Exploring Transformer and Multilabel Classification for Remote Sensing Image Captioning

Hitesh Kandala, Sudipan Saha, *Member, IEEE*, Biplab Banerjee, *Member, IEEE*, and Xiao Xiang Zhu, *Fellow, IEEE*

Abstract—**High-resolution remote sensing images are now available with the progress of remote sensing technology. With respect to popular remote sensing tasks, such as scene classification, image captioning provides comprehensible information about such images by summarizing the image content in human-readable text. Most existing remote sensing image captioning methods are based on deep learning-based encoder–decoder frameworks, using convolutional neural network or recurrent neural network as the backbone of such frameworks. Such frameworks show a limited capability to analyze sequential data and cope with the lack of captioned remote sensing training images. Recently introduced Transformer architecture exploits self-attention to obtain superior performance for sequence-analysis tasks. Inspired by this, in this work, we employ a Transformer as an encoder–decoder for remote sensing image captioning. Moreover, to deal with the limited training data, an auxiliary decoder is used that further helps the encoder in the training process. The auxiliary decoder is trained for multilabel scene classification due to its conceptual similarity to image captioning and capability of highlighting semantic classes. To the best of our knowledge, this is the first work exploiting multilabel classification to improve remote sensing image captioning. Experimental results on the University of California (UC)-Merced caption dataset show the efficacy of the proposed method. The implementation details can be found in https://gitlab.lrz.de/ai4eo/captioningMultilabel.**

Manuscript received 3 December 2021; revised 18 March 2022 and 24 May 2022; accepted 14 July 2022. Date of publication 11 August 2022; date of current version 26 August 2022. This work was supported in part by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Program (*So2Sat*) under Grant ERC-2016-StG-714087, in part by the Helmholtz Association through the Framework of Helmholtz Artificial Intelligence (AI) under Grant ZT-I-PF-5-01, in part by the Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)" and Helmholtz Excellent Professorship "Data Science in Earth Observation—Big Data Fusion for Urban Research" under Grant W2-W3-100, in part by the German Federal Ministry of Education and Research (BMBF) through the Framework of the International Future AI Lab "Artificial Intelligence for Earth Observation (AI4EO): Reasoning, Uncertainties, Ethics and Beyond" under Grant 01DD20001, and in part by the German Federal Ministry of Economics and Technology through the Framework of the "National Center of Excellence ML4Earth" under Grant 50EE2201C. *(Corresponding author: Xiao Xiang Zhu.)*

Hitesh Kandala and Biplab Banerjee are with the Indian Institute of Technology Bombay, Mumbai 400076, India (e-mail: hitesh_1603@iitb.ac.in; bbanerjee@iitb.ac.in).

Sudipan Saha is with the Department of Aerospace and Geodesy, Chair of Data Science in Earth Observation, Technical University of Munich, 85521 Munich, Germany (e-mail: sudipan.saha@tum.de).

Xiao Xiang Zhu is the Department of Aerospace and Geodesy, Chair of Data Science in Earth Observation, Technical University of Munich, 85521 Munich, Germany, and also with the Remote Sensing Technology Institute, German Aerospace Center (DLR), 82234 Weßling, Germany (e-mail: xiaoxiang.zhu@dlr.de).

Digital Object Identifier 10.1109/LGRS.2022.3198234

Index Terms—**Auxiliary task, image captioning, multitask learning, remote sensing, Transformer.**

I. Introduction

REMOTE sensing technology has made significant progress in the last decade, thus making high-quality remote sensing images available from a plethora of sensors. Despite this, the commonly studied remote sensing tasks, e.g., image segmentation and change detection, usually focus on object-level or pixel-level understanding without comprehensive semantic knowledge. Toward capturing more comprehensive global semantic information, image captioning is introduced in remote sensing that can generate intuitive textual descriptions summarizing the high-level semantic information [1], [2].

Image captioning is a challenging task, as it involves both understanding the content of the image and translating it to natural language. Early remote sensing image caption methods used template-based and retrieval-based models [3], [4]. Subsequently, they have been replaced by encoder–decoder-based methods. More recently, the visual attention mechanism has also been explored [5]. The Transformer further exploits the attention mechanism to model the sequence dependency and excludes the usage of recurrent units [6], [7]. In addition to traditional computer vision tasks, such as segmentation [8], Transformer-based architectures have also been adopted for computer vision image captioning [9]. Their works show the superior capability of Transformers to utilize long-range dependencies among the sequenced patches via the self-attention mechanism.

While Transformer can potentially improve image captioning [10], their performance may fall when sufficient training data are not available, as observed in [11]. Notably, the remote sensing image captioning datasets (RSICDs) are much smaller than those available in computer vision. Auxiliary/supplemental tasks and multitask learning are used to alleviate the lack of large training data by providing additional supervision, i.e., simultaneously using the same data for a different supplemental learning task during the training procedure [12]. The intuition behind their success is that the network learns to generalize better by adapting to multiple tasks. Such supplemental tasks are usually collected from related tasks; e.g., [13] uses image classification as an auxiliary task while generating synthetic images, and [14] explores multitask learning for human settlement extent regression

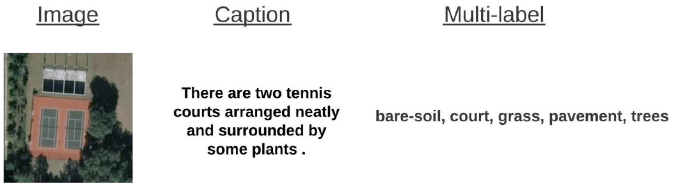


Image
Caption
Multi-label
There are two tennis courts arranged neatly and surrounded by some plants .
bare-soil, court, grass, pavement, trees

Fig. 1. Image with its caption and multilabel data, and the relationship between them is evident.

and local climate zone classification. Supplemental tasks can be both supervised [13] and unsupervised [15]. They have been used in several works related to image captioning in computer vision [16], [17]. Zhao *et al.* [16] used three related tasks of image captioning, multilabel classification, and syntax generation using the CNN-long short-term memory (LSTM) model, where all the three tasks share the CNN encoder, and the first and the last tasks share the LSTM decoder. Zhou *et al.* [17] jointly tackles two related visual-text tasks of image captioning and visual question answering. While syntax or visual question answers are not abundantly available in remote sensing, image labels are easily available. Furthermore, multilabel classification and image captions are conceptually similar [18], as both highlight the semantic classes, evident in the example shown in Fig. 1. Motivated by this, we propose to use multilabel image classification as a supplemental task along with Transformer-based remote sensing caption generation. Any other task, e.g., rotation prediction, could be used in practice. However, those tasks are focused on getting better discriminative visual features and do not highlight the semantic meaning, unlike multilabel classification, which makes the appropriate choice as a supplemental task to regulate the caption generation. Our proposed model benefits from the superior capability of the Transformer to exploit sequence information and the capability of multitask learning to perform training with limited data. The contributions of our work are as follows.

1) We propose a novel remote sensing image caption generation model that exploits recently popular Transformer architecture along with multilabel classification as a supplemental task. To the best of our knowledge, this is the first work jointly tackling multilabel classification and remote sensing image captioning.
2) We compare our method not only to the existing methods but also to other auxiliary tasks, showing that the chosen auxiliary task is most suitable for regulating the Transformer-based model.

II. Proposed Approach

A. Methodology

Given a remote sensing image I, remote sensing captioning generates its textual description—$S : S_1, S_2, \ldots, S_N$, where N is the total number of words in the sentence S. In practice, the training process is accomplished by training a model with model parameters θ_1 that maximizes the probability of the generated caption S given the input image I.

The proposed model is trained with the abovementioned objective function using a Transformer-based encoder. (Section II-B)and Transformer-based decoder (Section II-C). In addition, we use an auxiliary LSTM-based decoder (Section II-D) that ingests the bottleneck features directly from encoder and performs multilabel classification. The auxiliary decoder is trained to optimize the parameters θ_2 given input image I and its ground-truth (GT) corresponding labels $y_1, y_2, \ldots, y_M$. Note that parameters θ_1 and θ_2 share the encoder weights. The two tasks—caption generation and classification—are regulated by different loss functions, as described in Section II-E. The proposed framework is shown in Fig. 2.

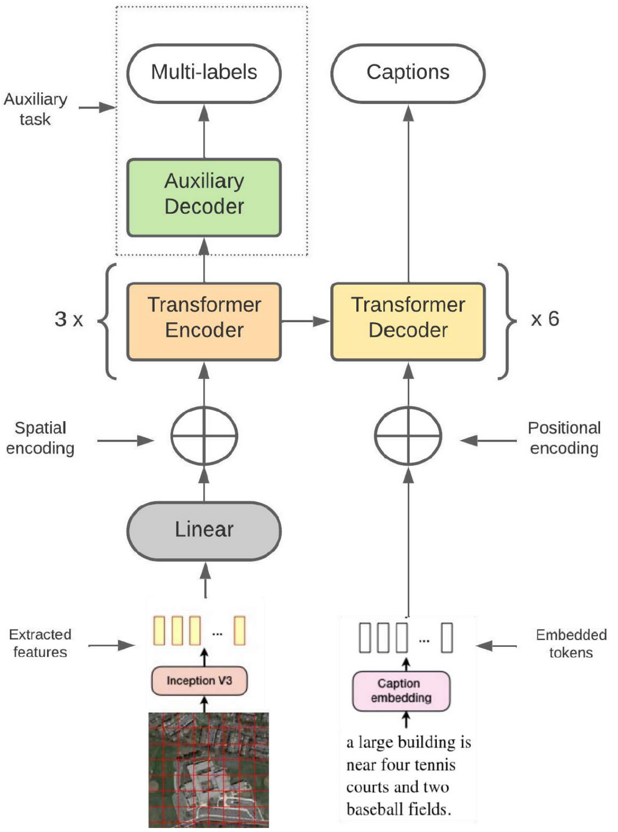


Fig. 2. Proposed multitask network, with a common transformer encoder and two decoders, one for caption generation and the other for multilabel classification.

B. Encoder

Encoders, mainly as an encoder–decoder pair, have seen their usage in different tasks, including autoencoder-based reconstruction and sequence-to-sequence learning tasks [19]. In general, the encoder part is composed of a sequence of convolution, pooling, and batch normalization layers to learn a concise feature representation of the whole image. Similarly, in sequence-to-sequence learning tasks, CNN layers are substituted by LSTM layers [19]. In this regard, Transformer encoders have been shown to learn better sequence feature representation than the LSTM. Also, recent advances have shown it to learn a rich feature space for images by exploiting its self-attention layers and highlighting the relevant parts of the image [6].

We use the combination of a CNN-based feature extractor that gives us a higher-level semantic feature and a Transformer encoder to get a good feature representation of the image. Since we have limited data, this combinatorial approach is suitable. As shown in Fig. 2, we split the image into 8 × 8 grid tiles and extract features for each tile using Inception v3 [20], extracting features after the mixed-seven layer. Since CNN gives us higher-level semantic features for the image, extracted features can be seen as a sequence representation. Instead of positional encoding, we do a spatial encoding based on the tile's position. In this way, we have the original image, except that it is replaced by a much higher representation. This is then passed through the multihead attention layers of Transformer for the extracted features to learn dependency on each other and give importance to relevant parts in the image. The multihead attention allows the model to attend to information from different representation subspaces at different positions. Thus, by the combination of Inception v3-based feature extraction and Transformer, we obtain a high-level semantic representation of the image while taking care of its spatial information and simultaneously building deeper relationships between the two using the self-attention heads of the Transformer.

The encoding component of the Transformer [6] is composed of a stack of encoders. While all the encoders are identical, they can be broken down into positional encoding, self-attention, and feedforward network layers. In the case of image input, the position of a tile in the image plays a determining role in understanding the sequence of the image that is imposed through a positional encoding layer. Output from the positional encoding layer goes to the self-attention layer. The self-attention mechanism computes the score by taking the dot product of the query vector with the key vector. This score is appropriately scaled and passed through the softmax layer to get the scores as probabilities. This attention mechanism can also be described as scaled dot-product attention [6]. The output from multihead attention is fed to the feedforward network. Since the feedforward layer expects only a single matrix (a vector for each word), the output obtained from multiple heads of self-attention is combined by additional weights. The output of the last encoder is transformed into key and value vectors, which are used by the decoder in its encoder–decoder attention layer.

For text encoding, we do not use any pre-trained model. We use byte pair encoding—subword-neural machine translation (NMT)—to build the vocabulary (dictionary) and use Moses tokenizer for captions.

C. Decoder

As mentioned previously, encoder–decoder pairs work hand-in-hand in sequence analysis tasks. Traditional LSTM decoders learn the feature representation by taking input in a sequential manner, limiting the efficiency of learning the long dependencies. Therefore, for the task of image captioning, we use a Transformer decoder that ingests the whole sentence at a time and uses its stacked self-attention layers to solve the abovementioned problem of learning long dependencies.

The Transformer decoder has similar architecture to the encoder. The output from the encoder is fed to the decoder using an encoder–decoder attention layer that works just like the multiheaded self-attention layer. Decoder layers use a masked self-attention sublayer to allow the model to attend to only earlier positions in the output sequence by masking the future positions. The decoder stack outputs are finally passed through a linear layer to produce a vocabulary size vector. This vector represents the probability of each word in the vocabulary being the following word in the sentence. In short, the decoder tries to find the probability of the next word, given the previous words and the spatial and semantic information.

D. Auxiliary Decoder

While our primary task is to generate captions for the input remote sensing images, we use multilabel classification as a supplemental task that helps us in improving the primary task by regularizing the features learned by the Transformer-based encoder. The output from the Transformer-based encoder is directly fed to the auxiliary decoder, bypassing the Transformer decoder. The encoder generates a high-level representation of the image packed in an embedded vector. For multilabel classification, we essentially need to decode this vector, as mentioned earlier, into labels to classify. So, for this task, we use a simple LSTM decoder to generate a feature vector, which is then passed through a sigmoid layer to obtain probabilities of labels.

To describe the setting of multitask learning, we reiterate the frameworks mentioned earlier and their combined flow. First, the image is fed into the encoder, which generates a rich and compact feature representation. At the same time, the sentence is passed through a masked self-attention layer to generate words sequentially. The output from the encoder goes to both the decoders. The transformer decoder uses the encoder output to learn codependencies between the text and the semantic information. While in the auxiliary decoder, it decodes the vector at every time step, which is then concatenated with the input at the next time step and then decoded again to generate a rich feature vector, which when passed through sigmoid generates a multilabel prediction $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_M$ for the image.

The model architecture is tabulated in Table I.

E. Loss Functions

For training the image captioning decoder, we use label-smoothed cross-entropy loss $\mathcal{L}_1$

$$\mathcal{L}_1 = (1-\epsilon)\left(-\sum_{i=1}^{N}\log(p(S_i|S_1,\ldots,S_{i-1}))\right) + \frac{\epsilon}{K}\beta \qquad (1)$$

where ϵ is a weight factor, β is the smooth loss, and K is the vocab size. β/K is the label smoothing loss [21], which tries to make one-hot label vector into a uniformly distributed vector to prevent model from overfitting and overconfidence.

For training the multilabel classification decoder, we use binary cross-entropy loss $\mathcal{L}_2$

$$\mathcal{L}_2 = -\frac{1}{M}\sum_{i=1}^{M} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i). \qquad (2)$$

We alternate between both the losses randomly at every mini-batch, which after some epochs generalizes to taking backpropagation through both the losses simultaneously. Alternatively, a combined weighted loss can be used.

TABLE I

MODEL FRAMEWORK: FEATURES EXTRACTED FROM INCEPTION V3 ARE FED TO TRANSFORMER-BASED ENCODER, THE OUTPUT FROM WHICH IS FED TO TWO DECODERS, ONE BASED ON TRANSFORMER FOR GENERATING CAPTION AND THE OTHER BASED ON LSTM FOR GENERATING MULTICLASS LABELS

Component	**Nature of layer**	**Layer**	**Input**
Inception v3	CNN-feature extraction	till mixed-7	Image
Encoder	Transformer-encoder	3	Inception v3 output
Decoder-1	Transformer-decoder	6	Encoder output
Decoder-2	LSTM-decoder	3	Encoder output

F. Overheads Over Traditional Captioning

Compared with the existing image captioning methods, the proposed method requires multilabel scene labels for the training images. However, compared with the target task, i.e., captioning, it is much less challenging to obtain scene labels, and both can be annotated simultaneously in any practical setting. Moreover, such labels are only required for training. For test/deployment, we do not require any prior knowledge about image labels. The overhead for computation time is negligible, as both decoders can be trained simultaneously.

III. EXPERIMENTAL VALIDATION

A. Test Dataset

We used the University of California (UC)-Merced captions dataset for experimental validation, extending the popular UC-Merced dataset. The UC-Merced dataset is a 21-class land use remote sensing image dataset, with 100 images per class. The images were manually extracted from large images from the United States Geological Survey (USGS) National Map Urban Area Imagery collection for various urban areas around the country [22]. The pixel resolution of this dataset is 0.3 m/pixel. Most images in the dataset are 256 × 256 pixels. The dataset was extended for multilabel classification in [23], with up to seven labels per image.

The UC-Merced captions dataset, introduced in [3], extended the UC-Merced dataset with five reference sentences per image. In the experiments, we have used 80% image captions as training data and 10% as validation data, and the rest 10% is used as test data.

Please note that other datasets, such as RSICD [1], are not suitable for our evaluation, as they do not have multiclass labels.

B. Compared Methods

To verify whether both Transformer-based architecture and auxiliary task-based training provide benefits, we compare the proposed method to both single-task networks and multitask learning with different auxiliary tasks.

Single-task networks compared are as follows: 1) LSTM (C) network, a CNN encoder and LSTM decoder model for image captioning and 2) Transformer (C) network, an encoder–decoder transformer model for image captioning. We also compare a variant of the proposed method with an LSTM-based decoder, LSTM (C + L) network, a model, consisting of a common CNN encoder and two LSTM decoders for image captioning and multilabel classification.

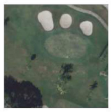
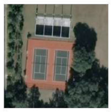
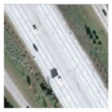
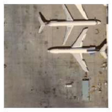
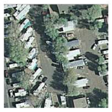
Image	Caption w/o multi-label	Caption with multi-label	Ground truth captions
	green turfs and some bunkers and withered trees in the golf course.	this is a part of a golf course with green turfs and some bunkers and trees.	This is a part of a golf course with green turfs and some bunkers and trees .
	four tennis courts arranged neatly with some plants surrounded.	there are two tennis courts arranged neatly and surrounded by some plants.	There are two tennis courts arranged neatly and surrounded by some plants .
	some cars are on the freeways.	two straight freeways closed to each other with some cars on them.	Two straight freeways parrallel forward with some cars on them .
	an airplane is stopped at the airport.	two airplanes are stopped at the airport.	Two airplanes are stopped at the airport .
	lots of mobile homes with plants surrounded in the mobile home park.	many houses arranged neatly with plants surrounded in the medium residential area.	Many mobile homes are closed to each other with some cars parked at the roadside in the mobile home park .

Fig. 3. Generated captions; the second column lists the captions generated w/o the use of multitask learning; the third column lists the captions generated from the proposed method, and the last column lists the GT captions. Our proposed model generates exact captions as GT for some images, semantically more meaningful captions sentences compared to without multitask learning.

We compare image reconstruction as an auxiliary task instead of the proposed auxiliary task. For this, we use the same architecture as the proposed method, except a CNN decoder for image reconstruction.

We further compare two recently introduced remote sensing image captioning methods: 1) scene attention-based method introduced in [24] and 2) multilabel attention in [5], and Transformer + reinforcement learning-based method [11].

C. Result

The proposed method can obtain meaningful textual descriptions of the remote sensing images, as shown in Fig. 3. A quantitative analysis of the results is tabulated in Table II. Results are shown using different popular indices, Bleu-1, Bleu-2, Bleu-3, Bleu-4, metric for evaluation of translation with explicit ordering (METEOR), recall-oriented understudy for gisting evaluation - longest common subsequence (ROUGE-L), and consensus-based image description evaluation (CIDEr) [25].

LSTM (C + L) outperforms LSTM (C), showing that multilabel classification as an auxiliary task is indeed helpful in improving the captioning, even for simpler LSTM-based architecture. Transformer (C) (i.e., transformer without an auxiliary task) performs similar to LSTM (C + L). Performance drops when using trivial auxiliary tasks, i.e., image reconstruction or angle prediction. This shows that such auxiliary tasks are not consistent with our primary task, i.e., image captioning. However, the proposed method (i.e., using multilabel classification as an auxiliary task) significantly improves the result over LSTM (C), LSTM (C + L), Transformer (C), and scene attention-based method.

TABLE II

QUANTITATIVE COMPARISON OF THE PROPOSED METHOD WITH DIFFERENT EXISTING FRAMEWORKS AND STATE-OF-THE-ART METHODS

Method	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**	**METEOR**	**ROUGE-L**	**CIDEr**
LSTM (C)	0.766	0.688	0.627	0.574	0.388	0.713	2.875
LSTM (C+L)	0.798	0.728	0.670	0.618	0.404	0.744	3.109
Transformer(C)	0.791	0.722	0.670	0.626	0.410	0.736	3.088
Transformer+Reconstruction	0.695	0.612	0.541	0.480	0.348	0.644	2.278
Scene attn	0.822	0.765	0.717	0.674	0.440	0.778	3.228
Multilabel attention+attribute	0.819	0.761	0.712	0.668	0.418	0.757	3.170
Transformer+Reinforcement learning	0.838	0.790	0.744	0.701	0.446	0.776	3.565
Proposed (MLP Decoder)	0.828	0.758	0.703	0.656	0.435	0.766	3.222
Proposed (LSTM Decoder)	0.846	0.785	0.737	0.696	0.444	0.785	3.553

Overall, we observe the following.

1) The Transformer-based model is beneficial compared with the LSTM-based model, as evident from the improved result of Transformer (C) in comparison with LSTM (C).
2) Multilabel classification as an auxiliary task is useful, as evident from the improved result of LSTM (C + L) in comparison with LSTM (C) and improved performance of the proposed model in comparison with Transformer (C).
3) Unsupervised auxiliary tasks, such as rotation, are not suitable for image captioning, as such tasks are semantically different from the primary task of captioning.
4) The proposed method does not need an additional step of reinforcement learning; however, it still obtains similar performance to Transformer + reinforcement learning.

IV. CONCLUSION

We presented a multitask model for remote sensing image captioning. Specifically, we chose multilabel classification as an auxiliary task to improve image captioning. The chosen auxiliary task is semantically similar to our primary task of image captioning. Our experiments show that it helps improve image captioning by outperforming single-task models. Though this is true for any choice of architecture that we may have, we provide evidence to show the superiority of Transformer-based architecture. Our future work will be toward a comprehensive summarizing of remote sensing time series by designing datasets and extending the proposed method for such time series.

REFERENCES

[1] 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.

[2] G. Hoxha, F. Melgani, and B. Demir, “Toward remote sensing image retrieval under a deep image captioning perspective,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4462–4475, 2020.

[3] 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.

[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] Z. Yuan, X. Li, and Q. Wang, “Exploring multi-level attention and semantic relationship for remote sensing image captioning,” *IEEE Access*, vol. 8, pp. 2608–2620, 2020.

[6] A. Vaswani *et al.*, “Attention is all you need,” 2017, *arXiv:1706.03762*.

[7] A. Dosovitskiy *et al.*, “An image is worth 16×16 words: Transformers for image recognition at scale,” 2020, *arXiv:2010.11929*.

[8] S. Zheng *et al.*, “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers,” 2021, *arXiv:2012.15840*.

[9] W. Liu, S. Chen, L. Guo, X. Zhu, and J. Liu, “CPTR: Full transformer network for image captioning,” 2021, *arXiv:2101.10804*.

[10] 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.

[11] X. Shen, B. Liu, Y. Zhou, and J. Zhao, “Remote sensing image caption generation via transformer and reinforcement learning,” *Multimedia Tools Appl.*, vol. 79, nos. 35–36, pp. 26661–26682, Sep. 2020.

[12] B. Shi, J. Hoffman, K. Saenko, T. Darrell, and H. Xu, “Auxiliary task reweighting for minimum-data learning,” 2020, *arXiv:2010.08244*.

[13] S. Saha, Y. T. Solano-Correa, F. Bovolo, and L. Bruzzone, “Unsupervised deep transfer learning-based change detection for HR multispectral images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 856–860, May 2021.

[14] C. Qiu, L. Liebel, L. H. Hughes, M. Schmitt, M. Korner, and X. X. Zhu, “Multitask learning for human settlement extent regression and local climate zone classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[15] P. Mazumder, P. Singh, and V. P. Namboodiri, “Improving few-shot learning using composite rotation based auxiliary task,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2654–2663.

[16] W. Zhao *et al.*, “A multi-task learning approach for image captioning,” in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1205–1211.

[17] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, “Unified vision-language pre-training for image captioning and VQA,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 13041–13049.

[18] J. Xu *et al.*, “Concept detection based on multi-label classification and image captioning approach—DAMO at ImageCLEF 2019,” in *Proc. CLEF*, 2019, pp. 1–10.

[19] S. Saha, F. Bovolo, and L. Bruzzone, “Change detection in image time-series using unsupervised LSTM,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[20] C. Szegedy *et al.*, “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[22] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst. (GIS)*, 2010, pp. 270–279.

[23] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, “Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, Feb. 2018.

[24] S. Wu, X. Zhang, X. Wang, C. Li, and L. Jiao, “Scene attention mechanism for remote sensing image caption generation,” in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–7.

[25] M. Artetxe, G. Labaka, E. Agirre, and K. Cho, “Unsupervised neural machine translation,” 2017, *arXiv:1710.11041*.