LMMs as Zero-Shot Teachers for Downstream Remote Sensing Classification Tasks*

Maria Tzelepi[1 [0000−0001−5359−8315]], Paraskevi Nousi[2 [0000−0002−3087−3174]], Nikolaos Dimitriou[1 [0000−0002−6650−7758]], and Dimitrios Tzovaras[1 [0000−0001−6915−6722]]

[1] Information Technologies Institute (ITI), Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece
{mtzelepi,nikdim,Dimitrios.Tzovaras}@iti.gr
[2] Swiss Data Science Center (SDSC), ETH Zurich, Zurich, Switzerland
paraskevi.nousi@sdsc.ethz.ch

Abstract. Remote sensing scene classification is a challenging task associated with many critical applications, such as urban planning and environmental conservation. To this end, there is an apparent need for methods that are both accurate and computationally efficient to deal with large volumes of Earth observation data like this. In this paper, we propose a novel Knowledge Distillation (KD) method for addressing remote sensing classification, leveraging Large Multimodal Models (LMM). Specifically, we propose to use an LMM as a *zero-shot teacher* in order to provide the student classifier with soft labels that capture similarities in the input data. To do so, we appropriately prompt the LMM to obtain the class scores which are then transformed to soft-labels. The LMM-generated soft labels aid the student network's training process and lead to improved generalization ability. The effectiveness of the proposed method is validated through extensive experiments in three well-known datasets and using various neural architectures.

Keywords: knowledge distillation · large multimodal models · zero-shot teacher · remote sensing · classification.

1 Introduction

Remote sensing scene classification has become a key element of modern Earth observation, enabling critical applications ranging from urban planning and land use monitoring to disaster response and environmental conservation [17]. To benchmark progress in this field, a wide range of datasets have been published and have since served as standard references for many researchers, providing

* This work has received funding from the EU's Horizon CL4 2024 program under Grant Agreement No 101178230 (ENCIRCLE project). This paper reflects only the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

diverse and challenging scenarios that test the robustness of classification algorithms across varying spatial resolutions and environmental conditions [35, 28, 4].

In the past, traditional Machine Learning (ML) methods such as Random Forests and Support Vector Machines were widely adopted to classify spectral data, as demonstrated in [18, 19]. While effective for smaller datasets, these methods often struggled with generalization due to the high variability and complexity of modern high-resolution imagery. This limitation led to a paradigm shift toward Deep Learning (DL), starting with Convolutional Neural Networks (CNNs) and more recently Visual Transformer (ViT) networks. DL-based methods have significantly improved classification accuracy by learning hierarchical features from raw data, even under challenging conditions such as cloud cover, which constitutes a major challenge in the field [22, 15, 21, 16].

Such models come with a significant drawback of heavy computational costs; to combat these, Knowledge Distillation (KD) has emerged as an effective strategy [8]. As proposed by Hinton et al. [11], KD allows for the transfer of *knowledge*, defined as class probabilities from a large, complex *teacher* model to a smaller, more efficient *student* model. The so-called soft labels provided by the teacher model by increasing the softmax temperature, provide the student model with valuable information as they implicitly reveal similarities over the data, enhancing in this way its generalization capabilities. Over the past years, apart from the response-based KD methods, that use the output of the teacher model as source of knowledge [11, 33, 31], several KD methods have been proposed that investigate other sources of knowledge, such as intermediate layers [20, 10]. Furthermore, several online distillation methods have been proposed where the teacher and student models are trained concurrently without the stage of pretraining the teacher model [2, 25, 25]. Additionally, beyond the traditional strategy of transferring the knowledge of a complex model to a smaller one, another successful line of research includes methods where the knowledge is transferred from teachers to students of identical capacity [7, 13]. Overall, KD has proven particularly successful in classification tasks, enabling the deployment of high-performance models on resource-constrained devices without significant loss of accuracy, and as such lends itself well to the remote sensing classification task.

Recent advances in the field of DL-based classification involve the integration of Large Language Models (LLMs) and Large Multimodal Models (LMMs) [32, 26, 30]. LMMs have recently consistently been demonstrating powerful performance in various downstream vision recognition tasks, overtaking previous DL-based approaches [6, 24]. In a recent work, a Multi-aspect KD framework was proposed that leverages LMMs to guide the training process [14]. Unlike traditional methods that rely solely on visual class logits, this approach utilizes the rich semantic understanding of LMMs to teach student models about various aspects of a scene, such as its spatial arrangement or object interactions. This combination of generative capabilities with visual classification represented a significant leap forward, offering a more nuanced understanding of remote sensing imagery than before.

In this paper, we combine the strengths of KD and LMMs and propose a novel LMM-assisted KD method for remote sensing classification tasks. Specifically, we propose to use LMMs as *zero-shot teachers*, to provide our student networks with soft labels which capture similarities present in the data and which can help guide the classifiers during the student's training process. This means that the proposed KD method does not require the training of a teacher model to acquire the soft labels; instead LMMs are employed in inference-mode to this aim. During test, only the student network is activated. Deep neural networks of different architectures are used, including ones of convolutional nature or the more recent attention-based transformers. Our experimental study showcases the effectiveness of the proposed method on several remote sensing classification benchmarks, highlighting the usefulness of LMMs in the field.

The remainder of the manuscript is organized as follows. Section 2 lists the details of the proposed KD method. Section 3 consists of the experimental study conducted to validate the proposed method, as well as details regarding its implementation. Finally, Section 4 concludes this paper and summarizes its findings.

2 Proposed Method

In this paper we propose a KD method for remote sensing classification tasks, exploiting LMMs as zero-shot teachers. The proposed method consists of two stages, as illustrated in Fig. 1. In the first stage we utilize an LMM (MiniGPT-4 [34] is used in our experiments) as the zero-shot teacher model, i.e., we employ it in inference-mode, in order to obtain the class scores. To do so, we prompt the LMM with each training image of the dataset along with the text query: *The dataset consists of the following classes: [class names]. The image belongs to one class but it may also bear similarities to the other classes. Give a score between 1 and 10 based on the similarity of the image with each of the classes. 10 is for the most similar. Reply only with the name of each class and the score you give (i.e., class: score).* In the subsequent stage the LMM-generated class scores are used to compute the soft labels.

More specifically, we first adapt a pretrained student backbone by introducing a classification head for the considered task. More expressly, considering a C-class remote sensing classification task, and the labeled data $\{\mathbf{x}_i, \mathbf{c}_i\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^D$ is an input vector and D its dimensionality, and $\mathbf{c}_i \in \mathcal{Z}^C$ corresponds to its C-dimensional one-hot class label. For an input space $\mathcal{X} \subseteq \mathbb{R}^D$ and an output space $\mathcal{F} \subseteq \mathbb{R}^C$, we consider as $\Phi(\cdot\,; \mathcal{W}) : \mathcal{X} \rightarrow \mathcal{F}$ the student model, with set of parameters $\mathcal{W}$, which transforms its input vector to a C-dimensional vector. That is, $\Phi(\mathbf{x}_i\,; \mathcal{W}) \in \mathcal{F}$ corresponds to the output vector of $\mathbf{x}_i \in \mathcal{X}$ given by the network Φ with parameters $\mathcal{W}$.

Instead of simply training the student model with hard labels (note that we train only the classification head for computational efficiency), using cross entropy loss, we propose to train it both with hard labels and soft labels using Kullback-Leibler divergence loss. In this way, we aim to enhance the student model's generalization ability, by taking into consideration the similarities en-

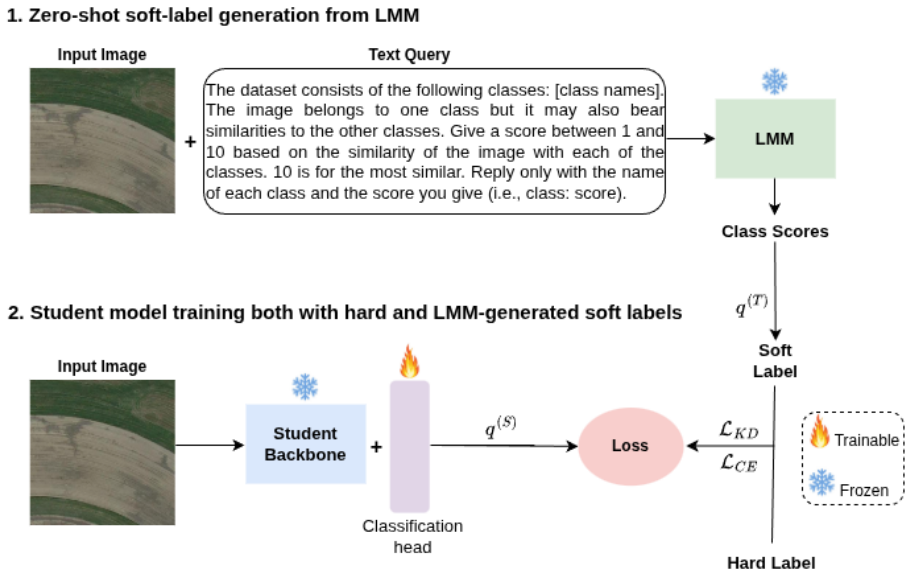


Fig. 1: Proposed method: In the first stage we prompt the LMM with the training images along with the depicted text query, in order to obtain the scores for each input image belonging to each of the classes of the dataset. In the second stage we first compute the soft labels from the LMM-generated scores using the softmax operation. Subsequently, we adapt the student model by introducing a classification head for the considered task, and with frozen the student's model backbone, we train the classification head both with hard labels using cross entropy loss, and the computed soft labels using the Kullback-Leibler divergence loss.

coded in the soft-labels. To do so, as mentioned, we first compute the soft labels by applying the softmax operation on the LMM-generated class scores $\mathbf{z}_i^{(T)}$. Formally expressed:

$$q_{ik}^{(T)} = \frac{\exp\left(\frac{z_{ik}^{(T)}}{\tau}\right)}{\sum_{j=1}^{C} \exp\left(\frac{z_{ij}^{(T)}}{\tau}\right)}, \tag{1}$$

where τ corresponds to the temperature parameter, $i = 1, \ldots, N$ corresponds to the i-th sample and $k = 1, \ldots, C$ corresponds to the k-th class. Correspondingly

the student's model softmax probabilities are computed as follows:

$$q_{ik}^{(S)} = \frac{\exp\left(\frac{z_{ik}^{(S)}}{\tau}\right)}{\sum_{j=1}^{C} \exp\left(\frac{z_{ij}^{(S)}}{\tau}\right)}. \tag{2}$$

where $\mathbf{z}_i^{(S)} = \Phi(\mathbf{x}_i; \mathcal{W}) \in \mathbb{R}^C$ are the output scores (logits) of the student model.

Subsequently, we train the student model both with the cross entropy loss, $\mathcal{L}_{\text{CE}}$ and the distillation loss, $\mathcal{L}_{\text{KD}}$ (using Kullback-Leibler divergence). Cross entropy loss is defined as follows:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N}\sum_{k=1}^{C} c_{ik} \log q_{ik}^{(S)}. \tag{3}$$

The distillation loss is computed as follows:

$$\mathcal{L}_{\text{KD}} = \frac{\tau^2}{N}\sum_{i=1}^{N}\sum_{k=1}^{C} q_{ik}^{(T)} \log\left(\frac{q_{ik}^{(T)}}{q_{ik}^{(S)}}\right). \tag{4}$$

The total loss for training the student model is computed as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{KD}}, \tag{5}$$

where the parameter λ controls the contribution of the two loss components.

During the test phase, the test images are propagated to the trained student model and the class prediction are produced through a forward pass. This means that the LMM-generated soft labels can be computed once only for the training images, offline, before the training stage of the student network begins and are not required for the test phase. Thus, the method runs as fast as the underlying student network does, while providing improved classification accuracy.

3 Experimental Evaluation

In this section we present the experiments performed in order to validate the performance of the proposed KD method. We first present the utilized datasets and evaluation metrics, followed by the utilized LMM and downstream models' descriptions, as well as the implementation details. Finally, the experimental results are provided.

3.1 Datasets and Evaluation Metrics

We use three datasets to evaluate the performance of the proposed KD method: RSSCN7 [35], UCMerced LandUse [28], and WHU-RS19-1 [4]. RSSCN7 contains 2247 training images and 567 test images divided into 7 classes. UCMerced LandUse contains 1701 training images and 441 test images divided into 21 classes.

WHU-RS19-1 contains 779 training images and 209 test images divided into 19 classes. We use test accuracy as evaluation metric (i.e., the ratio of correct predictions to the total number of predictions). We also provide qualitative results of the LMM-generated soft-labels.

3.2 Models

Large Multimodal Model GPT-4 [1] was the first multimodal language model, which could process both textual and visual inputs, and generate textual output. MiniGPT-4 [34] was soon thereafter published, as an open-source, lightweight alternative, as the technical details of GPT-4 remain unpublished. The architecture of MiniGPT-4 relies on a projection layer to synchronize a pretrained, frozen visual encoder with a frozen LLM. Specifically, it uses Vicuna LLM [3], and a ViT-G/14 [29] from EVA-CLIP [23] and a Q-Former network for its visual perception. In this paper we use MiniGPT-4 in the first stage of the proposed method for obtaining the class scores; however it should be noted that any other LMM, general-purpose or specialized toward the remote sensing task, can be utilized.

Downstream Models We use three models of varying complexity as downstream models in the experimental evaluation of the proposed method: ResNet-18 (11.7M) [9], TinyViT (5.7M) [27], MobileNet-v3-small (2.5M) [12]. All the models are pre-trained in the ImageNet dataset. [5] Keeping the backbone weights frozen, we introduce a trainable classification head (i.e., single linear layer), with neurons equal to the number of the classes for each dataset.

3.3 Implementation Details

All the models are trained using stochastic gradient descent, with the learning rate set to 0.001, for 100 epochs. The batch size is set to 32. The temperature in eq. (4) is set to 1. The parameter λ in eq. (5) is set to 0.01. The experiments conducted using an NVIDIA RTX A4500 with 20GB of GPU memory.

3.4 Experimental Results

In this section we provide the experimental results for validating the effectiveness of the proposed KD method. We compare the performance of the proposed method against the baseline training of the student model only with the hard labels, using the same experimental setup. Best results are printed in bold. In Table 1 we provide the experimental results on all the utilized datasets using the ResNet-18 student model. As it is demonstrated, the proposed method significantly improves the performance in two out three cases, while it does not provide improvements on the WHU-RS19-1 dataset. It should be noted that, as mentioned, in the conducted experiments we set the parameter of temperature $\tau = 1$ for consistency, since temperature tuning can be tedious [11]. However,

the optimal temperature value is dataset-dependent. For example, in the case of WHU-RS19-1 we can benefit from a higher temperature value; indeed, we achieve test accuracy 94.74% against 94.21% by setting $\tau = 2$.

Correspondingly, In Table 2 we provide the evaluation results using the TinyViT model, while in Table 3 the evaluation results using the MobileNet-v3-small. As it can be observed, the proposed KD method significantly improves the baseline training on all the considered cases. Furthermore, in Fig. 2 we provide test accuracy throughout the training epochs using the MobileNet-v3-small model on all the utilized cases, where the steadily improved performance of the proposed KD method against the baseline is validated.

Table 1: Evaluation results of the proposed KD method against baseline, in terms of test accuracy (%), on the all utilized datasets using the ResNet-18 model.

Method	**RSSCN7**	**UCMerced LandUse**	**WHU-RS19-1**
Baseline	85.89	92.38	94.21
Proposed	**87.50**	**93.57**	94.21

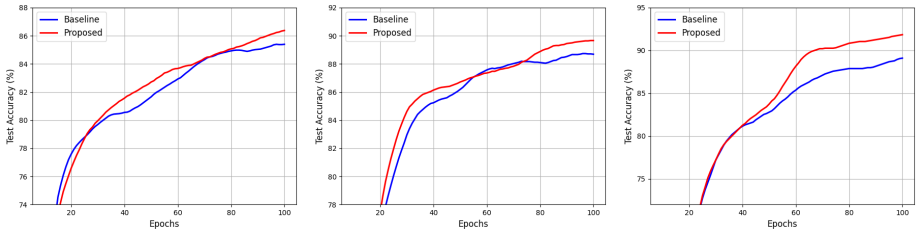
Table 2: Evaluation results of the proposed KD method against baseline, in terms of test accuracy (%), on the all utilized datasets using the TinyViT model.

Method	**RSSCN7**	**UCMerced LandUse**	**WHU-RS19-1**
Baseline	86.61	90.24	90.53
Proposed	**86.96**	**91.43**	**92.63**

Table 3: Evaluation results of the proposed KD method against baseline, in terms of test accuracy (%), on the all utilized datasets using the MobileNet-v3-small model.

Method	**RSSCN7**	**UCMerced LandUse**	**WHU-RS19-1**
Baseline	85.54	89.05	89.95
Proposed	**86.61**	**89.76**	**92.11**

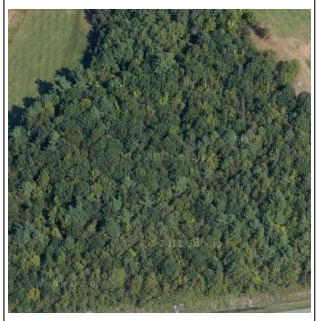
Finally, in Table 4 we provide some qualitative results on the LMM-generated soft labels. It is evident that using an LMM as a zero-shot teacher can produce meaningful soft labels that, as experimentally validated, improve the generalization ability of the student model.



(a) RSSCN7 (b) UCMerced LandUse (c) WHU-RS19-1

Fig. 2: Test accuracy throughout training epochs using the MobileNet-v3-small model.

Table 4: Qualitative Results: LMM-Generated Soft Labels. The ground-truth class for each image is underlined.

Image	**Soft Label**	**Image**	**Soft Label**
	Field: 0.23 Forest: 0.63 Grass: 0.08 Industry: 0.03 Parking: 0.01 Resident: 0.01 River: 0.01		Field: 0.12 Forest: 0.02 Grass: 0.04 Industry: 0.36 Parking: 0.22 Resident: 0.22 River: 0.02
	Field: 0.10 Forest: 0.04 Grass: 0.26 Industry: 0.01 Parking: 0.01 Resident: 0.15 River: 0.43		Field: 0.23 Forest: 0.03 Grass: 0.08 Industry: 0.01 Parking: 0.01 Resident: 0.01 River: 0.63

4 Conclusions

In this paper, we have proposed a novel LMM-based KD method for remote sensing classification, where the LMM acts as the teacher network to provide the student classifier with soft labels which capture similarities in the input data. The soft labels aid the student network's training process and lead to improved classification scores, as demonstrated in three well-known datasets and using three different neural architectures. The proposed method is fast, only requiring the LMM to generated soft labels once for the training set, which can be performed in an offline matter. The student network is trained to match these soft labels, with a double objective of cross entropy and KL-divergence. The result is an effectively trained efficient student network with improved accuracy over its non-KD baseline. The integration of LMMs in our framework also showcases the usefulness of such models in this field, and serves as a baseline method for future work in this direction.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Chen, D., Mei, J.P., Wang, C., Feng, Y., Chen, C.: Online knowledge distillation with diverse peers. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 3430–3437 (2020)
3. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), https://lmsys.org/blog/2023-03-30-vicuna/
4. Dai, D., Yang, W.: Satellite image classification via two-layer sparse coding with biased image representation. IEEE Geoscience and remote sensing letters **8**(1), 173–176 (2010)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Fu, D., Li, X., Wen, L., Dou, M., Cai, P., Shi, B., Qiao, Y.: Drive like a human: Rethinking autonomous driving with large language models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 910–919 (2024)
7. Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., Anandkumar, A.: Born again neural networks. In: International conference on machine learning. pp. 1607–1616. PMLR (2018)
8. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International journal of computer vision **129**(6), 1789–1819 (2021)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Heo, B., Lee, M., Yun, S., Choi, J.Y.: Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 3779–3787 (2019)

11. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
12. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1314–1324 (2019)
13. Lan, X., Zhu, X., Gong, S.: Self-referenced deep learning. In: Asian conference on computer vision. pp. 284–300. Springer (2018)
14. Lee, T., Bang, J., Kwon, S., Kim, T.: Multi-aspect knowledge distillation with large language model. arXiv preprint arXiv:2501.13341 (2025)
15. Li, Y., Zhang, H., Xue, X., Jiang, Y., Shen, Q.: Deep learning for remote sensing image classification: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8**(6), e1264 (2018)
16. Lu, X., Zheng, X., Yuan, Y.: Remote sensing scene classification by unsupervised representation learning. IEEE Transactions on Geoscience and Remote Sensing **55**(9), 5148–5157 (2017)
17. Mehmood, M., Shahzad, A., Zafar, B., Shabbir, A., Ali, N.: Remote sensing image classification: A comprehensive review and applications. Mathematical problems in engineering **2022**(1), 5880959 (2022)
18. Pal, M.: Random forest classifier for remote sensing classification. International journal of remote sensing **26**(1), 217–222 (2005)
19. Pal, M., Mather, P.M.: Support vector machines for classification in remote sensing. International journal of remote sensing **26**(5), 1007–1011 (2005)
20. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: hints for thin deep nets (2014). arXiv preprint arXiv:1412.6550 **3** (2014)
21. Song, S., Yu, H., Miao, Z., Zhang, Q., Lin, Y., Wang, S.: Domain adaptation for convolutional neural networks-based remote sensing scene classification. IEEE Geoscience and Remote Sensing Letters **16**(8), 1324–1328 (2019)
22. Sun, H., Lin, Y., Zou, Q., Song, S., Fang, J., Yu, H.: Convolutional neural networks based remote sensing scene classification under clear and cloudy environments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 713–720 (2021)
23. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023)
24. Tzelepi, M., Mezaris, V.: Improving multimodal hateful meme detection exploiting lmm-generated knowledge. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 202–211 (2025)
25. Tzelepi, M., Tefas, A.: Efficient training of lightweight neural networks using online self-acquired knowledge distillation. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2021)
26. Wu, J., Gan, W., Chen, Z., Wan, S., Yu, P.S.: Multimodal large language models: A survey. In: 2023 IEEE International Conference on Big Data (BigData). pp. 2247–2256. IEEE (2023)
27. Wu, K.T., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., Tinyvit, L.Y.: Fast pretraining distillation for small vision transformers/k. Wu, J. Zhang, H. Peng, M. Liu, B. Xiao, J. Fu, L. Yuan. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **13681** (2022)
28. Yang, Y., Newsam, S.: 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. pp. 270–279 (2010)

29. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12104–12113 (2022)
30. Zhang, D., Yu, Y., Dong, J., Li, C., Su, D., Chu, C., Yu, D.: Mm-llms: Recent advances in multimodal large language models. arXiv preprint arXiv:2401.13601 (2024)
31. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 11953–11962 (2022)
32. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 **1**(2) (2023)
33. Zhou, H., Song, L., Chen, J., Zhou, Y., Wang, G., Yuan, J., Zhang, Q.: Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. arXiv preprint arXiv:2102.00650 (2021)
34. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
35. Zou, Q., Ni, L., Zhang, T., Wang, Q.: Deep learning based feature selection for remote sensing scene classification. IEEE Geoscience and remote sensing letters **12**(11), 2321–2325 (2015)