Zero Shot Classification for Change Detection in Satellite Imaginary

Kürşat Kömürcü
Institute of Computer Science,
Vilnius University,
Vilnius, Lithuania
kursat.komurcu@mif.stud.vu.lt
0009-0006-1149-8686

Linas Petkevičius
Institute of Computer Science,
Vilnius University,
Vilnius, Lithuania
linas.petkevicius@mif.vu.lt
0000-0003-2416-0431

***Abstract*—This research investigates the zero-shot classification using the Comparative Language-Image Pre-Training (CLIP) model for change detection in satellite imagery. Since traditional supervised learning methods require extensive labeled datasets for all classes of objects in satellite images, which are often scarce or unavailable, zero-shot learning offers a promising alternative. Through detailed analysis of three different satellite image datasets (LEVIR-CD, DSIFN, and S2Looking), the study evaluates the ability of the zero-shot classification model to identify changes without prior exposure to specific target classes and compares it with traditional tree-based supervised learning algorithms. These findings highlight the potential of zero-shot learning as a powerful tool for monitoring, managing, and responding to global changes, underscoring its importance for remote sensing and Earth observation applications where fast, accurate change detection is critical. The research is accompanied by the data and reproducible code at Github repository.**

***Index Terms*—Zero Shot Classification, CLIP Model, Satellite Image Analysis, Change Detection**

I. Introduction

Zero-shot classification is a problem setup in machine learning, aiming to classify new classes that were not seen during the model training phase [25]. This approach is particularly useful in satellite imagery, where the ability to detect and classify unseen objects or changes are challenging due large variety of possible options [10], [23], [24].

Recent studies have demonstrated enhancing zero-shot object detection frameworks by incorporating class embedding vectors, synthesizing robust region features for unseen objects to achieve intra-class diversity and inter-class separability, and improving detection performance in remote sensing imagery [12]. Augmenting zero-shot detection training with image labels from datasets like ImageNet [5] has shown to significantly enhance the alignment of detector output to the embedding space, demonstrating a promising direction for improving zero-shot detection capabilities [14].

The application of zero-shot learning to remote sensing and satellite imagery has been successfully applied in studies introducing techniques for zero-shot scene classification in high spatial resolution images [4], [15], [16] and urban changes [8], [9]. These techniques leverage embedding vectors and directed graphs to recognize unseen scene classes [15]. Furthermore, vision-language models have been explored for zero-shot classification of remote sensing images, where pre-trained models grasp the associations between image-text pairs, demonstrating superior accuracy over existing solutions [13], [16], [17], [20].

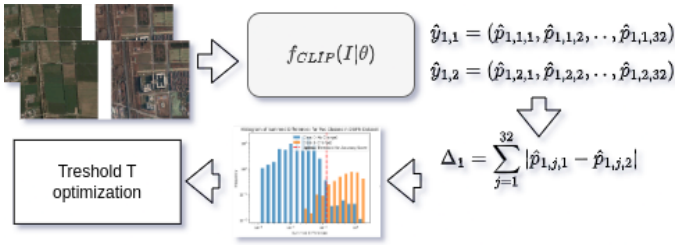

Fig. 1. The pipeline of zero-shot learning: using pre-trained CLIP model from learned embeddings to classity in selected number of classes and using it differences for threshold optimization.

Additional zero-shot learning methods focusing on transferable object proposal mechanisms and vision-language knowledge distillation present innovative approaches to overcome the challenges of domain shift in zero-shot detection [11], [21].

Recent advancements in zero-shot classification for change detection in satellite imagery highlight the integration of semantic information, the exploitation of pretrained models, and the synthesis of robust embedding features as key strategies. These studies provide the foundation for future studies in satellite image processing and remote sensing tasks. Thus we propose zero-shot learning approach for change detection see Fig 1. The method contains reusage of existing foundational models [19] and classification predictions to construct the change class distribution.

In the study, the methodology will be delineated in Section II, followed by an exploration of the datasets utilized in Section III. Subsequently, Section IV will present the results obtained from the study. Finally, the conclusion will be provided in Section V.

IEEE Member numbers: 94481679 and 96751685
This project was funded by the European Union (project No S-MIP-23-44) under the agreement with the Research Council of Lithuania (LMTLT).

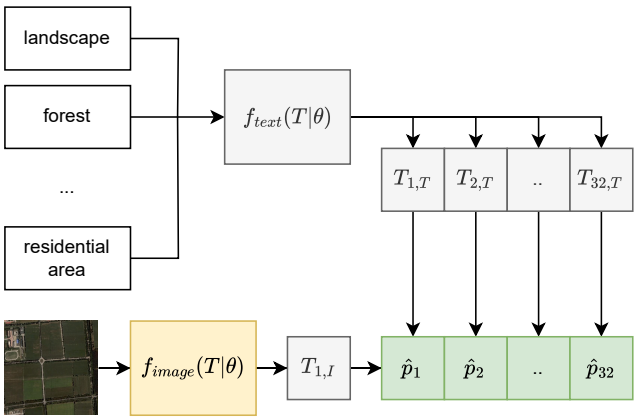

Fig. 2. CLIP model inference.

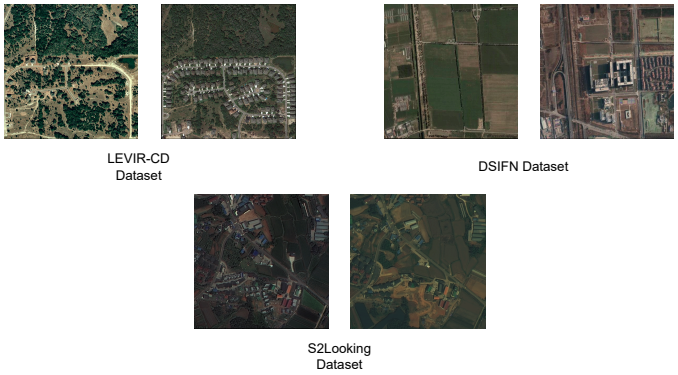

Fig. 3. Examples of Image Pairs for Each Datasets LEVIR-CD, DSIFN, S2Looking.

II. Methodology

In this section of our study, we present fundamental model we employed - Contrastive Language–Image Pre-training (CLIP) to analyze satellite imagery [19]. CLIP, a model developed by researchers of OpenAI, is designed to form expressive features of images in the context of natural language descriptions, making it particularly suited for tasks such as zero-shot classification where the objective is to classify classes that were not seen during the model's training phase [19]. Model consist from two encoder models $f_{image}(I|\theta_{image}) = T_{image} : \mathbb{R}^{d_w \times d_h \times d_c} \rightarrow \mathbb{R}^{d_e}, f_{text}(T|\theta_{text}) = T_{text} : \mathbb{R}^{K \times d_Z} \rightarrow \mathbb{R}^{d_e}$, where f_{text}, f_{image} neural networks encoding raw data to embeddings vectors T_{text} and T_{image}, respectively. The θ represents neural networks unknown parameters, $d_w \times d_h \times d_c$ - image dimensions, $K \times d_Z$ - text input. The model is pretrained on predicting $\hat{Y} = T_{image}^T T_{text}$ identification of corresponding pairs of images/text.

As input to the text encoder model, we provided an array containing the selected classes names common in satellite imaginary of size $K = 32$ objects which are *landscape, forest, building, road, vehicle, bridge, river, lake, farmland, airport, runway, ship, railway, parking lot, cloud, wind turbine, stadium, school, hospital, industrial site, park, beach, mountain, glacier, desert, volcano, crater, island, wetland, quarry, dam* and *residential area*. These objects could potentially be identified within the satellite images. This array served as the textual descriptions against which the model evaluates the imagery, predicting the likelihood $\hat{Y} = T_{text}$ of each object's presence within the image. By transforming outputted logits via softmax to probabilities for each of these objects, indicating their presence within each image see Fig 2.

This process was systematically applied across all images within our datasets. By doing so, we were able to assess the model's ability to identify changes in satellite imagery, many of which were not included in the training dataset of the model.

Upon completing the analysis with the CLIP model, the next step involved processing the model's output probabilities for each satellite image. The core of our methodology was to determine the presence of any significant changes within each image based on a assessment of probabilities assigned to potential classes identified by the CLIP model, by calculating difference (1), for each image i in dataset. Where $\hat{Y}_{i,k} = (\hat{p}_{i,k,1}, \hat{p}_{i,k,2}, ..., \hat{p}_{i,k,32})$, and $k = 1, 2$, the reference and query images, respectively.

To achieve this, we first summed the probabilities difference across all identified objects for each image to create a composite likelihood score. This score was intended to reflect the overall presence of change-indicative features within the image, as recognized by the CLIP model.

The critical part of our methodology was to classify images into two categories: 'change' and 'no change'. This classification was based on a threshold optimization process, which aim to identify the optimal probability threshold T that distinguishes between the two categories based on selected F1-score metric. The optimization was formulated as follows:

1) **Summation of differences:** For each image, sum the probabilities differences of all potential objects detected by the CLIP model.

$$\Delta_i = \sum_{j=1}^{n} |\hat{p}_{i,j,1} - \hat{p}_{i,j,2}| \tag{1}$$

where Δ_i is the sum of probabilities diffrerence for the i-th image and $n = 32$ is the total classes.

2) **Threshold Optimization:** Determine the optimal threshold T by evaluating a range of threshold values to maximize classification metrics. For each candidate threshold value, classify images as 'change' if their summed probability exceeds the threshold, or 'no change' otherwise.

$$C_i(T) = \begin{cases} 1 & \text{if } \Delta_i > T \\ 0 & \text{otherwise} \end{cases}$$

where $C_i(T)$ represents the classification of the i-th image under threshold T, with 1 indicating 'change' and 0 indicating 'no change'.

3) **Evaluation Metrics:** For each threshold T, compute key evaluation metrics such as F1 score, recall, precision, and accuracy. These metrics assess the performance

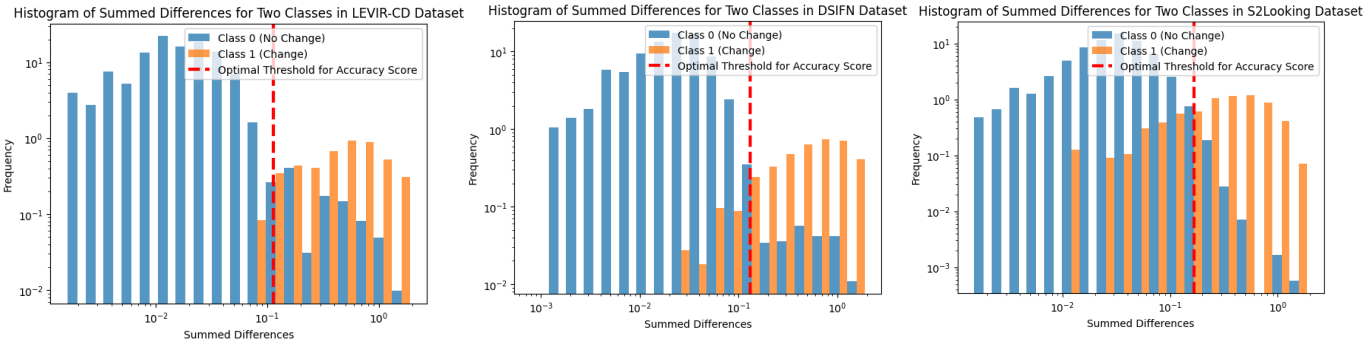

Fig. 4. Histogram of Classes for Each Datasets in Log Scale

of each threshold in accurately classifying images into 'change' and 'no change' categories.

$$F1(T), Recall(T), Precision(T), Accuracy(T)$$

4) **Optimal Threshold Selection:** Identify the threshold that maximizes the desired metrics.

$$T_{\text{opt},F1} = \arg\max_T F1(T)$$

$$T_{\text{opt},R} = \arg\max_T Recall(T)$$

$$T_{\text{opt},P} = \arg\max_T Precision(T)$$

$$T_{\text{opt},Acc} = \arg\max_T Accuracy(T)$$

This optimized thresholds T_{opt} is then used to classify all images in the dataset, providing a systematic and quantitatively justified method for detecting changes within satellite imagery. Through this proposed methodology, we ensure that our classification process is both robust and aligned with the overarching goal of identifying significant changes in the observed landscapes.

Following the zero-shot classification based on threshold optimization, we compared the results with those obtained from several tree-based machine learning algorithms which datasets divided by 70% train 30% test data and total amount of image pairs are 746 for LEVIR-CD dataset, 6601 for DSIFN dataset, 6501 for S2Looking dataset with augmented images. This comparative analysis aimed to assess the efficacy of our zero-shot learning approach against traditional supervised learning methods in the context of satellite image classification. The tree-based algorithms selected for this comparison were:

- **Decision Tree:** A fundamental tree-based model that partitions the dataset into subsets based on feature values, making decisions from the root to leaf nodes [18].
- **Random Forest:** An ensemble of decision trees that improves classification performance by reducing overfitting through averaging multiple decision trees trained on different parts of the dataset [1].
- **Gradient Boosting:** An additive model that sequentially adds weak decision trees to improve the model by focusing on instances that were misclassified in previous rounds [7].
- **XGBoost:** An optimized distributed gradient boosting library that provides a highly efficient implementation of the gradient boosting framework [3].

This comparative study was designed to highlight the potential advantages of employing a zero-shot learning framework for change detection in satellite imagery, particularly in scenarios where labeled data for certain classes might not be available or are scarce.

III. Datasets

In this study, we utilized three prominent datasets for evaluating our zero-shot classification methodology for change detection in satellite imagery: LEVIR-CD, DSIFN, and S2Looking. Illiustrative examples of images can be seen Fig 3. Each dataset offers a unique set of challenges and characteristics, making them ideal for assessing the robustness and effectiveness of our approach across different scenarios and environments.

A. LEVIR-CD

The LEVIR-CD dataset is specifically designed for building change detection and consists of 637 high-resolution aerial image pairs, each with a size of 1024×1024 pixels. The images, obtained from Google Earth, cover various urban and rural areas, providing a diverse set of building structures in different developmental stages. This dataset is particularly valuable for evaluating the model's ability to detect changes in man-made structures within complex landscapes [2].

B. DSIFN

The DSIFN (Deeply Supervised Image Fusion Network) dataset, although primarily curated for image fusion tasks, presents a unique opportunity for change detection studies. It comprises multi-temporal, multi-spectral, and multi-resolution images, offering a comprehensive dataset for examining the effectiveness of zero-shot learning models in identifying subtle changes that may not be immediately apparent in single-temporal or single-spectral data [6].

TABLE I
ALGORITHM COMPARISON FOR TRAINING DATA

	LEVIR-CD				DSIFN				S2Looking				Average			
	Acc	F1	Rec	Prec	Acc	F1	Rec	Prec	Acc	F1	Rec	Prec	Acc	F1	Rec	Prec
Threshold Optimization	0.9369	0.9439	1.0	0.9496	0.9648	0.9669	1.0	0.9664	0.9507	0.9535	1.0	0.9985	0.9508	0.9547	1.0	0.9715
Decision Tree	0.875	0.8703	0.8468	0.8952	0.9606	0.9609	0.9570	0.9647	0.9374	0.9419	0.9428	0.9410	0.9244	0.9243	0.9155	0.9336
Random Forest	0.9464	0.9473	0.9729	0.9230	0.9757	0.9764	0.9950	0.9586	0.9697	0.9718	0.9695	0.9741	0.9639	0.9651	0.9791	0.9519
Gradient Boosting	0.9508	0.9502	0.9459	0.9545	0.9732	0.9739	0.9870	0.9611	0.9723	0.9741	0.9676	0.9806	0.9654	0.9660	0.9668	0.9654
XGBoost	0.9553	0.9553	0.9639	0.9469	0.9782	0.9788	0.9930	0.9650	0.9743	0.9760	0.9714	0.9807	0.9692	0.97	0.9761	0.9642

TABLE II
ALGORITHM COMPARISON FOR TESTING DATA

	LEVIR-CD				DSIFN				S2Looking				Average			
	Acc	F1	Rec	Prec	Acc	F1	Rec	Prec	Acc	F1	Rec	Prec	Acc	F1	Rec	Prec
Threshold Optimization	0.9375	0.9669	1.0	0.9767	0.9352	0.9665	1.0	0.9642	0.94	0.9690	1.0	0.9960	0.9375	0.9674	1.0	0.9789
Decision Tree	0.8593	0.9203	0.8888	0.9541	0.9294	0.9623	0.9504	0.9746	0.947	0.9727	0.9488	0.9978	0.9119	0.9517	0.9293	0.9755
Random Forest	0.9296	0.9620	0.9743	0.95	0.9617	0.9802	0.9969	0.9640	0.971	0.9852	0.9729	0.9979	0.9541	0.9758	0.9813	0.9706
Gradient Boosting	0.8984	0.9432	0.9230	0.9642	0.9588	0.9785	0.9907	0.9667	0.974	0.9868	0.9759	0.9979	0.9437	0.9695	0.9632	0.9762
XGBoost	0.9296	0.9613	0.9572	0.9655	0.9764	0.9877	0.9969	0.9787	0.978	0.9888	0.9799	0.9979	0.9613	0.9792	0.978	0.9807

C. S2Looking

The S2Looking dataset is a large-scale dataset designed for remote sensing scene change detection, featuring over 5,000 pairs of high-resolution images. These images span a wide range of geographical areas and environmental conditions, from urban development to natural disasters. The dataset's variety in scenes and change types makes it an excellent resource for testing the generalizability of change detection algorithms across different domains and change scenarios [22].

To mitigate the issue of imbalanced labels within the LEVIR-CD, DSIFN, and S2Looking datasets, we employed a data augmentation strategy to equalize the distribution of images depicting changes and no changes. This augmentation process was implemented through a custom Python function that dynamically adjusts each image's rotation and scale. By randomly rotating the images within a range of -90 to 90 degrees and resizing them with a scale factor between 0.5 and 1.5, we generated diverse variations of the original datasets. This approach not only enhanced the balance between the classes but also enriched the datasets with a wider range of perspectives and scales, further challenging and thereby improving our model's robustness and ability to generalize across different types of change detection scenarios.

Datasets comprised pairs of images along with their corresponding masks for segmentation purposes were utilized. Subsequent to the application of the Contrastive Language–Image Pre-training (CLIP) model to each image, the resulting probability arrays were documented and stored in CSV format files. Following this automated process, manual labeling was conducted on the images within these CSV files, categorizing them as '1' to indicate change, and '0' to signify no change.

These datasets were selected for their relevance to the study's objectives and their potential to provide comprehensive insights into the performance of zero-shot classification algorithms for satellite imagery change detection. By leveraging these diverse datasets which can be seen in image pair examples on Fig 3, we aim to demonstrate the versatility and adaptability of our proposed methodology across various settings and challenges in satellite imagery analysis.

IV. RESULTS

Our study systematically evaluates the performance of the zero-shot classification methodology for change detection in satellite imagery. Utilizing the Contrastive Language–Image Pre-training (CLIP) model, we explored its efficacy across three distinct datasets: LEVIR-CD [2], DSIFN [6], and S2Looking [22]. Additionally, we provide a comparative analysis with traditional supervised learning algorithms to underscore the zero-shot approach's relative performance.

For a comprehensive perspective, we compared the zero-shot classification's results with those obtained from several tree-based supervised learning algorithms with different metrics as train data in Table I as test data in Table II.

The histograms in Figure 4 illustrate the frequency distribution of the summed differences for classes 0 and 1 across the LEVIR-CD, DSIFN, and S2Looking datasets, with the optimal threshold for accuracy delineated, showcasing the effectiveness of the zero-shot classification method in differentiating between the change classes.

In Figure 5, we present an illustrative example of change detection using the CLIP model on satellite imagery. The left side shows the original paired images, while the graph on the right quantifies the changes detected with increased probability of *industrial site*.

In our analysis of the performance metrics, it is evident that the zero-shot classification using the CLIP model yields competitive accuracy when benchmarked against traditional algorithms. Particularly, the XGBoost algorithm demonstrates high efficacy, as reflected by its F1 score and precision across all datasets for both training and testing data. However, the zero-shot approach stands out in its ability to maintain consistent performance without the need for extensive training data, a notable advantage in scenarios where labeled data is scarce or expensive to procure. The histograms further reveal the zero-shot method's robustness, indicating a significant

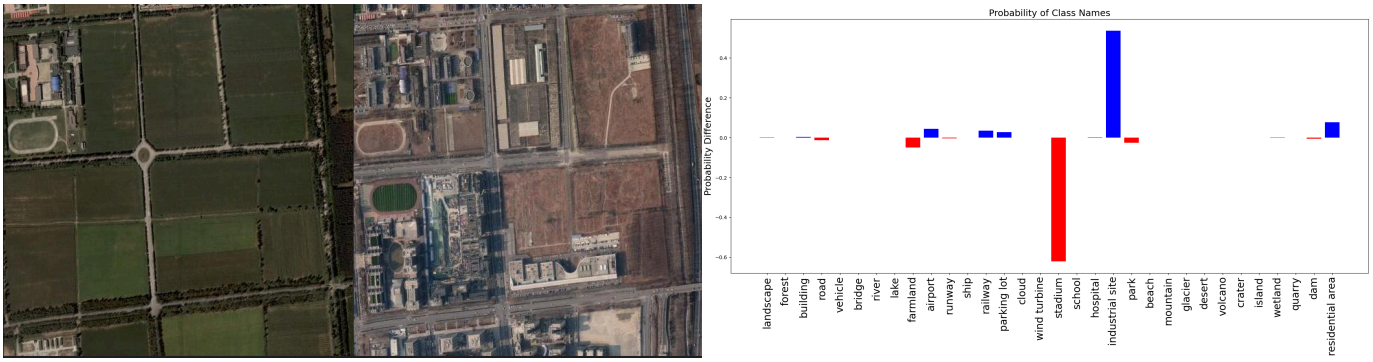

Fig. 5. A Change Detection Example for an Image Pair

frequency of accurate classifications at the optimal threshold. This threshold optimization appears to be a critical factor in enhancing the model's performance, suggesting that the zero-shot methodology could be finely tuned to achieve greater efficacy in change detection tasks. These findings suggest that while supervised methods continue to be reliable, the zero-shot learning provides a training-free method for remote sensing applications where adaptability and quick deployment are crucial.

V. Conclusion

This study we propose zero-shot classification, using the Contrastive Language–Image Pre-training (CLIP) embedding, for detecting changes in satellite imagery. Our findings, derived from evaluations across the LEVIR-CD, DSIFN, and S2Looking datasets, demonstrate the model's capability to identify changes without prior training on specific target classes, by achieving results equivalent of supervised learning machine learning methods. The research is accompanied by the data and reproducible code at Github repository.

References

[1] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[2] Pengyu Chen, Bo Du Chen, Wei Li, and Haifeng Lu. A large-scale dataset for remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[3] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[4] Gong Cheng, Xingxing Xie, Junwei Han, Lei Guo, and Gui-Song Xia. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:3735–3756, 2020.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Lei Fang, Shutao Li, and Jon Atli Benediktsson. Dsifn: Deeply supervised image fusion network. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020.

[7] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[8] Tautvydas Fyleris, Andrius Kriščiūnas, Valentas Gružauskas, and Dalia Čalnerytė. Deep learning application for urban change detection from aerial images. In *GISTAM 2021: proceedings of the 7th international conference on geographical information systems theory, applications and management, April 23-25, 2021*, volume 1, pages 15–24. SciTePress, 2021.

[9] Tautvydas Fyleris, Andrius Kriščiūnas, Valentas Gružauskas, Dalia Čalnerytė, and Rimantas Barauskas. Urban change detection from aerial images using convolutional neural networks and transfer learning. *ISPRS International Journal of Geo-Information*, 11(4):246, 2022.

[10] Dalia Grendaitė and Linas Petkevičius. Identification of algal blooms in lakes in the baltic states using sentinel-2 data and artificial neural networks. *IEEE Access*, 12:27973–27988, 2024.

[11] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Zero-shot detection via vision and language knowledge distillation. *ArXiv*, abs/2104.13921, 2021.

[12] Peiliang Huang, Junwei Han, De Cheng, and Dingwen Zhang. Robust region feature synthesizer for zero-shot object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7612–7621. IEEE, 2022.

[13] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint arXiv:2311.17179*, 2023.

[14] Katharina Kornmeier, Ulla Scheler, and P. Herrmann. Augmenting zero-shot detection training with image labels. *ArXiv*, abs/2306.06899, 2023.

[15] Aoxue Li, Zhiwu Lu, Liwei Wang, T. Xiang, and Ji-Rong Wen. Zero-shot scene classification for high spatial resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 55:4157–4167, 2017.

[16] Xiang Li, Congcong Wen, Yuan Hu, and Nan Zhou. Rs-clip: Zero shot remote sensing scene classification via contrastive vision-language supervision. *International Journal of Applied Earth Observation and Geoinformation*, 124:103497, 2023.

[17] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *arXiv preprint arXiv:2306.11029*, 2023.

[18] J. R. Quinlan. *Induction of Decision Trees*, volume 1. Kluwer Academic Publishers, 1986.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[20] Mohamad Mahmoud Al Rahhal, Y. Bazi, Hebah Elgibreen, and M. Zuair. Vision-language models for zero-shot classification of remote sensing images. *Applied Sciences*, 2023.

[21] Yilan Shao, Yanan Li, and Donghui Wang. Zero-shot detection with transferable object proposal mechanism. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3666–3670. IEEE, 2019.

[22] Xiang Shen, Tongwen Wu, Zhengxia Zou, Xiaoqing Zhang, Zhiqiang Zhou, and Wen Wang. S2looking: A satellite side-looking dataset for building change detection. *Remote Sensing*, 12(16):2643, 2020.

[23] Chufeng Tan, Xing Xu, and Fumin Shen. A survey of zero shot

detection: methods and applications. *Cognitive Robotics*, 1:159–167, 2021.

[24] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019.

[25] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.