

Enhancing Remote Sensing Vision-Language Models for Zero-Shot Scene Classification

Karim El Khoury^{*1} Maxime Zanella^{*1,2} Benoît Gérin^{*1} Tiffanie Godelaine^{*1}
 Benoît Macq¹ Saïd Mahmoudi² Christophe De Vleeschouwer¹ Ismail Ben Ayed³

¹UCLouvain, Belgium ²UMons, Belgium ³ÉTS Montreal, Canada

Abstract—Vision-Language Models for remote sensing have shown promising uses thanks to their extensive pretraining. However, their conventional usage in zero-shot scene classification methods still involves dividing large images into patches and making independent predictions, i.e., inductive inference, thereby limiting their effectiveness by ignoring valuable contextual information. Our approach tackles this issue by utilizing initial predictions based on text prompting and patch affinity relationships from the image encoder to enhance zero-shot capabilities through transductive inference, all without the need for supervision and at a minor computational cost. Experiments on 10 remote sensing datasets with state-of-the-art Vision-Language Models demonstrate significant accuracy improvements over inductive zero-shot classification. Our source code is publicly available on Github: <https://github.com/elkhouryk/RS-TransCLIP>

Index Terms—remote sensing, scene classification, vision-language models, zero-shot, transductive inference

I. INTRODUCTION

Remote Sensing (RS) imagery has become an effective tool for monitoring the surface of the Earth. It has given rise to several applications, ranging from environmental monitoring [1, 2], to precision agriculture [3, 4], as well as emergency disaster response [5, 6]. All of these tasks require precise and quick scene classification to extract useful insights from highly complex visual data.

Linking images with text descriptions has been an effective approach for learning granular visual representations [7, 8]. While this idea seemed powerful, pioneering works in the field of RS [9, 10] were limited by computational budgets and the quantity of available RS data, both of which have been significant bottlenecks for generalization and robustness capabilities [11]. More recently, Vision-Language Models (VLMs) like CLIP [12] have overcome these limitations by leveraging a new pretraining paradigm that uses large-scale image-text pair datasets for unsupervised contrastive learning. These models have demonstrated high capability for numerous downstream tasks, including efficient zero-shot image classification by prompting arbitrary candidate class descriptions, e.g., "a satellite photo of a [class].", sometimes even surpassing supervised competitors [12]. Inspired by these

^{*} The authors have contributed equally to this work.

Acknowledgments – M.Z. and B.G. are funded by the Walloon region under grant No. 2010235 (ARIAC by DIGITALWALLONIA4.AI). T.G. is funded by MedReSyst part of the Walloon Region and EU-Wallonie 2021-2027 program.

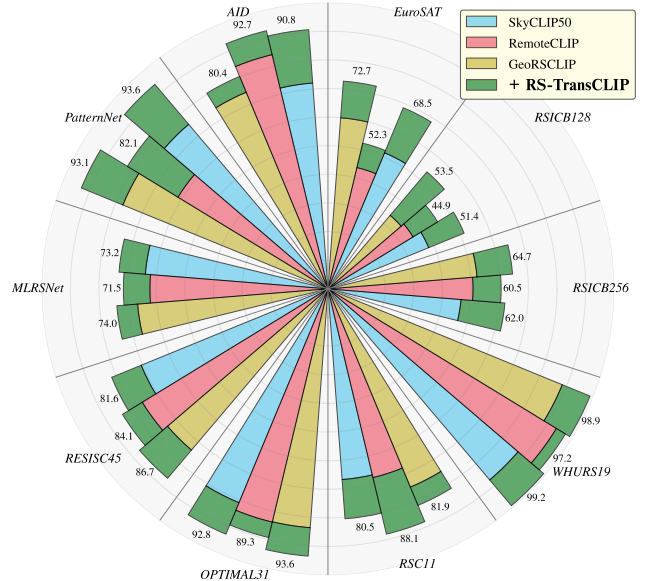


Fig. 1: Top-1 accuracy of RS-TransCLIP, on ViT-L/14 RS VLMs, for zero-shot scene classification across 10 datasets.

promising results, the RS community has worked on developing large image-text RS datasets [13–16] leading to rapid progress in zero-shot scene classification benchmarks [17].

In remote sensing scene classification, both the large size of the images and the need for granular information pose challenges. To make high-resolution inference tractable, it is common practice to divide the images into smaller patches and generate predictions for each patch individually; this is known as *inductive* inference. Another paradigm known as *transductive* inference [18, 19], has shown that jointly considering multiple instances at prediction time can improve the prediction accuracy by accounting for the statistical distribution of instances in the embedding space [20, 21]. Despite its large potential, *transductive* inference has been largely overlooked in RS within the context of VLMs. We aim to address this gap by introducing an efficient transductive method that operates exclusively within the embedding space, i.e., in a black box setup after feature extraction.

In a zero-shot classification setting, class-specific textual prompts are mapped to a shared embedding space generating

individual pseudo-label for each image patch. In a traditional *inductive* inference process, predictions are generated by utilizing initial pseudo-labels to identify the most confident class, with each patch predicted individually. In contrast, our work envisions *transductive* inference in a zero-shot classification setting. As shown in Fig. 2, this approach leverages the data structure within the feature space to account for instance relations, enabling collective prediction of all points simultaneously. Our proposed objective function can be viewed as a regularized maximum-likelihood estimation, constrained by a Kullback-Leibler divergence penalty that integrates the aforementioned initial pseudo-labels and a Laplacian term that constraints similar patches to have similar predictions.

Contribution: We introduce RS-TransCLIP, a transductive algorithm that enhances RS VLMs without requiring any labels, only incurring a negligible computational cost to the overall inference time. Fig. 1 highlights the significant boost that RS-TransCLIP offers on state-of-the-art RS VLMs.

II. RELATED WORK

A. Vision-Language Models for Remote Sensing

Due to foundation models being trained on natural images, there is an active research effort to build domain-specific versions of these models. This is prevalent in the medical imaging where VLMs have shown promising results [22, 23] in improving image-text retrieval and few-shot classification. The RS community has followed suit, working on creating extensive image-text datasets by scraping and filtering public satellite and UAV imagery sources [13–16]. This has led to the development of several fine-tuned VLMs on various downstream tasks [24–28], with many of them showing strong performances in zero-shot scene classification [11, 13, 15].

B. Transductive inference in Vision-Language Models

In the few-shot literature, transduction leverages both the few labeled samples and unlabeled test data outperforming inductive methods [29–32]. However, when applied to VLMs, these transductive methods face significant performance drops [20, 21] since they are based solely on the vision features. This motivated very recent transductive methods in computer vision to explicitly leverage the textual modality alongside image embeddings – a capability not present before the emergence of VLMs [20, 21, 33]. Building on these advances and the transductive-inference zero-shot objective described in [21], our work enhances the predictive accuracy of pretrained RS VLMs without the need of any supervision.

III. METHOD

The transductive approach employed by RS-TransCLIP is based on the hypothesis that the data structure within the feature space can be modeled as a mixture of Gaussian distributions. As a result, the RS-TransCLIP objective function integrates this hypothesis alongside affinity relationships among patches and initial text-based pseudo-labels to minimize prediction deviation. The intuition behind the proposed transductive approach is depicted in Fig. 2.

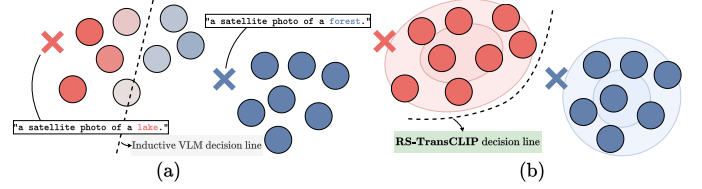


Fig. 2: (a) VLMs assign each image to its closest text embedding and (b) RS-TransCLIP exploits the image-text structure to enhance the predictions without any additional labels.

A. Variable Definition

In an *inductive* approach, predictions are made individually using only the initial pseudo-label $\hat{\mathbf{y}}$. Conversely, in the proposed *transductive* approach, predictions are made simultaneously by modeling the feature space using three variables, $\mathbf{z}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ which can be split into two categories:

Assignment variables — where $\mathbf{z}$ is defined as:

$$\mathbf{z}_i = (z_{i,k})_{1 \leq k \leq K} \in \Delta_K, \quad \forall i \in \mathcal{Q}$$

with K the number of classes, $\mathcal{Q}$ the sample indices set and Δ_K the K -dimensional probability simplex (prediction space).

Gaussian Mixture Model (GMM) variables — where the mean $\boldsymbol{\mu}$ and the covariance $\boldsymbol{\Sigma}$ are defined as:

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_k \in \mathbb{R}^d)_{1 \leq k \leq K} \quad \boldsymbol{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_d)$$

with d the embedding dimension. Note that $\boldsymbol{\Sigma}$ is shared among classes to decrease the number of parameters.

B. RS-TransCLIP objective function

The goal is to minimize the objective function $\mathcal{L}$ composed of three terms: an unsupervised *GMM clustering* term, an affinity-based *Laplacian regularization* term and a divergence-driven *Kullback-Leibler (KL) regularization* term. The terms of $\mathcal{L}$, written in Eq. (1), are detailed hereafter:

$$\begin{aligned} \mathcal{L}(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = & \underbrace{-\frac{1}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} \mathbf{z}_i^\top \log(\mathbf{p}_i)}_{\text{GMM clustering}} \\ & \underbrace{- \sum_{i \in \mathcal{Q}} \sum_{j \in \mathcal{Q}} w_{ij} \mathbf{z}_i^\top \mathbf{z}_j}_{\text{Laplacian regularization}} + \underbrace{\sum_{i \in \mathcal{Q}} \text{KL}(\mathbf{z}_i || \hat{\mathbf{y}}_i)}_{\text{KL regularization}} \end{aligned} \quad (1)$$

GMM clustering — The goal of this term is maximizing the similarity between the assignment variables $\mathbf{z}_i$ and the likelihood $\mathbf{p}_i$. In our case, we model the likelihood of target data as a balanced mixture of K multivariate Gaussian distributions. Each distribution represents a class k with an associated mean vector $\boldsymbol{\mu}_k$ and a covariance matrix $\boldsymbol{\Sigma}$. Defining $\mathbf{f}_i \in \mathbb{R}^d$ as the image embedding of sample i , we set $p_{i,k}$ the probability that sample i is generated by the Gaussian distribution of class k :

$$p_{i,k} \propto \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{f}_i - \boldsymbol{\mu}_k)^\top \Sigma^{-1}(\mathbf{f}_i - \boldsymbol{\mu}_k)\right)$$

Laplacian regularization — The aim of this term is to favor pairs of samples with high affinity to have similar assignment variables. In our case, we define non-negative affinities w_{ij} using the cosine similarities between image embeddings of each sample (see line 2 in Algorithm 1). Note that affinities can be tailored for each specific use-case, provided the affinity matrix is positive semi-definite. This ensures the concavity of the term, which in turn guarantees the convergence of the decoupled updates (refer to [21] for details).

KL regularization — The purpose of this term is to prevent the assignment variables to deviate significantly from the initial pseudo-labels. In our case, we obtain the pseudo-labels $(\hat{\mathbf{y}}_i)_{1 \leq i \leq Q}$ by applying the softmax function to the vector whose components are obtained by computing the dot product between the image embeddings $\mathbf{f}_i$ and all text embeddings $\mathbf{t}_k \in \mathbb{R}^d$, scaled by the temperature factor τ used during VLM pretraining (see line 1 in Algorithm 1). This allows us to integrate the text-knowledge into the optimization process.

C. Solving procedure

We refer to [21] for the derivation and optimization details of the convergence procedure for the objective function $\mathcal{L}$. The pseudo-code for the RS-TransCLIP procedure is outlined in Algorithm 1. Note that image and text embeddings are only computed once at the start. After the affinity w_{ij} and the pseudo-labels $\hat{\mathbf{y}}_i$ are determined, the assignment variables $\mathbf{z}_i$ and the GMM variables $\boldsymbol{\mu}_k$ and Σ are then initialized and updated, according to the update rules listed in Eq. (2), (3) and (4) respectively. The update rules vary depending on the two variable categories:

Iterative decoupled updates. — The assignment variable $\mathbf{z}_i$ is updated at each iteration l as it depends on its neighbors $\mathbf{z}_j$. Note that the update rule of $\mathbf{z}_i$ can be parallelized, which makes the convergence procedure computationally efficient.

$$\mathbf{z}_i^{(l+1)} = \frac{\hat{\mathbf{y}}_i \odot \exp(\log(\mathbf{p}_i) + \sum_{j \in Q} w_{ij} \mathbf{z}_j^{(l)})}{(\hat{\mathbf{y}}_i \odot \exp(\log(\mathbf{p}_i) + \sum_{j \in Q} w_{ij} \mathbf{z}_j^{(l)}))^\top \mathbb{1}_K} \quad (2)$$

Closed-form updates. — With $\mathbf{z}_i$ fixed, obtained following the iterative decoupled updates, we can calculate the closed-form updates for GMM variables $\boldsymbol{\mu}_k$ and Σ .

$$\boldsymbol{\mu}_k = \frac{\sum_{i \in Q} z_{i,k} \mathbf{f}_i}{\sum_{i \in Q} z_{i,k}} \quad (3)$$

$$\text{diag}(\Sigma) = \frac{\sum_{i \in Q} \sum_k z_{i,k} (\mathbf{f}_i - \boldsymbol{\mu}_k)^2}{|Q|} \quad (4)$$

Algorithm 1: RS-TransCLIP procedure

Input: $\mathbf{f}, \mathbf{t}, \tau$
1 $\hat{\mathbf{y}}_i \leftarrow \text{softmax}(\tau \mathbf{f}_i^\top \mathbf{t}) \quad \forall i;$
2 $w_{ij} = \mathbf{f}_i^\top \mathbf{f}_j \quad \forall i, j;$
3 $\mathbf{z}_i \leftarrow \hat{\mathbf{y}}_i \quad \forall i;$
4 Initialize $\boldsymbol{\mu}_k \quad \forall k$, and $\text{diag}(\Sigma)$; ▷ See *
5 **while not converged do**
6 // *Iterative decoupled updates*
7 **for** $l = 1 : \dots$ **do**
8 Update $\mathbf{z}_i^{(l+1)} \quad \forall i$; ▷ See Eq. (2)
9 **end**
10 // *Closed-form updates*
11 Update $\boldsymbol{\mu}_k \quad \forall k$; ▷ See Eq. (3)
12 Update $\text{diag}(\Sigma)$; ▷ See Eq. (4)
13 **end**
14 **return** $\mathbf{z}$

* $\boldsymbol{\mu}_k$ is initialized by averaging the image embeddings of the 8 most confident samples according to the pseudo-labels, while $\text{diag}(\Sigma)$ is initialized by setting each element to $1/d$.

IV. EXPERIMENTS

A. Experimental setup

We test RS-TransCLIP on four VLMs: CLIP [12], RemoteCLIP [11], SkyCLIP [15], and GeoRSCLIP [13] — all with various model architectures to generate their respective image embeddings. Using RS text-prompt templates from [13], 106 individual text embeddings were averaged out to get a single textual embedding per class. The zero-shot scene classification performance is evaluated on 10 RS benchmark datasets: AID, EuroSAT, MLRSNet, OPTIMAL31, PatternNet, RESISC45, RSC11, RSICB128, RSICB256, and WHURS19 [34–42]. Note that none of the chosen VLMs were fine-tuned on any of the listed datasets. TABLE I presents the zero-shot top-1 accuracy, *without* and *with* the addition of RS-TransCLIP.

B. Zero-shot classification — without RS-TransCLIP

First, we assess the top-1 accuracy *without* RS-TransCLIP, evaluating it in an *inductive* inference scenario based on the initial pseudo-labels $\hat{\mathbf{y}}_i$ (see line 1 in Algorithm 1). We notice that for smaller backbones like ViT-B/32, RemoteCLIP, GeoRSCLIP and SkyCLIP50 outperform CLIP. However, for larger backbones like ViT-L/14, CLIP is surprisingly competitive on various benchmarks in comparison to the RS VLMs. A clear trend of larger models performing better indicates promising potential in scaling both model and dataset sizes.

C. Zero-shot classification — with RS-TransCLIP

Second, we observe the top-1 accuracy *with* RS-TransCLIP, evaluating it in a *transductive* inference scenario based on the obtained assignment variables $\mathbf{z}_i$ when solving $\mathcal{L}$ (see Algorithm 1). We can clearly see a massive performance improvement across all benchmarks and models. We find that the addition of RS-TransCLIP provides average gains ranging from 9.9% up to 17.1% across all benchmarks and models.

TABLE I: Top-1 accuracy for zero-shot scene classification without (white) and with (blue) RS-TransCLIP on 10 RS datasets.

	Model	AID	EuroSAT	MLRSNet	OPTIMAL31	PatternNet	RESISC45	RSC11	RSICB128	RSICB256	WHURS19	Average
ResNet-50	CLIP	55.4	28.3	45.0	64.5	46.4	52.8	56.7	23.4	30.4	71.3	47.4
	+ RS-TransCLIP	69.6	48.1	54.2	79.6	69.0	69.6	77.8	34.3	46.8	95.9	64.5
	Δ	+14.2	+19.8	+9.3	+15.2	+22.6	+16.7	+21.0	+10.8	+16.4	+24.6	+17.1
	RemoteCLIP	89.1	26.7	43.0	64.0	43.6	51.6	67.0	15.0	36.4	95.4	53.2
	+ RS-TransCLIP	93.3	34.4	58.0	85.0	53.6	72.9	87.2	19.1	48.2	98.4	65.0
	Δ	+4.2	+7.8	+15.0	+21.0	+10.0	+21.2	+20.2	+4.1	+11.8	+3.0	+11.8
ViT-B/32	CLIP	66.4	45.3	51.2	73.0	59.6	60.7	55.5	27.7	40.3	81.1	56.1
	+ RS-TransCLIP	80.7	49.0	64.2	82.9	76.6	74.1	67.0	33.2	46.4	90.3	66.5
	Δ	+14.3	+3.6	+13.0	+9.9	+16.9	+13.4	+11.5	+5.6	+6.0	+9.3	+10.4
	GeoRSCLIP	70.3	53.4	65.0	79.6	75.8	68.8	68.3	29.0	46.5	88.8	64.5
	+ RS-TransCLIP	78.2	69.0	71.9	87.3	94.5	79.5	78.6	42.8	61.8	98.7	76.2
	Δ	+7.9	+15.5	+6.9	+7.7	+18.6	+10.7	+10.3	+13.8	+15.3	+10.0	+11.7
	RemoteCLIP	91.7	35.5	56.3	77.6	55.9	68.1	61.8	26.0	41.5	95.2	61.0
	+ RS-TransCLIP	95.6	51.0	65.8	87.8	70.7	79.4	79.7	31.1	49.2	97.9	70.8
	Δ	+3.9	+15.5	+9.5	+10.3	+14.8	+11.2	+17.9	+5.1	+7.7	+2.7	+9.9
	SkyCLIP50	70.3	52.6	63.2	79.5	73.8	66.7	61.2	39.0	47.1	91.0	64.5
	+ RS-TransCLIP	78.7	64.5	73.2	85.2	87.6	77.3	77.1	49.4	59.1	97.8	75.0
	Δ	+8.3	+11.9	+10.1	+5.8	+13.8	+10.6	+15.9	+10.4	+11.9	+6.8	+10.5
ViT-L/14	CLIP	69.7	60.1	64.1	80.6	74.7	71.3	67.3	37.9	47.2	85.5	65.8
	+ RS-TransCLIP	84.2	71.9	74.5	92.4	91.8	82.2	80.5	43.9	50.5	99.1	77.1
	Δ	+14.4	+11.9	+10.4	+11.7	+17.1	+10.9	+13.2	+5.9	+3.3	+13.6	+11.3
	GeoRSCLIP	74.4	59.9	66.7	83.7	77.4	73.8	75.0	33.7	52.2	88.5	68.5
	+ RS-TransCLIP	80.4	72.7	74.0	93.6	93.1	86.7	81.9	53.5	64.7	98.9	79.9
	Δ	+6.0	+12.8	+7.3	+9.9	+15.7	+12.9	+6.9	+19.9	+12.4	+10.4	+11.4
	RemoteCLIP	84.1	43.6	62.2	83.8	61.4	76.0	67.8	34.8	50.7	93.5	65.8
	+ RS-TransCLIP	92.7	52.3	71.5	89.3	82.1	84.1	88.1	44.9	60.5	97.2	76.3
	Δ	+8.6	+8.7	+9.3	+5.5	+20.7	+8.1	+20.3	+10.1	+9.8	+3.7	+10.5
	SkyCLIP50	72.1	51.5	64.0	80.9	75.3	70.5	66.8	38.0	46.6	87.5	65.3
	+ RS-TransCLIP	90.8	68.5	73.2	92.8	93.6	81.6	80.5	51.4	62.0	99.2	79.4
	Δ	+18.7	+17.0	+9.2	+11.9	+18.3	+11.1	+13.7	+13.4	+15.4	+11.7	+14.0
ViT-H/14	GeoRSCLIP	76.3	68.3	67.4	84.8	82.7	73.8	77.4	43.1	56.5	90.4	72.1
	+ RS-TransCLIP	83.8	91.2	78.1	94.5	96.2	88.0	83.3	54.8	72.8	99.7	84.2
	Δ	+7.5	+22.9	+10.7	+9.7	+13.5	+14.2	+5.9	+11.7	+16.3	+9.3	+12.1

Interestingly, RS-TransCLIP produces notable improvements even when the inductive model’s top-1 accuracy performance is already high. For example, when GeoRSCLIP ViT-H/14 is applied to WHURS19, the top-1 accuracy increases from 90.4% to 99.7%. Similarly, for the same model applied to PatternNet, the top-1 accuracy improves from 82.7% to 96.2%. This shows RS-TransCLIP’s applicability for tasks where these VLMs are already effective, without any labels.

We also notice that, for the ViT-L/14 backbone, RS-TransCLIP offers slightly higher gains to SkyCLIP50 compared to CLIP and RemoteCLIP, allowing it to outperform them when combined with transduction. RS-TransCLIP also demonstrates its applicability to more robust models, bringing an average gain of 12.1% on GeoRSCLIP ViT-H/14.

D. RS-TransCLIP computational cost

We evaluated the computational cost of RS-TransCLIP using three datasets of varying sizes. As shown in TABLE II, the feature extraction time increases with the number of image patches while the additional load from RS-TransCLIP remains minimal. Thus, by not requiring optimization of model parameters or input prompts [43], our transductive method ensures fast inference all while boosting model accuracy.

TABLE II: RS-TransCLIP run time on top of CLIP ViT-L/14, evaluated with 24GB NVIDIA GeForce RTX 4090 GPU.

RS dataset	Total patches	Features encoding time	+ RS-TransCLIP time
WHURS19	$\sim 10^3$	~ 8 seconds	~ 0.3 seconds
AID	$\sim 10^4$	~ 40 seconds	~ 2 seconds
MLRSNet	$\sim 10^5$	~ 6 minutes	~ 25 seconds

V. CONCLUSION

In this work, we proposed RS-TransCLIP, a transductive algorithm that enhances RS VLMs with minimal extra computational cost. By leveraging initial pseudo-labels and patch affinities, our method improves zero-shot capabilities through *transductive* inference, demonstrating significant accuracy improvements over inductive zero-shot classification and showing its wide applicability beyond natural images [21]. Future works will study RS-TransCLIP’s performance concerning text-prompt variability, given VLMs’ high sensitivity to input text prompts. Moreover, adapting RS-TransCLIP to a few-shot setting to incorporate labeled data will be explored in human-in-the-loop scenarios.

REFERENCES

- [1] H. Chen, C. Lan, *et al.*, “Land-cover change detection using paired openstreetmap data and optical high-resolution imagery via object-guided transformer,” *arXiv preprint arXiv:2310.02674*, 2023.
- [2] Q. Yuan, H. Shen, *et al.*, “Deep learning in environmental remote sensing: Achievements and challenges,” *Remote Sens. Environ.*, vol. 241, p. 111716, 2020.
- [3] W. H. Maes and K. Steppe, “Perspectives for remote sensing with unmanned aerial vehicles in precision agriculture,” *Trends Plant Sci.*, vol. 24, no. 2, pp. 152–164, 2019.
- [4] S. K. Phang, T. H. A. Chiang, *et al.*, “From satellite to uav-based remote sensing: A review on precision agriculture,” *IEEE Access*, 2023.
- [5] H. Xia, J. Wu, *et al.*, “A deep learning application for building damage assessment using ultra-high-resolution remote sensing imagery in turkey earthquake,” *Int. J. Disaster Risk Sci.*, vol. 14, no. 6, pp. 947–962, 2023.
- [6] K. El Khoury, T. Godelaine, *et al.*, “Streamlined hybrid annotation framework using scalable codestream for bandwidth-restricted uav object detection,” *arXiv preprint arXiv:2402.04673*, 2024.
- [7] M. B. Sariyildiz, J. Perez, *et al.*, “Learning visual representations with caption annotations,” in *ECCV*, pp. 153–170, Springer, 2020.
- [8] A. Joulin, L. Van Der Maaten, *et al.*, “Learning visual features from large weakly supervised data,” in *ECCV*, pp. 67–84, Springer, 2016.
- [9] T. Abdullah, Y. Bazi, *et al.*, “Texts: Deep bidirectional triplet network for matching text to remote sensing images,” *Remote Sens.*, vol. 12, no. 3, p. 405, 2020.
- [10] M. M. A. Rahhal, Y. Bazi, *et al.*, “Deep unsupervised embedding for remote sensing image retrieval using textual cues,” *Appl. Sci.*, vol. 10, no. 24, p. 8931, 2020.
- [11] F. Liu, D. Chen, *et al.*, “Remoteclip: A vision language foundation model for remote sensing,” *IEEE Trans. Geosci. Remote Sens.*, 2024.
- [12] A. Radford, J. W. Kim, *et al.*, “Learning transferable visual models from natural language supervision,” in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139 of *Proc. Mach. Learn. Res.*, pp. 8748–8763, PMLR, 2021.
- [13] Z. Zhang, T. Zhao, *et al.*, “Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing,” *arXiv preprint arXiv:2306.11300*, 2024.
- [14] C. Pang, J. Wu, *et al.*, “Towards helpful and honest remote sensing large vision language model,” *arXiv preprint arXiv:2403.20213*, 2024.
- [15] Z. Wang, R. Prabha, *et al.*, “Skyscript: A large and semantically diverse vision-language dataset for remote sensing,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, pp. 5805–5813, 2024.
- [16] D. Muhtar, Z. Li, *et al.*, “Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model,” *arXiv preprint arXiv:2402.02544*, 2024.
- [17] X. Li, C. Wen, *et al.*, “Vision-language models in remote sensing: Current progress and future trends,” *IEEE Geosci. Remote Sens. Mag.*, vol. 12, no. 2, pp. 32–66, 2024.
- [18] V. Vapnik, “An overview of statistical learning theory,” *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–999, 1999.
- [19] T. Joachims, “Transductive inference for text classification using support vector machines,” in *ICML*, vol. 99, pp. 200–209, 1999.
- [20] S. Martin, Y. Huang, *et al.*, “Transductive zero-shot and few-shot clip,” in *CVPR*, pp. 28816–28826, 2024.
- [21] M. Zanella, B. Gérin, *et al.*, “Boosting vision-language models with transduction,” *arXiv preprint arXiv:2406.01837*, 2024.
- [22] S. Zhang, Y. Xu, *et al.*, “Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs,” *arXiv preprint arXiv:2303.00915*, 2024.
- [23] S. Eslami, G. de Melo, *et al.*, “Does clip benefit visual question answering in the medical domain as much as it does in the general domain?,” 2021.
- [24] J. Luo, Z. Pang, *et al.*, “Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding,” *arXiv preprint arXiv:2406.10100*, 2024.
- [25] W. Zhang, M. Cai, *et al.*, “Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain,” *arXiv preprint arXiv:2401.16822*, 2024.
- [26] U. Mall, C. P. Phoo, *et al.*, “Remote sensing vision-language foundation models without annotations via ground remote alignment,” in *ICLR*, 2024.
- [27] Y. Hu, J. Yuan, *et al.*, “Rsgpt: A remote sensing vision language model and benchmark,” *arXiv preprint arXiv:2307.15266*, 2023.
- [28] Y. Bazi, L. Bashmal, *et al.*, “Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery,” *Remote Sens.*, vol. 16, no. 9, 2024.
- [29] G. S. Dhillon, P. Chaudhari, *et al.*, “A baseline for few-shot image classification,” in *ICLR*, 2019.
- [30] M. Boudiaf, I. Ziko, *et al.*, “Information maximization for few-shot learning,” *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 2445–2457, 2020.
- [31] J. Liu, L. Song, *et al.*, “Prototype rectification for few-shot learning,” in *ECCV*, pp. 741–756, Springer, 2020.
- [32] I. Ziko, J. Dolz, *et al.*, “Laplacian regularized few-shot learning,” in *ICML*, PMLR, 2020.
- [33] M. Zanella, F. Shakeri, *et al.*, “Boosting vision-language models for histopathology classification: Predict all at once,” in *International Workshop on Foundation Models for General Medical AI*, pp. 153–162, Springer, 2024.
- [34] G.-S. Xia, J. Hu, *et al.*, “Aid: A benchmark data set for performance evaluation of aerial scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, 2017.
- [35] P. Helber, B. Bischke, *et al.*, “Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification,” in *IGARSS*, pp. 204–207, 2018.
- [36] X. Qi, P. Zhu, *et al.*, “Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding,” *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 337–350, 2020.
- [37] Q. Wang, S. Liu, *et al.*, “Scene classification with recurrent attention of vhr remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, 2019.
- [38] W. Zhou, S. Newsam, *et al.*, “Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval,” *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 197–209, 2018.
- [39] G. Cheng, J. Han, *et al.*, “Remote sensing image scene classification: Benchmark and state of the art,” *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.
- [40] L. Zhao, P. Tang, *et al.*, “Feature significance-based multibag-of-visual-words model for remote sensing image scene classification,” *J. Appl. Remote Sens.*, vol. 10, 2016.
- [41] H. Li, X. Dou, *et al.*, “Rsi-cb: A large-scale remote sensing image classification benchmark using crowdsourced data,” *Sensors*, vol. 20, no. 6, 2020.
- [42] G.-S. Xia, W. Yang, *et al.*, “Structural high-resolution satellite image indexing,” *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 38, 2010.
- [43] M. Zanella and I. Ben Ayed, “On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning?,” in *CVPR*, pp. 23783–23793, June 2024.