REDUNDANCY-AWARE PRETRAINING OF VISION-LANGUAGE FOUNDATION MODELS IN REMOTE SENSING

Mathis Jürgen Adler*[1,2], Leonard Hackel*[1,2], Gencer Sumbul[3], Begüm Demir[1,2]

[1] TU Berlin, Germany [2] BIFOLD, Germany [3] EPFL, Switzerland

***Abstract*—The development of foundation models through pretraining of vision-language models (VLMs) has recently attracted great attention in remote sensing (RS). VLM pretraining aims to learn image and language alignments from a large number of image-text pairs. Each pretraining image is often associated with multiple captions with redundant information due to repeated or semantically similar phrases that result in increased pretraining and inference time of VLMs. To overcome this, we introduce a weighted feature aggregation (WFA) strategy for VLM pretraining in RS. Our strategy aims to extract and exploit complementary information from multiple captions per image, while reducing redundancies through feature aggregation with importance weighting. To calculate adaptive importance weights for different captions of each image, we propose two different techniques: i) non-parametric uniqueness; and ii) learning-based attention. In the first technique, importance weights are calculated based on the bilingual evaluation understudy (BLEU)-scores of the captions to emphasize unique sentences while removing the influence of repetitive sentences. In the second technique, importance weights are learned through an attention mechanism instead of relying on hand-crafted features. The effectiveness of the proposed WFA strategy with the two techniques is analyzed in terms of downstream performance on text-to-image retrieval in RS. Experimental results show that the proposed strategy enables efficient and effective pretraining of VLMs in RS. Based on the experimental analysis, we derive guidelines for the proper selection of techniques, considering downstream task requirements and resource constraints. The code of this work is publicly available at https://git.tu-berlin.de/rsim/redundancy-aware-rs-vlm.**

***Index Terms*—Vision-language models, foundation models, redundancy-aware pretraining, remote sensing.**

I. INTRODUCTION

Inspired by the success of large language model (LLM)-based foundation models (FMs) in natural language processing (NLP) such as LLama [1] and GPT-3 [2], vision-language models (VLMs) in remote sensing (RS) have recently exhibited substantial advancements [3]–[6]. The VLMs learn image-language alignments from a large number of image-text (i.e., image-caption) pairs and are then fine-tuned with a small amount of labeled data. Existing VLMs in RS demonstrate remarkable capability in several image-language tasks, such as image-text retrieval [4], [7], visual question answering [5], zero-shot classification [6] and image captioning [5].

For example, in [4], continual pretraining of a contrastive language-image pretraining (CLIP) model for RS alignment is presented, showing that the RS-specific alignment improves performance over the domain-agnostic CLIP on various RS-specific segmentation, detection and retrieval tasks.

1) Four airplanes are parked at the airport.
2) There are some planes and cars in the airport.
3) Four different kinds of airplanes are in the airport.
4) Four different sizes of airplanes are in the airport.
5) Here are some airplanes and cars in the airport.

Fig. 1: An example image from the UCM-captions dataset [9] with highly redundant (shown in red) and unique (shown in teal) information within its five captions.

In [5], a small-scale, albeit high quality RS image captioning dataset is constructed to fine-tune a VLM for RS. In this work, it is shown that pretraining on small-scale high-quality data can improve the performance of RS-VLMs in downstream tasks (such as captioning and visual question answering) compared to pretraining on large but automatically annotated datasets with lower variety and quality. For a comprehensive overview of VLMs in RS, we refer the reader to [8].

To pretrain VLMs in RS, various datasets have recently been developed and made publicly available [4]–[6]. Most large-scale VLM pretraining datasets include existing image captioning datasets such as UCM-captions [9] and NWPU-Captions [10]. Fig. 1 shows an example image from the UCM-captions dataset that contains five captions per image. The five sentences share almost the same information, with one sentence (sentence 5) containing no unique information at all. Although existing VLMs in RS have demonstrated remarkable potential in aligning image and text pairs, they can not efficiently handle redundant information. A widely used strategy to exploit such data in RS VLM pretraining is to replicate each image as many times as the number of its captions (denoted as Replication strategy in this paper) [3], [4], [6]. We would like to note that, when a model processes all the image caption pairs individually, redundant information has to be processed multiple times. Thus, this strategy results in an increase in the pretraining and inference time of VLMs (without having any benefit in the performance of the downstream tasks).

To address this issue, in this paper we present a weighted

*These authors contributed equally to this work

feature aggregation (WFA) strategy for VLM pretraining in RS. Our strategy exploits complementary information from all captions of an image, while reducing redundancies through feature aggregation with importance weighting. For importance estimation, we introduce two techniques: i) non-parametric uniqueness; and ii) learning-based attention.

We compare both techniques of our WFA strategy with several baseline strategies, and derive some guidelines for their use.

II. Methodology

Let $\mathcal{X} = \{(I_i, \mathcal{C}_i)\}_{i=1}^{N}$ be a pretraining set including N pairs, where I_i is the ith image and $\mathcal{C}_i$ is the set of captions associated to I_i. Each $\mathcal{C}_i$ contains M_i individual captions, i.e., $\mathcal{C}_i = \{c_{i,1}, c_{i,2}, \ldots, c_{i,M_i}\}$, where $c_{i,j}$ is the jth caption that is associated with image I_i. For the sake of simplicity, $\mathcal{C}_i$, I_i, M_i and $c_{i,j}$ are denoted as $\mathcal{C}$, I, M and c_j, respectively, in the rest of the paper. VLM pretraining is generally achieved by CLIP [11] that employs a text encoder $\phi_{\mathcal{C}}$, an image encoder ϕ_I and a projection head ϕ_p. For a given pretraining image I, the image encoder extracts image features $H_I = \phi_I(I)$, while the LLM-based text encoder extracts text features $H_{c_j} = \phi_{\mathcal{C}}(c_j)$ from each caption c_j individually and a projection head employs an additional feature alignment step. The normalized temperature-scaled cross entropy (NT-Xent) loss function $\mathcal{L}_{\text{NTX}}$ [12] is utilized to model image-text alignment.

We assume that the captions in $\mathcal{C}$ can contain not only partially complementary but also redundant information. This increases the complexity of VLM pretraining if I is replicated M times as in the widely used Replication strategy. As an alternative to the Replication strategy, the mean of caption features can be associated with the image feature, where multiple captions contribute equally to the overall representation. However, as shown in Fig. 1, certain captions may be partially or entirely redundant, and thus provide limited descriptive value. To overcome this, we propose a WFA strategy for VLM pretraining that utilizes feature aggregation based on the importance of all captions in $\mathcal{C}$ weighted with respect to their informational content. This is achieved by weighted averaging of text features as follows:

$$H_{\mathcal{C}} = \sum_{j=1}^{M} \alpha_j \cdot H_{c_j}, \tag{1}$$

where $H_{\mathcal{C}}$ is the aggregated text feature and α_j is the weighting factor for caption feature H_{c_j}.

To calculate adaptive importance weights of captions in $\mathcal{C}$ (e.g., α_j for c_j), we propose two techniques: i) non-parametric uniqueness; and ii) learning-based attention. The weighting factor α_j is estimated on the basis of the proposed techniques. By increasing or decreasing the contributions of individual captions, the proposed techniques aim to maximize information content and effectively reduce redundancies. The proposed techniques are explained in detail in the following, and their overview is shown in Fig. 2.

1) Non-Parametric Uniqueness: This technique aims to measure the importance of captions in $\mathcal{C}$ based on their uniqueness compared to each other. We assume that uniqueness of captions can be estimated based on their linguistic distinction to each other. To this end, we utilize the widely-used bilingual evaluation understudy (BLEU) [13] score for measuring linguistic alignment, i.e., the opposite of distinction. In detail, we define a uniqueness score $\mathrm{u}(\mathcal{C}, j)$ for caption c_j based on the BLEU score with respect to all other captions of $\mathcal{C}$. Specifically, we use BLEU-4 for this calculation. The uniqueness of the text c_j is its BLEU-4 score subtracted from one as follows:

$$\mathrm{u}(\mathcal{C}, j) = 1 - \mathrm{BLEU}_4(c_j, \mathcal{C} \setminus c_j). \tag{2}$$

where $\mathrm{BLEU}_4(c_j, \mathcal{C} \setminus c_j)$ is the BLEU-4 score of the jth caption in the set of texts $\mathcal{C}$ against all other texts $\mathcal{C} \setminus c_j$. The result is normalized using the softmax operation to obtain the weighting factors α_j^{u} as:

$$\begin{aligned} \alpha_j^{\mathrm{u}} &= \sigma(\mathbf{u})_j \\ &= \frac{\exp \mathrm{u}(\mathcal{C}, j)}{\sum_{k=1}^{M} \exp \mathrm{u}(\mathcal{C}, k)}. \end{aligned} \tag{3}$$

where $\mathbf{u} = \langle \mathrm{u}(\mathcal{C}, 1), \mathrm{u}(\mathcal{C}, 2), \ldots, \mathrm{u}(\mathcal{C}, M) \rangle$ is the vector of all uniqueness scores. The weighting factors can be used as in (1) to get the aggregated features as follows:

$$\begin{aligned} H_{\mathcal{C}}^{\mathrm{u}} &= \sum_{j=1}^{M} \alpha_j^{\mathrm{u}} \cdot H_{c_j} \\ &= \sum_{j=1}^{M} \sigma(\mathrm{u}(\mathcal{C}, j))_j \cdot H_{c_j}. \end{aligned} \tag{4}$$

Since $\mathrm{u}(\mathcal{C}, j)$ is non-parametric and unlearned, it is computationally efficient to estimate (i.e., it can be calculated offline and re-used in different iterations of the pretraining process).

2) Learning-based Attention: This technique aims to estimate a weighting factor α_j^{att} for caption c_j based on the caption features H_{c_j}. To this end, H_{c_j} is transformed via multiple transformer encoder blocks [14] and an additional linear layer into a pseudo-weight α_j' as follows:

$$\begin{aligned} \alpha_j' &= \mathrm{att}(H_{c_j}) \\ &= \mathrm{FC}\left(\mathrm{TransformerEncoder}\left(H_{c_j}\right)\right), \end{aligned} \tag{5}$$

where $\mathrm{TransformerEncoder}$ is a transformer encoder with two layers as defined in [14] and FC is a fully-connected layer followed by a non-linearity. The pseudo-weights a_j' are calculated for all text embeddings $H_{c_j}, j = 1, 2, \ldots, M$ independently. We would like to note that this technique uses intra-caption attention only to calculate weights. This means that it uses attention on every caption individually. Therefore, this technique is independent of the number of captions within the set of texts for image I and allows for different images to have different numbers of associated captions. The final

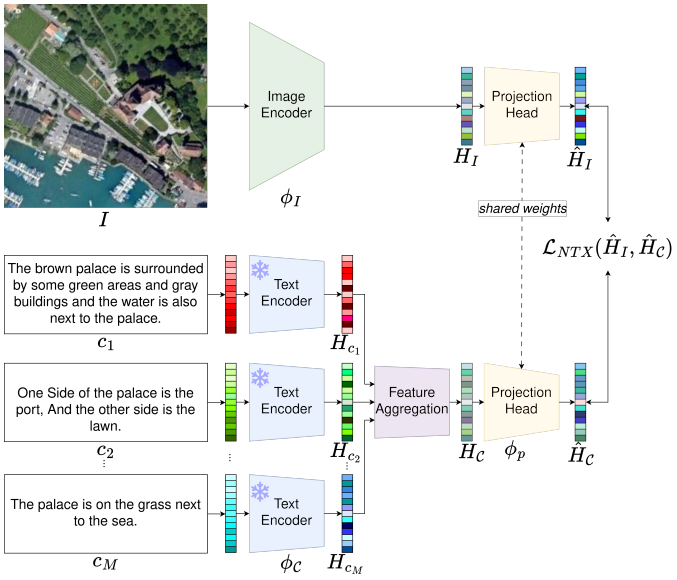

Fig. 2: An illustration of the proposed weighted feature aggregation (WFA) strategy for pretraining the considered VLM (which consists of an image encoder ϕ_I, a text encoder $\phi_\mathcal{C}$, and a projection head ϕ_p shared for image and text) with NT-Xent loss $\mathcal{L}_{\text{NTX}}$. The proposed WFA strategy exploits complementary information from all captions of an image at once, while reducing redundancies through feature aggregation with importance weighting.

weighting factors α_j^{att} are then calculated by applying the softmax operation over all pseudo-weights as follows:

$$\alpha_j^{\text{att}} = \sigma(\boldsymbol{\alpha'})_j, \tag{6}$$

where $\boldsymbol{\alpha'} = \langle \alpha'_1, \alpha'_2, \ldots, \alpha'_M \rangle$ is the vector of all pseudo-weights. Therefore, the aggregated text feature $H_\mathcal{C}^{\text{att}}$ can be obtained as follows:

$$\begin{aligned} H_\mathcal{C}^{\text{att}} &= \sum_{j=1}^{M} \alpha_j^{\text{att}} \cdot H_{c_j} \\ &= \sum_{j=1}^{M} \sigma(\text{att}(H_{c_k})|_{k=1,2,\ldots,M})_j \cdot H_{c_j}. \end{aligned} \tag{7}$$

After obtaining the aggregated text features, they are fed into the projection layer together with the image features as follows: $\hat{H}_\mathcal{C} = \phi_p(H_\mathcal{C})$ and $\hat{H}_I = \phi_p(H_I)$. Then, for pretraining the considered VLM model, the resulting representations are used for image-text alignment through $\mathcal{L}_{\text{NTX}}(\hat{H}_\mathcal{C}, \hat{H}_I)$ [12]. The pretrained VLM model can be used to obtain $\hat{H}_I$ and, depending on the selected technique, $\hat{H}_\mathcal{C}^{\text{u}}$ or $\hat{H}_\mathcal{C}^{\text{att}}$, which can be used for downstream tasks.

III. Experimental Results

In the experiments, for the VLM pretraining, we combine the following four image-text datasets: i) GAIA [15]; ii) RSICD [16]; iii) RSITMD [17]; and iv) NWPU-Captions [10]. The resulting pretraining dataset comprises 88,194 images with 440,970 captions. We divide this dataset into training (90%) and validation (10%) subsets.

As the image encoder of the considered VLM, we utilize the DOFA Base model [19] pretrained on RS data with the input image size of 224×224. As the text encoder, we utilize a small variant of the pretrained bidirectional encoder representations from transformers (BERT) [20] model. For text-level embeddings from BERT's token-level embedding space, we average the sum of the last four hidden states following [21]. We utilize a multi-layer perceptron (MLP)-based projection head with one hidden layer. The text encoder weights are kept frozen during pretraining, while we fine-tune the image encoder, projection head and, for the learned technique, weights for weighted averaging. We pretrain the considered VLM for a maximum of 1000 epochs by using the AdamW [22] optimizer and early stopping with a patience of 5 epochs based on the BLEU-4 score on the validation set. The learning rate, weight decay, temperature and batch size are set to 10^{-5}, 10^{-4}, 0.5, and 200, respectively. VLM pretraining of the considered model takes less than 24 hours on a single NVIDIA H100 GPU.

We compare the results of our WFA strategy with those obtained by five different strategies to exploit various captions in VLM training: i) Replication (which replicates images to match the number of captions); ii) Concatenation (which concatenates multiple captions of an image as a single text); iii) Random Selection (which randomly selects one caption per image), iv) Mean Feature (which averages the text features) [18]; and v) Localized Weight Generation (LG-WF) (which learns the weights to merge text features through a single linear layer) [18]. To the best of our knowledge, all these strategies, except Replication, are applied for the first time in the context of VLMs in RS.

In the experiments, as a downstream task we consider text-to-image retrieval. To this end, we exploit the UCM-captions [9] and Sydney-captions [9] datasets. Each image within the considered datasets is associated with five captions, resulting in 10,500 captions for UCM-captions and 157,500 captions for Sydney-captions. Each caption of each image is selected as a query text, while image retrieval is applied to the respective dataset. Results of each strategy are provided in terms of: i) floating point operations per second (FLOPS), ii) BLEU-4; and iii) mAP scores. The scores are obtained in 10,500 trials performed with 10,500 selected query texts from the UCM-captions dataset, while they are attained in 157,500 trials performed with 157,500 selected query texts from the Sydney-captions dataset. The retrieval performance was assessed on the top-5 and top-20 retrieved images. It is worth noting that the datasets considered in the evaluation were neither used during pretraining nor for fine-tuning. Accordingly, the retrieval performance can be considered as zero-shot performance. We take into account only the inference complexity to calculate the FLOPS for each strategy. It is worth emphasizing that FLOPS are input-size dependent and the considered datasets have varying caption lengths. Therefore, we consider the average caption length for FLOPS calculation in our computational

TABLE I: Text-to-image retrieval results on UCM and Sydney in terms of BLEU-4 (%), mean average precision (mAP) (%) and FLOPS (B) obtained by different VLM strategies. The results are obtained for top-5 and top-20 retrieved images. Learning-based feature aggregation (LFA) indicates whether the considered importance weighting technique introduces additional learned parameters or not. * indicates our strategy and techniques. Best results for LFA/no-LFA are given in bold.

Strategy	**LFA**	**BLEU-4**@5(%) ↑		**mAP**@5(%) ↑		**BLEU-4**@20(%) ↑		**mAP**@20(%) ↑		**FLOPS (B)** ↓
		UCM	Sydney	UCM	Sydney	UCM	Sydney	UCM	Sydney	
Replication	✗	27.8	25.9	49.6	22.9	28.2	26.5	48.6	23.0	171.92
Concatenation	✗	13.6	11.3	30.1	31.3	16.2	8.0	33.0	30.0	34.40
Random Selection	✗	23.8	30.5	44.6	58.7	24.8	28.7	44.6	58.2	**34.38**
Mean Feature [18]	✗	35.6	15.4	62.5	34.9	33.2	11.3	62.4	33.4	35.80
WFA* (Uniqueness)	✗	**40.8**	**40.6**	**67.5**	**88.0**	**42.3**	**41.0**	**69.3**	**85.6**	35.80
LG-WF [18]	✓	38.4	21.0	64.1	42.0	37.8	21.7	65.6	42.1	**35.80**
WFA* (Attention)	✓	**49.9**	**44.2**	**82.0**	**87.6**	**45.8**	**42.8**	**78.2**	**86.3**	36.47

complexity assessment.

Table I shows the results of the proposed WFA with non-parametric uniqueness technique (denoted as WFA (Uniqueness)), the proposed WFA with learning-based attention technique (denoted as WFA (Attention)) and all the other strategies used for the comparison for the UCM and Sydney datasets. In the table, we group the strategies into LFA strategies (which introduce additional learned parameters to the VLM) and no-LFA strategies (which do not introduce additional learned parameters to the VLM). From the table one can observe that WFA (Attention) in general achieves the highest BLEU-4 and mAP scores at the cost of a slight increase in FLOPS for both datasets. For example, it outperforms the Replication strategy by 32.4% and 65.3% in mAP@5 for the UCM and Sydney datasets, respectively. In the case of mAP@20, the improvement with respect to the Replication strategy is 29.6% and 53.3% for the UCM and Sydney datasets, respectively. Among no-LFA approaches, WFA (Uniqueness) demonstrates the best performance. As an example, it yields improvements of 20.7% and 52.6% in mAP@20 compared to Replication for the UCM and Sydney datasets, respectively. In addition, WFA (Uniqueness) also surpasses the LFA-based LG-WF strategy for both datasets. Concerning computational complexity, the Replication strategy has the highest complexity with 171.92 billion FLOPS. This is due to its requirement to apply a forward pass for every sample, which includes the computations for both the image and text encoders. Therefore, to process all five captions that correspond to one image, the image and text encoders have to be used five times each. In contrast, all other strategies including our WFA need only a single forward pass per image input, as the extracted image features can be reused between text inputs. They all require FLOPS only in the range of 34.38 – 36.47 billion.

IV. Conclusion

In this paper, we have presented WFA, a strategy for VLM pretraining in RS that uses feature aggregation of captions based on importance estimation. In addition, we have introduced two importance estimation techniques: i) non-parametric uniqueness; and ii) learning-based attention that calculate adaptive weights for different captions. Our techniques optimize the text inputs during pretraining by enhancing their unique features, while reducing redundant content. Thus, they enable more efficient and effective VLM pretraining in RS. We have compared the proposed WFA strategy and techniques with five pretraining strategies, focusing on the text-to-image retrieval task as a downstream task. The experimental results demonstrate the success of our strategy and techniques. In particular, we would like to highlight that, compared to the Replication strategy (which is widely used for VLM pretraining in RS), our strategy significantly reduces the pretraining and inference time, while improving the text-to-image retrieval accuracy. Based on our experimental analysis, we have derived some guidelines as follows:

1) If sufficient computational resources are available, the WFA strategy with the learning-based attention technique is suggested to be selected, as it is the best-performing approach among all the considered ones.
2) If the compute resources are limited and the pretraining set is highly redundant, the WFA strategy with the Uniqueness technique can be selected. This strategy has lower computational requirements (since it relies on the WFA-weights computed before the training process) and results in a good downstream performance.
3) The widely used Replication strategy shows inferior performance compared to other strategies, while having a high computational cost. Thus, we suggest not to use it.

As future developments of our work, we plan to: i) evaluate the performance of our techniques under different downstream tasks, including visual question answering and captioning; and ii) scale up VLM pretraining with more data and deeper architectures.

References

[1] H. Touvron, T. Lavril, G. Izacard, *et al.*, "Llama: Open and efficient foundation language models.," *arXiv preprint 2302.13971*, 2023.

[2] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[3] Y. Bazi, L. Bashmal, M. M. Al Rahhal, *et al.*, "Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery," *Remote Sensing*, vol. 16, no. 9, p. 1477, 2024.

[4] F. Liu, D. Chen, Z. Guan, *et al.*, "Remoteclip: A vision language foundation model for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.

[5] Y. Hu, J. Yuan, C. Wen, *et al.*, "Rsgpt: A remote sensing vision language model and benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 224, pp. 272–286, 2025.

[6] W. Zhang, M. Cai, T. Zhang, *et al.*, "Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–20, 2024.

[7] X. Zhang, W. Li, X. Wang, *et al.*, "A fusion encoder with multi-task guidance for cross-modal text–image retrieval in remote sensing," *Remote Sensing*, vol. 15, no. 18, p. 4637, 2023.

[8] X. Li, C. Wen, Y. Hu, *et al.*, "Vision-language models in remote sensing: Current progress and future trends," *IEEE Geoscience and Remote Sensing Magazine*, vol. 12, no. 2, pp. 32–66, 2024.

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

[10] Q. Cheng, H. Huang, Y. Xu, *et al.*, "Nwpu-captions dataset and mlca-net for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.

[11] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," *International Conference on Machine Learning*, pp. 8748–8763, 2021.

[12] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," *Advances in Neural Information Processing Systems*, vol. 29, pp. 1857–1865, 2016.

[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311–318, 2002.

[14] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 6000–6010, 2017.

[15] A. Zavras, D. Michail, X. X. Zhu, *et al.*, "GAIA: A global, multi-modal, multi-scale vision-language dataset for remote sensing image analysis," *arXiv preprint 2502.09598*, 2025.

[16] 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.

[17] Z. Yuan, W. Zhang, K. Fu, *et al.*, "Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.

[18] Z. Wang, Y. Wu, K. Narasimhan, and O. Russakovsky, "Multi-query video retrieval," *European Conference on Computer Vision*, pp. 233–249, 2022.

[19] Z. Xiong, Y. Wang, F. Zhang, *et al.*, "Neural plasticity-inspired multimodal foundation model for earth observation," *arXiv preprint 2403.15356*, 2024.

[20] I. Turc, M.-W. Chang, K. Lee, and K. Toutanova, "Well-read students learn better: On the importance of pre-training compact models," *arXiv preprint 1908.08962*, 2019.

[21] G. Mikriukov, M. Ravanbakhsh, and B. Demir, "Unsupervised contrastive hashing for cross-modal retrieval in remote sensing," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4463–4467, 2022.

[22] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *International Conference on Learning Representations*, 2019.