to extract the image information and RNN or LSTMs are used to combine the textual information in the deep multimodal neural network model. This model is deep in both space (CNN) and time (RNN) and the multimodal means combining the visual information with the textual information.

*1) Training stage:* At the training stage, $(I, S)$ is the training pair from the HSR image-captions datasets and the proposed model is trained to predict every word in the image description after given an image. So the unrolled form of RNN or LSTMs are used to train the network with both images and their descriptions. And in more details, the training stage is iterating the following formulations from t=1 to N:

$$b_0 = CNN(I), \quad (10)$$
$$h_t = g(W_1 w_t + W_2 h_{t-1} + b_0 \cdot \mathrm{I}(t = 1)), \quad (11)$$
$$p(w_{t+1}) = softmax(W_3 h_t), \quad (12)$$

where $b_0$ is the image feature vector, $g(\cdot)$ is computational process of RNN or LSTMs. Note that $w_1$ is a special START vector and $w_N$ is an END vector, which represent the start and end of the sentence, respectively.

As it can be seen in Fig. 3, at each training step, the input word $x_t$ and the previous hidden state $h_{t-1}$ are combined to get the hidden state $h_t$ at this time. Then $h_t$ goes through the softmax function to calculate the probability distribution of the next word and the word of the highest probability is chosen to be the predicted word. And then the predicted word is the input for the next timestep. Finally, repeat the above steps until the network predicts the END vector.

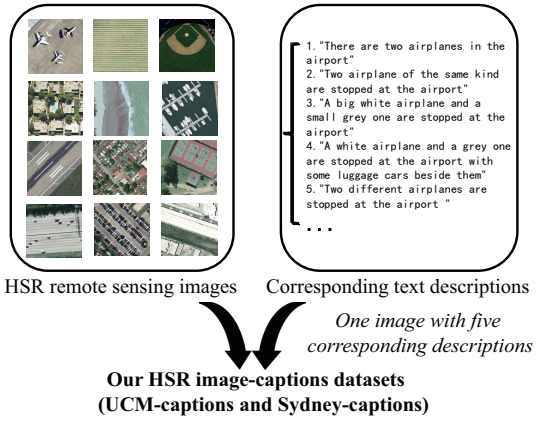


Fig. 4. The HSR image-captions datasets.

The loss function of the proposed model is the sum of the negative log likelihood of the correct word at each step as follows: $loss(I,S) = -\sum_{t=1}^{N} \log p(w_t)$ . We minimize the above loss function at the training stage to get the best parameters of the proposed model.

*2) Inference stage:* The inference stage of the proposed method is similar to the training stage: the test image is input to the trained model then the model predicts one new word at a timestep. Finally, all the predicted words constitute a description of the test image (see Fig. 3).

## IV. Experiments

In this section, the datasets used in the experiments and the evaluation metrics for the experiment results are firstly described. Then the deep multimodal neural network is used to generate description of HSR images in two different datasets. Finally, the results of our experiments will be discussed.

### A. Datasets and Evaluation Metrics

Semantic understanding of HSR images by a deep multimodal framework is an innovative problem in computer vision. And a dataset of HSR images with corresponding sentences is needed for experiments. However, no such dataset exists. Therefore two new datasets of HSR images with corresponding descriptions are built in this paper which based on the HSR remote sensing datasets: UCM dataset and Sydney dataset. The UCM dataset totally has 2100 HSR images which are divided into 21 challenging scene categories and the Sydney dataset contains 7 different scene categories and totally has 613 HSR images. Then every HSR image is annotated with 5 reference sentences in the two datasets (see Fig. 4). Thus two HSR image-captions datasets are built: one has totally 2100 HSR remote sensing images with 10500 descriptions ( UCM-captions dataset) and the other has 613 images with 3065 captions (Sydney-captions dataset). In the following experiments 80% image-captions in the datasets are used as training data, 10% as validation data and the rest 10% as test data.

TABLE I
Evaluation of image caption predict on the UCM-captions dataset. B-n is BLEU score uses for n-gram. High is good in all metrics.

| | | B-1 | B-2 | B-3 | B-4 | METEOR | CIDEr |
|---|---|---|---|---|---|---|---|
| RNN | AlexNet | 57.5 | 45.5 | 30.7 | 18.7 | 17.5 | 35.6 |
| | VGG-16 | 60.1 | 50.7 | 32.8 | 20.8 | 19.3 | 42.8 |
| | VGG-19 | 60.3 | 51.1 | 33.1 | 21.2 | 19.5 | 42.9 |
| | GoogLe Net | 59.2 | 51.3 | 33.5 | 20.7 | 19.1 | 42.2 |
| LSTM | AlexNet | 61.2 | 48.2 | 32.7 | 19.8 | 18.6 | 38.5 |
| | VGG-16 | 63.5 | 53.2 | 37.5 | 21.3 | 20.3 | 44.5 |
| | VGG-19 | **63.8** | **53.6** | **37.7** | 21.9 | **20.6** | **45.1** |
| | GoogLe Net | 63.7 | 53.1 | 37.1 | **22.2** | 20.1 | 44.7 |

TABLE II
Evaluation of image caption predict on the Sydney-captions dataset.

| | | B-1 | B-2 | B-3 | B-4 | METEOR | CIDEr |
|---|---|---|---|---|---|---|---|
| RNN | AlexNet | 46.2 | 34.6 | 18.5 | 17.7 | 16.3 | 28.9 |
| | VGG-16 | 51.3 | 37.5 | 20.4 | 19.3 | 18.5 | 32.2 |
| | VGG-19 | 51.9 | 37.4 | 20.9 | 19.7 | 18.8 | 32.5 |
| | GoogLe Net | 51.6 | 39.1 | 20.5 | 18.9 | 18.1 | 32.4 |
| LSTM | AlexNet | 51.7 | 36.2 | 19.6 | 19.5 | 17.9 | 32.5 |
| | VGG-16 | 54.6 | 39.5 | 22.3 | 21.2 | 20.5 | 37.2 |
| | VGG-19 | **54.8** | **39.8** | **22.8** | 21.5 | **20.8** | **37.9** |
| | GoogLe Net | 54.6 | 39.1 | 21.2 | **21.7** | 19.9 | 37.5 |

Three evaluation metrics are used in the experiments: BLEU, METEOR and CIDEr scores. Every metric evaluates how well the predicted description matches with the five reference sentences in the dataset. In detail, the BLEU score analyzes the co-occurrences of n-grams between the predicted and reference sentences with a brevity penalty. Besides, METEOR score is calculated by building an alignment between the words in the predicted and reference sentences. Finally, the CIDEr score calculates the *Term Frequency Inverse Document Frequency* (TF-IDF) weighting for each n-gram to compute the final score. The three metrics have their own advantages so we use them all in the experiments.

### B. Image caption generate

In this section, the deep multiomodal recurrent neural network is used to generate descriptions of HSR images. Three different CNNs: AlexNet [9], VGGNet (16-layers net and 19-layers net) [10] and GoogLeNet [11] are tested. All the networks are pre-trained on the huge ImageNet dataset and the last full-connected layer output is used as the image feature vector. Then RNN and LSTMs are utilized in the word inference part. Different CNNs with RNN or LSTMs are combined to find the best combination.

### C. Results

The results of image caption generation by the deep multimodal neural network which evaluated by different metrics are illustrated in Table I and Table II. These two tables compare the results of different combinations of deep CNNs with RNN

or LSTMs for image caption generation on the UCM-captions dataset and Sydney-captions dataset.

According to the two tables, the generation results roughly increase from the AlexNet to the VGGNet to the GoogLeNet, just like the feature extracting ability of different CNNs gradually enhancing. Then we can see that the result of using LSTMs is better than using RNN, and the combination of VGG-19 Net with LSTMs almost achieves the best results in all metrics. Besides, in experiments we find that although training LSTMs is slower than RNN, LSTMs gets better results than RNN, and furthermore LSTMs could solve the problem of vanishing or exploding gradients perfectly.

Fig. 5 shows some intuitive caption generation results of the HSR images. The proposed model could generate pretty good descriptions of HSR images and even distinguish the number of objects in the image (see the two images with different numbers of storage tanks). However, there are still some wrong results (see the last two images) in the experiments. The reason is that the datasets used in the experiments are not big enough to train a model of strong ability to distinguish very similar images. In the future, we will build some bigger HSR image-captions datasets in order to train a better model for semantic understanding of HSR images.

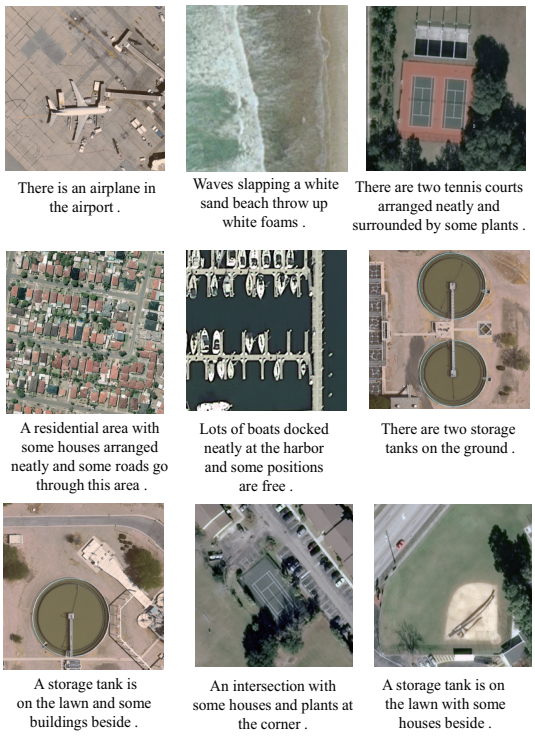


Fig. 5. (Best viewed in colors and magnification.) The result of HSR image caption generation.

## V. Conclusion

In this paper, a deep multimodal neural network model is proposed to solve the problem of understanding HSR remote sensing images in the semantic level. Different CNNs with RNN or LSTMs are combined in the experiments and the VGG19-layers network with LSTMs is the best combination for HSR remote sensing image caption generation.

## References


[1] J. Inglada, "Automatic recognition of man-made objects in high resolution optical remote sensing images by svm classification of geometric image features," *ISPRS journal of photogrammetry and remote sensing*, vol. 62, no. 3, pp. 236–248, 2007.

[2] L. Eikvil, L. Aurdal, and H. Koren, "Classification-based vehicle detection in high-resolution satellite images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 64, no. 1, pp. 65–72, 2009.

[3] J. Lin, Q. Wang, and Y. Yuan, "In defense of iterated conditional mode for hyperspectral image classification," in *Multimedia and Expo (ICME), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1–6.

[4] H. Lv, X. Lu, and Y. Yuan, "Data-dependent semi-supervised hyperspectral image classification," in *Signal and Information Processing (ChinaSIP), 2013 IEEE China Summit & International Conference on*. IEEE, 2013, pp. 664–668.

[5] M. Fu, Y. Yuan, and X. Lu, "Unsupervised feature learning for scene classification of high resolution remote sensing image," in *Signal and Information Processing (ChinaSIP), 2015 IEEE China Summit and International Conference on*. IEEE, 2015, pp. 206–210.

[6] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 53, no. 4, pp. 2175–2184, 2015.

[7] R. Trias-Sanz, G. Stamon, and J. Louchet, "Using colour, texture, and hierarchial segmentation for high-resolution remote sensing," *ISPRS Journal of Photogrammetry and remote sensing*, vol. 63, no. 2, pp. 156–168, 2008.

[8] B. Johnson and Z. Xie, "Unsupervised image segmentation evaluation and refinement using a multi-scale approach," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 4, pp. 473–483, 2011.

[9] 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.

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

[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[12] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2010, pp. 270–279.

[13] Y. Yuan, M. Fu, and X. Lu, "Substance dependence constrained sparse nmf for hyperspectral unmixing," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 53, no. 6, pp. 2975–2986, 2015.

[14] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 595–603.

[15] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov *et al.*, "Devise: A deep visual-semantic embedding model," in *Advances in Neural Information Processing Systems*, 2013, pp. 2121–2129.

[16] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[17] 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.

[18] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[19] X. Chen and C. Lawrence Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2422–2431.

[20] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.