

MiniCPM-V LLaMA Model for Image Recognition: A Case Study on Satellite Datasets

Kürşat Kömürcü  and Linas Petkevičius , *Member, IEEE*

Abstract—This study evaluates the performance of the MiniCPM-V model on four distinct satellite image datasets: MAI, RSICD, RSSCN7, and a newly created merged dataset that combines these three. The merged dataset was developed to expand the generalization and variation of data distribution associated with the labeling and training processes inherent in satellite image analysis. We systematically collected prediction results for each individual dataset and conducted a comparative analysis against results reported in previous studies to benchmark the model's effectiveness. The findings indicate that large language models (LLMs), such as MiniCPM-V, exhibit promising capabilities in the realm of satellite image recognition. On the RSSCN7 dataset, MiniCPM-V achieved an accuracy of 70.57% , while on RSICD it reached 62.19% , on MAI 7.01% , and on the merged dataset 43.49% . Specifically, the model demonstrated mostly high accuracy (more than 80%) in identifying a majority of object classes across the datasets. Also, we identified, it underperformed in accurately classifying certain object categories and recognizing all objects in multilabeled images, which suggests that while the model is robust overall, there are specific areas where its performance can be enhanced. Despite these limitations, the successful recognition of most objects underscores the potential of LLMs in advancing satellite imagery analysis. These results highlight the significant potential of integrating LLMs into remote sensing applications, offering a foundation for future research aimed at improving classification accuracy and expanding the range of detectable object classes by having caption level textual information.

Index Terms—Image recognition, LLama, large language models (LLMs), MiniCPM, remote sensing, satellite imagery, visual language models (VLMs).

I. INTRODUCTION

SATELLITE imagery plays a pivotal role in a wide range of applications, including environmental monitoring, disaster management, and urban planning [1]. For instance, advanced models have been developed for automatic weather classification from remote sensing images, enhancing environmental monitoring capabilities [2]. The vast amounts of data collected by satellites require advanced image recognition techniques to extract meaningful insights effectively and efficiently [3]. In recent years, traditional models that process satellite imagery have primarily relied on visual pixel information alone to interpret

data and generate insights [4]. However, with the rise of large language models (LLMs), new opportunities have emerged for enhancing these tasks by leveraging the model's extensive ability to process both visual and textual data [5].

Visual-language models (VLMs) have gained significant attention for their ability to process both textual and visual data simultaneously, enabling advancements in image captioning, visual question answering (VQA), and object recognition. Unlike traditional computer vision models, which rely solely on pixel-based feature extraction, VLMs integrate natural language understanding to improve classification and reasoning tasks. Recently, large multimodal language models (MLLMs) such as Llama [6], MiniGPT-4 [5], LLaVA (Large Language and Vision Assistant) [7], Otter [8], and PandaGPT [9] have demonstrated strong capabilities in processing and interpreting images through structured text-based queries.

Among these models, LLaMA-based MLLMs have gained particular interest due to their open-source nature and efficient scaling strategies. LLaVA, for instance, builds upon LLaMA by integrating a pretrained vision encoder, allowing it to answer questions about images with a high degree of contextual awareness [7]. Similarly, MiniGPT-4 aligns a vision transformer with a LLaMA-based text decoder to generate detailed image descriptions and reason about visual content [5]. While these models have achieved remarkable success in general-purpose vision-language tasks.

Traditional satellite image recognition techniques rely on deep learning architectures, such as CNNs and SVMs, which have been effective for high-resolution multispectral images [10]. However, these methods focus solely on pixel-based classification and lack the ability to incorporate textual descriptions or contextual metadata, limiting their semantic understanding. In addition, remote sensing images often exhibit high variability in resolution, scale, and spectral characteristics, making it difficult for conventional methods to generalize across datasets [11]. Ensemble learning techniques have been explored to improve classification accuracy [12], but they typically require large-scale labeled datasets and significant computational resources [13].

The integration of vision-language modeling into satellite image recognition provides several advantages over these traditional approaches. First, VLMs can process multimodal inputs, allowing them to leverage textual information alongside image features to improve classification performance. Second, because these models are pretrained on massive datasets, they can reduce reliance on manually labeled satellite imagery,

Received 27 November 2024; revised 1 February 2025; accepted 27 February 2025. Date of publication 3 March 2025; date of current version 25 March 2025. This work was supported by the European Union (project No S-MIP-23-45) under the agreement with the Research Council of Lithuania (LMTLT). (Corresponding author: Linas Petkevičius.)

The authors are with the Institute of Computer Science, Vilnius University, LT-08303 Vilnius, Lithuania (e-mail: linas.petkevicius@mif.vu.lt).

Digital Object Identifier 10.1109/JSTARS.2025.3547144

which is often limited or costly to obtain. Building on these strengths, MiniCPM-V extends the capabilities of LLaMA-based models by optimizing them for multimodal satellite image classification [14]. Unlike general-purpose multimodal models, MiniCPM-V is specifically designed to handle structured prompts for remote sensing data, making it well-suited for multilabel classification and object recognition tasks in satellite imagery.

This study aims to evaluate the performance of the MiniCPM-V model in satellite image recognition tasks, focusing on its ability to classify and recognize patterns in satellite datasets. Specifically, we chose MiniCPM-Llama3-V 2.5 version and we investigate how well the model can generalize across different types of satellite imagery, and compare its performance with traditional deep learning models, such as convolutional neural networks. By exploring MiniCPM-Llama3-V 2.5's capabilities in this context, we seek to contribute to the growing body of research on the application of LLMs to satellite image analysis, and to highlight the potential of these models in improving the accuracy and efficiency of satellite image recognition tasks.

The main contributions of this article are as follows.

- a) This study explores the use of the MiniCPM-V model, capable of processing both visual and textual data, as a novel approach for satellite image analysis, comparing its performance with existing methods.
- b) The MiniCPM-V model is evaluated across multiple satellite image datasets [multi-scene aerial image (MAI), RSICD, RSSCN7] and a combined dataset to analyze its generalization capabilities and the impact of data diversity on model performance.
- c) The study examines the model's performance in multilabel classification tasks, addressing challenges such as class imbalance and overlap, and discussing potential solutions for improving model accuracy.

The rest of this article is organized as follows. Section I, the model is presented. In Sections II and III, the data and its preparation is presented. Section IV, methodology is presented. In Section V, the results are presented followed by discussion. Finally, Section VI concludes this article.

II. MODEL

The MiniCPM-V model [14] is a smaller, optimized variant of the CPM (Chinese Pre-trained Language Model) family, specifically designed for multimodal tasks that require the integration of both visual and textual data. Built on transformer-based architecture, MiniCPM-V uses both Vision Transformers (ViTs) [15] and text transformers [16], allowing it to handle tasks like image classification, image captioning, and VQA with natural language prompts.

ViTs [15] are used for process visual data by splitting images into patches, embedding each patch into a vector space, and feeding them through transformer layers. This method allows the model to capture spatial relationships and features within images, similar to how transformers handle textual data. Unlike traditional CNNs, ViTs [15] in MiniCPM-V [14] can model long-range dependencies across the image, enhancing its ability to recognize complex patterns.

MiniCPM-V [14] is designed to be more computationally efficient than larger models in the CPM family, with fewer parameters but similar performance. This makes it suitable for environments, where computational resources are limited, such as real-time image classification or deployment on edge devices. The model's parameter-sharing techniques help maintain a balance between model size and accuracy, ensuring that it can be used effectively in practical applications. In this study, MiniCPM-V [14] was used for satellite image classification. The input consisted of satellite images combined with structured prompts to guide the model in recognizing specific patterns or objects in the images. The model was able to integrate visual and textual information, improving its accuracy in classifying diverse landscapes and objects in satellite imagery.

We selected MiniCPM-V for its efficiency in handling vision-language tasks while maintaining a balance between computational cost and performance. Unlike larger models, MiniCPM-V is optimized for multimodal learning, making it suitable for satellite image recognition without requiring extensive computational resources [14]. Among the MiniCPM model variants listed in Table I, MiniCPM-Llama3-V 2.5 stands out as the most effective option for our study. This model achieves the highest OpenCompass score (65.1) and exhibits superior performance across multiple benchmarks, including MME (2024.6), MMB test (en) (77.2), and LLaVA Bench (86.7). In addition, it demonstrates the lowest hallucination rate in the Object HalBench (10.3/5.0), indicating strong reliability in object recognition tasks. Compared to MiniCPM-V 1.0 and 2.0, the Llama3-V 2.5 variant significantly improves accuracy while maintaining a relatively compact size (8.5B parameters). Given its enhanced multimodal understanding, balanced computational efficiency, and robust classification accuracy, MiniCPM-Llama3-V 2.5 was selected as the optimal model for our satellite image recognition task. The performance metrics in Table I were obtained from [14], which provides a comprehensive evaluation of the MiniCPM family and its capabilities in various multimodal benchmarks.

III. DATASETS

In this study, three well-known datasets commonly used for remote sensing image recognition and classification were utilized: MAI, ¹ RSICD, ² and RSSCN7. ³ These datasets provide a diverse range of aerial scenes and are frequently used in research related to remote sensing and image processing.

The Multi-scene Aerial Image Dataset (MAI) is a dataset [39] designed for understanding aerial scenes, particularly focusing on recognizing various scene types. The dataset contains a wide variety of images aimed at multiscene recognition tasks. It was introduced in the context of prototype-based memory networks for scene classification, demonstrating the effectiveness of this method in identifying different scene types in aerial images. To facilitate the progress of aerial scene interpretation in the wild, we yield a new dataset, MAI dataset, by collecting and labeling 3923 large-scale images from Google Earth imagery that covers

¹[Online]. Available: <https://github.com/Hua-YS/Prototype-based-Memory-Network>

²[Online]. Available: https://github.com/201528014227051/RSICD_optimal

³[Online]. Available: <https://github.com/palewithout/RSSCN7>

TABLE I
EXPERIMENTAL RESULTS ON GENERAL MULTIMODAL BENCHMARKS

Model	Size	Open-Compass [17]	MME [18]	MMB test (en) [19]	MMB test (en) [19]	MMMU val [20]	Math-Vista [21]	LLaVA Bench [7]	RW QA [22],[23],[24]	Obj HalBench (Res./Men.) [25],[26]
<i>Proprietary Models</i>										
GPT-4V (2023.11.06)	—	63.5	1711.5	77.0	74.4	53.8	47.8	93.1	63.0	13.6/7.3*
Gemini Pro	—	62.9	2148.9	73.6	74.3	48.9	45.8	79.9	60.4	—
Claude 3 Opus	—	57.7	1586.8	63.3	59.2	54.9	45.8	73.9	48.4	—
<i>Open-source Models</i>										
DeepSeek-VL-1.3B [27]	1.7B	46.2	1531.6	66.4	62.9	33.8	29.4	51.1	49.7	16.7/9.6*
Mini-Gemini [28]	2.2B	—	1653.0	—	—	31.7	—	—	—	—
Yi-VL-6B [29]	6.7B	48.9	1915.1	68.4	66.6	40.3	28.8	51.9	53.5	19.4/11.7*
Qwen-VL-Chat [30]	9.6B	51.6	1860.0	61.8	56.3	37.0	33.8	67.7	49.3	43.8/20.0*
Yi-VL-34B [29]	34B	52.2	2050.2	72.4	70.7	45.1	30.7	62.3	54.8	20.7/14.0*
Phi-3-vision-128k-instruct [31]	4.2B	—	—	—	—	40.4	44.5	64.2*	58.8*	—
XTuner-Llama-3-8B-v1.1 [32]	8.4B	53.3	1818.0	71.7	63.2	39.2	40.0	69.2	—	—
CogVLM-Chat [33]	17B	54.2	1736.6	65.8	55.9	37.3	34.7	73.9	60.3	26.4/12.6*
Bunny-Llama-3-8B [34]	8.4B	54.3	1920.3	77.0	73.9	41.3	31.5	61.2	58.8	—
DeepSeek-VL-7B [27]	7.3B	54.6	1765.4	73.8	71.4	38.3	36.8	77.8	54.2	11.4/6.5*
LLaVA-NeXT-Llama3-8B [35]	8.4B	—	1971.5	—	—	41.7	—	80.1	60.0	—
Idetics2 [36]	8.0B	57.2	1847.6	75.7	68.6	45.2	52.2	49.1	60.7	—
Cambrian-8B [37]	8.3B	58.8	1802.9	74.6	67.9	41.8	47.0	71.0	60.0	—
CogVLM2-19B-Chat [33]	19B	62.3	1869.5	73.9	69.8	42.6	38.6	83.0	62.9	—
LLaVA-NeXT-Yi-34B [38]	34B	62.7	2006.5	81.1	79.0	48.8	40.4	81.8	66.0	—
Cambrian-34B [37]	34B	64.9	2049.9	80.4	79.2	50.4	50.3	82.0	67.1	—
MiniCPM-V 1.0 [14]	2.8B	47.5	1650.2	64.1	62.6	38.3	28.9	51.3	51.2	21.6/11.5
MiniCPM-V 2.0 [14]	2.8B	54.5	1808.6	69.1	66.5	38.2	38.7	69.2	55.8	14.5/7.8
MiniCPM-LLaMA3-V 2.5 [14]	8.5B	65.1	2024.6	77.2	74.2	45.8	54.3	86.7	63.5	10.3/5.0

RW QA: RealWorldQA, Obj HalBench (Res./Men.): Object HalBench With response/mention-Level Hallucination Rates, *: Our tested results with official checkpoints. The best open-source results are highlighted in bold [14].

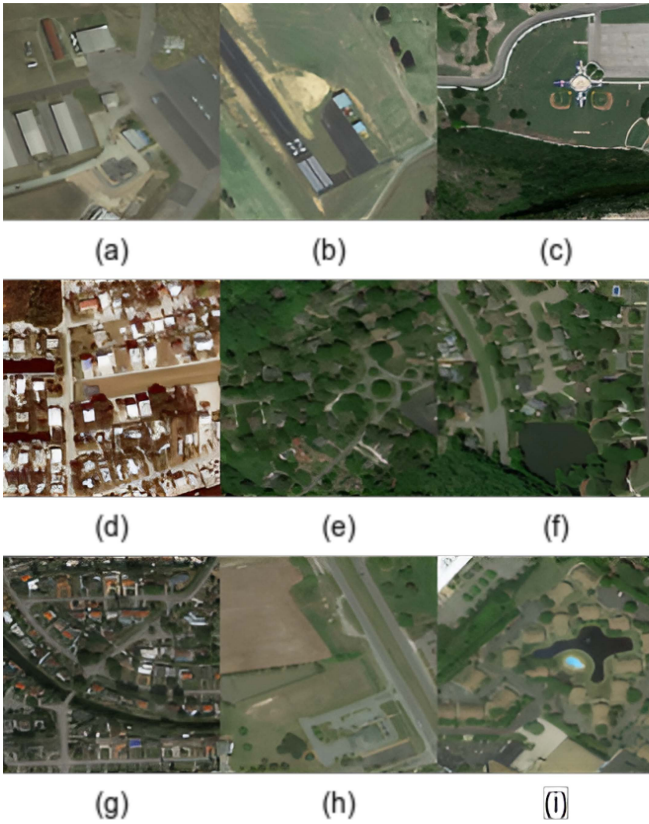


Fig. 1. Example images for MAI Dataset (a) “apron, parking lot, residential, runway,” (b) “apron, residential, runway,” (c) “baseball field, parking lot, river, park,” (d) “residential, runway,” (e) “residential, roundabout,” (f) “residential, lake, park,” (g) “residential, bridge, roundabout,” (h) “commercial, farmland, residential,” (i) “commercial, parking lot, residential, lake.”

the United States, Germany, and France. The size of each image is 512×512 , and spatial resolutions vary from 0.3 m/pixel to 0.6 m/pixel. The dataset has 24 classes, including apron, baseball, beach, commercial, farmland, woodland, parking lot, port, residential, river, storage tanks, sea, bridge, lake, park, roundabout, soccer field, stadium, train station, works, golf course, runway, sparse shrub, and tennis court see Fig. 1 and Table II.

The RSICD (Remote Sensing Image Captioning Dataset) is a dataset [40] specifically developed for generating textual descriptions of remote sensing images. RSICD contains a wide array of remote sensing images, which are used to train models that generate accurate captions for these images. This dataset is instrumental in improving the development of image captioning models in remote sensing contexts. The total number of sentences in RSICD is 24333, and the total words of these sentences are 3323. The dataset has 30 classes, including airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, school, square, parking, playground, pond, viaduct, port, railway station, resort, river, sparse residential, storage tanks, and stadium see Fig. 2 and Table III.

The RSSCN7 (Remote Sensing Scene Classification Dataset) is another dataset [41] used in deep learning-based scene classification tasks. It consists of a variety of scene types captured through remote sensing imagery and is widely employed to test the performance of deep learning-based feature selection methods for scene classification. The dataset provides a series of high-resolution images aimed at classifying different types of remote sensing scenes. The dataset RSSCN7 contains 2800 remote sensing sceneimages, which are from seven typical scene categories, namely, the grassland, forest, farmland, parking lot, residential region, industrial region, and river/lake. For each category, there are 400 images collected from the Google Earth, which are sampled on four different scales with 100 images per scale. Each image has a size of 400×400 pixels see Fig. 3.

To ensure consistency and fairness across the merged dataset, labels from the MAI, RSICD, and RSSCN7 datasets were systematically standardized by grouping synonymous or overlapping categories under unified labels. For example, “dense residential,” “medium residential,” “sparse residential,” and “resident” were consolidated into a single category named “Residential.” Similarly, “soccer field” and “playground” were unified under the label “Football field,” while “woodland” from MAI and “forest” from RSICD and RSSCN7 were combined as “Forest.” “Parking lot” and “parking” were merged into

TABLE II
NUMBER OF IMAGES FOR EACH CLASS IN THE MAI DATASET [39]

Class	Number	Class	Number	Class	Number
Residential	2387	Parking Lot	2007	Woodland	1610
Commercial	1610	Farmland	1222	Bridge	878
River	764	Lake	756	Park	638
Sparse Shrub	336	Soccer Field	302	Roundabout	281
Baseball Field	271	Runway	230	Storage Tanks	219
Apron	211	Works	186	Beach	165
Stadium	136	Tennis Court	114	Sea	80
Golf Course	75	Port	10	Train Station	9

TABLE III
NUMBER OF IMAGES FOR EACH CLASS IN THE RSICD DATASET [40]

Class	Number	Class	Number	Class	Number
Airport	420	Farmland	370	Playground	1031
Bare Land	310	Forest	250	Pond	420
Baseball Field	276	Industrial	390	Viaduct	420
Beach	400	Meadow	280	Port	389
Bridge	459	Medium Residential	290	Railway Station	260
Center	260	Mountain	340	Resort	290
Church	240	Park	350	River	410
Commercial	350	School	300	Sparse Residential	300
Dense Residential	410	Square	330	Storage Tanks	396
Desert	300	Parking	390	Stadium	290



Fig. 2. Example images for RSICD Dataset (a) airport, (b) bare land, (c) baseball field, (d) beach, (e) bridge, (f) center, (g) church, (h) commercial, (i) dense residential.

“Parking,” and “train station” was standardized as “Railway station” to align with other datasets. In addition, ambiguous or inconsistently represented labels such as “grass,” “meadow,” and “river/lake” were excluded due to their variability across

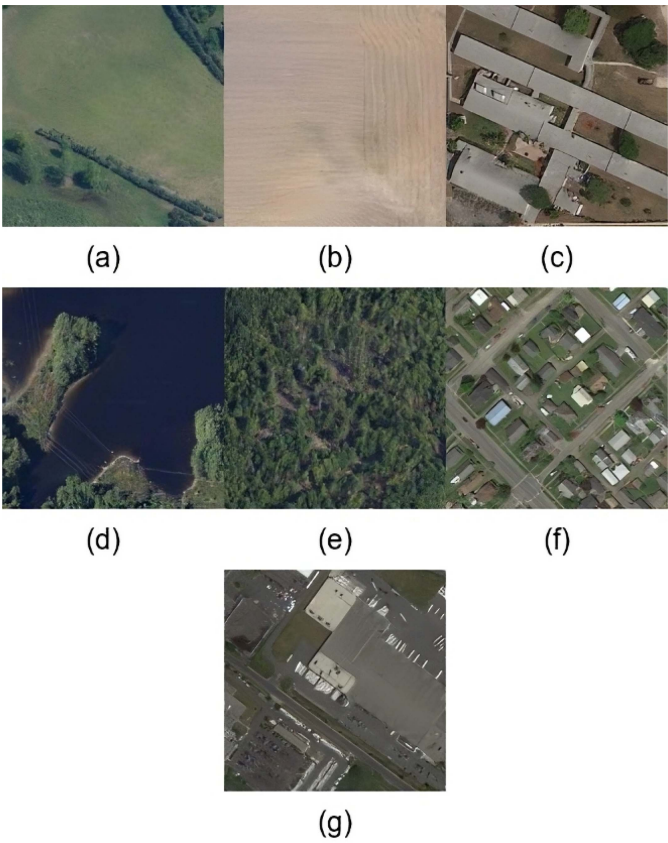


Fig. 3. Example images for RSSCN7 Dataset (a) grass, (b) field, (c) industry, (d) river/lake, (e) forest, (f) resident, (g) parking.

datasets, which could introduce noise and affect the reliability of the results. This careful standardization ensured the merged dataset retained its diversity while providing a clear and consistent framework for model evaluation.

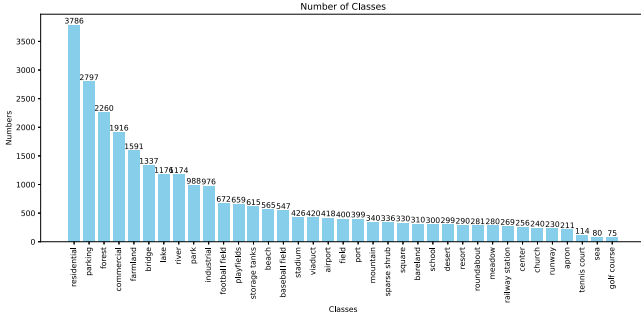


Fig. 4. Distribution of the classes in the merged dataset.

IV. METHODOLOGY

Each dataset meta-data was first converted into a CSV file format. For each dataset, we created a CSV file containing the names of the images along with their corresponding labels. This structure allowed for efficient handling and processing of the image data during the classification tasks. This preprocessing step ensured that the datasets were formatted in a uniform manner, enabling seamless integration into the classification pipeline.

For the classification model, we utilized the MiniCPM-V model [14]. We ran this model on the Google Colab platform using an L4 GPU. The input to this model included both the images and specific prompts designed to guide the classification process. These prompts provided context and additional instructions to enhance the model's ability to correctly classify each image. By leveraging a structured prompt-based approach, we ensured that the model received explicit guidance on object recognition, reducing ambiguity in classification. Data and reproduceable code can be found online.⁴

Using this approach, we leveraged the MiniCPM-V model to classify each image in the datasets and subsequently collected the classification results. After this process, we merged three datasets and converted synonymous labels into unified category names to maintain consistency. Specifically, the label *residential* was used in place of *dense residential*, *medium residential*, *sparse residential*, and *resident*; *football field* replaced *soccer field* and *playground*; *forest* was substituted for *woodland*; *parking* replaced *parking lot*; *industrial* replaced *works*; *railway station* replaced *train station*; and *lake* was used instead of *pond*. Moreover, we did not include *grass* and *river/lake* labels in the MAI dataset [39] in our merged dataset due to the unclear nature of their images (see Fig. 4).

The MiniCPM-V classification process follows a structured workflow, outlined in Algorithm 1. The methodology is designed to preprocess data, apply classification, and collect results systematically.

To ensure the MiniCPM-V model effectively classifies images, we used a set of structured prompts tailored to each dataset.

- 1) Identify the following categories in the satellite image and list them in a comma-separated format enclosed in double quotation marks: “apron, baseball field, beach, commercial, farmland, woodland, parking lot, port,

Algorithm 1: MiniCPM-V Image Classification Pipeline.

- 1: **Input:** Satellite image dataset D , MiniCPM-V model M , Prompts P
- 2: **Output:** Classified image labels L
- 3: Convert dataset D metadata into CSV format
- 4: **for** each image I in D **do**
- 5: Generate classification prompt P_I for I
- 6: Feed image I and prompt P_I into model M
- 7: Collect classification result L_I
- 8: **end for**
- 9: Merge dataset results and standardize labels
- 10: Return classified labels L

residential, river, storage tanks, sea, golf course, runway, sparse shrub, tennis court, bridge, lake, park, roundabout, soccer field, stadium, train station, works”. Return the identified categories in double quotation marks without any explanations or additional text.

- 2) Identify the object in this satellite image. Respond with only one word from the following list: airport, bareland, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, meadow, medium residential, mountain, park, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storage tanks, viaduct. Do not use any other words or phrases.
- 3) What do you see in this satellite image? Do not answer more than one word. Reply with only one word from these options: green areas, field, industry, river/lake, forest, resident, parking.
- 4) Identify the satellite image from the following list: field, industrial, river, lake, forest, residential, parking, airport, bareland, baseball field, beach, bridge, center, church, commercial, desert, farmland, meadow, mountain, park, playfields, port, railway station, resort, school, square, stadium, storage tanks, viaduct, apron, sea, golf course, runway, sparse shrub, tennis court, roundabout, football field. Note: Do not include any other words than the list and do not include any other additional information in your response. The image may contain one or multiple objects. Decide whether the image contains one or more objects and please list only the names of the detected object(s), if there are more than one objects and separate them by commas.

The MiniCPM-V model's performance was optimized by adjusting its hyperparameters to better align with the characteristics of the datasets. Two key hyperparameters, *temperature* and *sampling*, were carefully selected to control the model's output behavior, ensuring a balance between prediction diversity and determinism.

For the MAI, RSICD, and RSSCN7 datasets, the temperature was set to 0.7. This value allowed the model to explore diverse prediction possibilities while maintaining focus on the most probable labels, a crucial factor for multilabel classification tasks. On the other hand, for the merged dataset, a lower

⁴[Online]. Available: <https://www.kaggle.com/datasets/kursatkomurcu/minicpm-v-satellite-object-recognition/data>

TABLE IV
TOP-1/TOP-5 OVERALL METRICS OF THE DATASETS

Dataset	Accuracy	Precision	Recall	F1
MAI	0.0701/0.9783	0.4850/0.6180	0.6116/0.5000	0.5410/0.1100
RSICD	0.6219	0.6784	0.5836	0.5575
RSSCN7	0.7057	0.4257	0.4117	0.4084
Merged	0.4349	0.6026	0.4551	0.5186

temperature of 0.1 was applied. This stricter setting ensured deterministic and consistent predictions across the diverse and complex class distributions inherent in the merged dataset.

The sampling parameter was configured to influence the variability of the model's predictions. For the MAI, RSICD, and RSSCN7 datasets, sampling was enabled to encourage a broader exploration of potential labels, particularly useful in multilabel scenarios. Conversely, for the merged dataset, sampling was disabled to prioritize the most probable outputs, thereby reducing noise and enhancing reliability in predictions.

These adjustments were guided by empirical observations to ensure optimal alignment between the model's predictions and the specific demands of each dataset, without the need for fine-tuning the model's internal weights. By focusing on hyperparameter selection and dataset standardization, the MiniCPM-V model was effectively adapted for satellite image recognition tasks involving varying data complexities and label distributions. The inclusion of multilabel datasets allowed for a comprehensive evaluation of the model's capability to handle complex object recognition scenarios, further demonstrating the effectiveness of LLMs in satellite image classification.

V. RESULTS

Our study systematically evaluates the performance of the MiniCPM-V [14] model for satellite image recognition based purely on language information from VLM (see Table IV). We explored its efficacy across three distinct datasets: MAI [39], RSICD [40], RSSCN7 [41] and the merged dataset.

For the MAI dataset, the model achieved a notably low Top-1 accuracy of 0.0701, despite moderate precision 0.485 and recall 0.6116, culminating in an F1 score of 0.541. This indicates that while the model identified a significant portion of relevant instances (high recall), it struggled to correctly predict the majority class labels, as evidenced by the low accuracy. The Top-5 accuracy, however, was significantly higher at 0.9783, showing that the true labels were often present within the top five predictions. Despite this, the Top-5 precision and recall dropped to 0.618 and 0.5, respectively, resulting in an F1 score of 0.11. This suggests that while the model could identify relevant predictions, ranking these predictions accurately remains a challenge.

In contrast, the RSICD dataset showed improved performance with a Top-1 accuracy of 0.6219, precision of 0.6784, recall of 0.5836, and an F1 score of 0.5575. The Top-5 accuracy for RSICD remained at the same level 0.6219, but the precision and recall values dropped to 0.311 and 0.5, respectively, resulting in an F1 score of 0.3834. These results highlight that while the model performed reasonably well in ranking the top predictions, its ability to effectively utilize the additional predictions in a Top-5 setting was limited.

TABLE V
TOP-1 RESULTS FOR MAI DATASET

Class	Accuracy	Precision	Recall	F1
Apron	0.7204	0.1517	0.9147	0.2603
Baseball Field	0.9177	0.4502	0.8672	0.5927
Beach	0.9342	0.3542	0.6848	0.4669
Bridge	0.7798	0.5150	0.2745	0.3581
Commercial	0.6834	0.6769	0.3959	0.4996
Farmland	0.8218	0.6574	0.8936	0.7575
Golf Course	0.9207	0.1590	0.7333	0.2613
Lake	0.8022	0.4769	0.2725	0.3468
Park	0.8101	0.3803	0.2665	0.3134
Parking Lot	0.7693	0.7471	0.8301	0.7864
Port	0.9115	0.0172	0.6000	0.0334
Residential	0.7859	0.8480	0.7897	0.8178
River	0.7930	0.4579	0.3416	0.3913
Roundabout	0.6821	0.1683	0.8719	0.2821
Runway	0.9026	0.3128	0.5522	0.3994
Sea	0.9217	0.1397	0.5500	0.2228
Soccer Field	0.9014	0.3990	0.5563	0.4647
Sparse Shrub	0.8399	0.1933	0.2738	0.2266
Stadium	0.9118	0.1983	0.5074	0.2851
Storage Tanks	0.8792	0.2643	0.6530	0.3763
Tennis Court	0.9212	0.2269	0.7105	0.3439
Train Station	0.9286	0.0037	0.1111	0.0071
Woodland	0.7507	0.7008	0.6851	0.6928
Works	0.8871	0.0759	0.1237	0.0741

TABLE VI
COMPARISON FOR RSICD DATASET

Model	Recall	Reference
RemoteCLIP	0.3635	[42]
GeoRSLIP-FT	0.3887	[43]
AMFMN	0.1553	[44]
HarMA	0.3895	[45]
MiniCPM-Llama3-V 2.5	0.5836	[14] (Our Experiment)

The RSSCN7 dataset demonstrated the best performance across the board, with a Top-1 accuracy of 0.7057 and corresponding precision 0.4257, recall 0.4117, and F1 score 0.4084.

When evaluating the model on the merged dataset, which combines the challenges of all individual datasets, the Top-1 accuracy was moderate at 0.4349, with precision 0.6026, recall 0.4551, and an F1 score 0.5186. In the Top-5 setting, the accuracy improved to 0.7023, although precision 0.1547, recall 0.5, and the F1 score 0.2363 showed declines. These results suggest that while the model is capable of identifying relevant predictions in a broader set of data, achieving high precision across multiple labels remains challenging.

For the MAI and merged datasets, the lower Top-1 accuracy scores are due to their multilabeled nature—the MAI dataset is entirely multilabeled, while the merged dataset is partially so. In multilabel classification, accuracy reflects the model's ability to correctly predict each individual label, which inherently complicates the task. However, the significantly higher Top-5 accuracy values demonstrate that the model performs considerably better when evaluated with leniency in ranking predictions (see Tables V and VIII).

A. Results for MAI Dataset

Hua et al. [39] employed various machine learning and deep learning models on the entire MAI dataset, which comprises 100 000 images. They achieved maximum overall precision,

TABLE VII
RESULTS FOR RSICD DATASET

Class	Accuracy	Precision	Recall	F1
Airport	0.9961	0.9107	0.9952	0.9511
Bare Land	0.9760	0.5500	0.8516	0.6684
Baseball Field	0.9886	0.6959	0.9783	0.8133
Beach	0.9940	0.8678	0.9850	0.9227
Bridge	0.9885	0.8101	0.9477	0.8735
Center	0.9385	0.1284	0.2731	0.1747
Church	0.9875	0.8239	0.5458	0.6566
Commercial	0.9562	0.0077	0.0029	0.0042
Dense Residential	0.9242	0.3306	0.9951	0.4964
Desert	0.9860	0.8848	0.5633	0.6884
Farmland	0.9892	0.7800	0.9486	0.8561
Forest	0.9902	0.7871	0.7840	0.7856
Industrial	0.9648	1.0000	0.0129	0.0254
Meadow	0.9763	1.0000	0.0750	0.1395
Medium Residential	0.9311	0.1603	0.3759	0.2247
Mountain	0.9938	0.9928	0.8059	0.8896
Park	0.9740	0.6618	0.3857	0.4874
Parking	0.9810	0.9946	0.4718	0.6400
Playfields	0.9246	0.4112	0.5673	0.4768
Playground	0.9658	0.0000	0.0000	0.0000
Pond	0.9826	0.7146	0.9119	0.8013
Port	0.9959	0.9279	0.9589	0.9431
Railway Station	0.9950	0.9952	0.7923	0.8822
Resort	0.9821	0.7363	0.5103	0.6029
River	0.9920	0.9401	0.8415	0.8880
School	0.9757	0.9048	0.1271	0.2229
Sparse Residential	0.9728	0.5714	0.0268	0.0511
Square	0.9750	0.9385	0.1848	0.3089
Stadium	0.9609	0.3957	0.8966	0.5491
Storage Tanks	0.9934	0.8491	0.9949	0.9163
Viaduct	0.9926	0.9380	0.8643	0.8996

TABLE VIII
COMPARISON FOR RSSCN7 DATASET

Model	Accuracy	Reference
DBN	0.7700	[41]
GLNet	0.9507	[46]
CPM	0.5000	[47]
MiniCPM-Llama3-V 2.5	0.7057	[14] (Our Experiment)

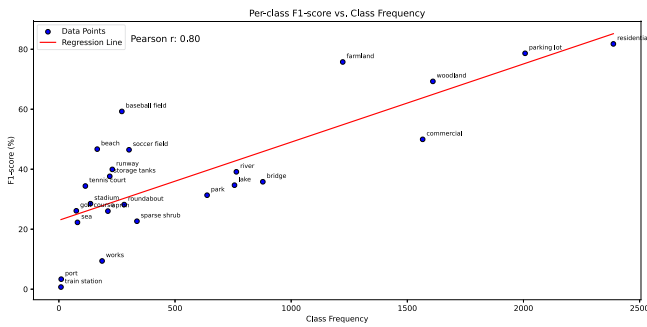


Fig. 5. MAI dataset F1 scores-class frequency graphics.

recall, and F1 scores of 0.801, 0.665, and 0.713, respectively. In addition, they reported only average precision results for each class. In contrast, our study utilized a subset of 3923 images from the MAI dataset. Although our results are lower than those reported in [39] (see Tables IV and V). Furthermore, we observed a positive correlation with 0.8 p -value between class frequencies and F1 scores (see Fig. 5).

As shown in Table V, the accuracy, precision, recall, and F1 scores for each class vary significantly. For example, the

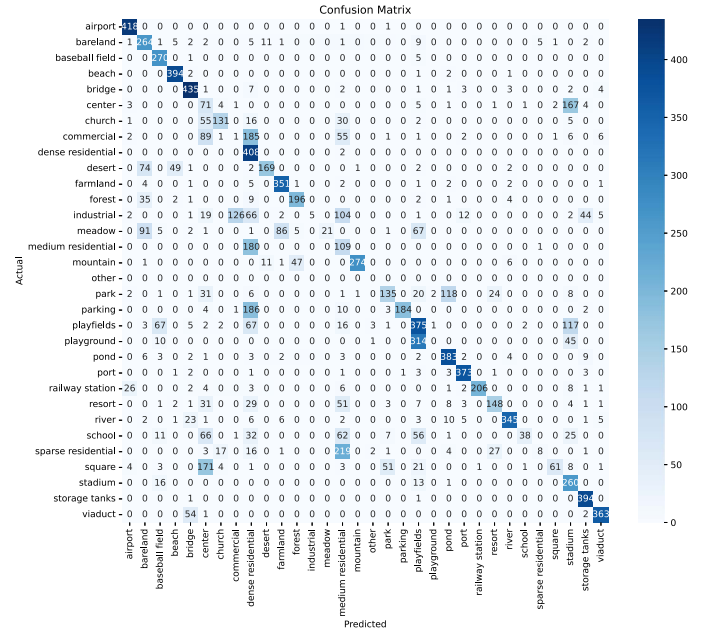


Fig. 6. Confusion matrix of RSICD dataset classification.

“Parking Lot” and “Residential” classes exhibit relatively high F1 scores, indicating that the model performs well in these categories. On the other hand, classes such as “Port,” “Train Station,” and “Works” have very low F1 scores. This could be due to the under representation of these classes in the dataset, potential confusion with other similar classes, or the model’s difficulty in distinguishing specific features of these classes. In addition, while the model demonstrates acceptable performance for certain classes like “Commercial,” “Woodland,” and “Farmland,” lower recall and precision rates were observed for some classes. These findings highlight the impact of data imbalance and overlap between classes on the overall model performance. Therefore, creating a more balanced dataset or employing techniques such as data augmentation could potentially improve the model’s performance. Furthermore, the prediction outcomes for sample images from the MAI dataset can be found in Fig. 9, which demonstrates the model’s performance visually, highlighting both false positives and false negatives.

B. Results for RSICD Dataset

Table VI presents the recall performance of various models on the RSICD dataset. Among the evaluated models, RemoteCLIP [42] achieved a recall of 0.3635, while GeoRSClip-FT [43] slightly improved this metric to 0.3887. The AMFMN model [44], however, demonstrated a significantly lower recall of 0.1553. In contrast, our proposed LLaMA MiniCPM-V model attained a substantially higher recall of 0.5836, outperforming all compared models by a considerable margin.

Moreover, our model achieved high metric values for each class (see Fig. 6 and Table X). This indicates that the MiniCPM-V model not only excels in overall recall, but also maintains strong performance across individual classes. The superior recall of the LLaMA MiniCPM-V model highlights its enhanced

TABLE IX
RESULTS FOR RSSCN7 DATASET

Class	Accuracy	Precision	Recall	F1
Green Areas	0.1300	0.1250	0.0162	0.0288
Field	0.9325	0.2000	0.1865	0.1930
Industry	0.6250	0.1667	0.1042	0.1282
River/Lake	0.8325	0.1250	0.1041	0.1136
Forest	0.8850	0.2500	0.2212	0.2347
Resident	0.6400	0.1250	0.0800	0.0976
Parking	0.8950	0.2000	0.1790	0.1889

TABLE X
RESULTS FOR THE MERGED DATASET

Class	Accuracy	Precision	Recall	F1
Airport	0.9899	0.7366	0.9234	0.8195
Apron	0.9873	0.4805	0.1754	0.2569
Bare Land	0.9832	0.7077	0.1484	0.2453
Baseball Field	0.9918	0.6790	0.8355	0.7492
Beach	0.9903	0.9052	0.7947	0.8464
Bridge	0.9463	0.8859	0.3717	0.5237
Center	0.9753	0.2130	0.2305	0.2214
Church	0.9918	0.7429	0.6500	0.6933
Commercial	0.8913	0.6239	0.1143	0.1932
Desert	0.9892	0.7489	0.5886	0.6592
Farmland	0.9279	0.8689	0.2791	0.4225
Field	0.8921	0.1694	0.9075	0.2855
Football Field	0.9631	0.5314	0.6429	0.5818
Forest	0.9109	0.8287	0.4239	0.5609
Golf Course	0.9957	0.5114	0.6000	0.5521
Industrial	0.9299	0.4258	0.5994	0.4979
Lake	0.9454	0.6904	0.3963	0.5035
Meadow	0.9785	0.0000	0.0000	0.0000
Mountain	0.9953	0.9888	0.7765	0.8699
Park	0.9411	0.4941	0.1700	0.2530
Parking	0.8815	0.7970	0.3847	0.5189
Playfields	0.9598	0.3396	0.0273	0.0506
Port	0.9960	0.9299	0.8972	0.9133
Railway Station	0.9942	0.8981	0.7212	0.8000
Residential	0.8750	0.6929	0.7979	0.7417
Resort	0.9859	0.7766	0.2517	0.3802
River	0.9491	0.7300	0.4284	0.5400
Roundabout	0.9437	0.1267	0.4021	0.1927
Runway	0.9858	0.4702	0.3087	0.3727
School	0.9816	0.4583	0.1833	0.2619
Sea	0.9951	0.3846	0.0625	0.1075
Sparse Shrub	0.9979	0.2500	0.0238	0.0435
Square	0.9811	0.7727	0.0515	0.0966
Stadium	0.9701	0.4275	0.5329	0.4744
Storage Tanks	0.9849	0.8399	0.7252	0.7784
Tennis Court	0.9827	0.1993	0.5175	0.2878
Viaduct	0.9745	0.2000	0.7100	0.0138

capability to identify relevant instances within the RSICD dataset, which may be attributed to its advanced feature extraction and classification mechanisms. These results suggest that MiniCPM-V is highly effective in capturing the diverse and complex patterns present in satellite imagery, leading to improved identification and classification of relevant classes. The consistent high performance across classes underscores the potential of the MiniCPM-V model for applications requiring robust and reliable satellite image recognition.

C. Results for RSSCN7 Dataset

Table VIII demonstrates that our experimental model achieves a lower accuracy compared to other approaches. Notably, the LLaMA MINICPM-V model [14] incorrectly classified many

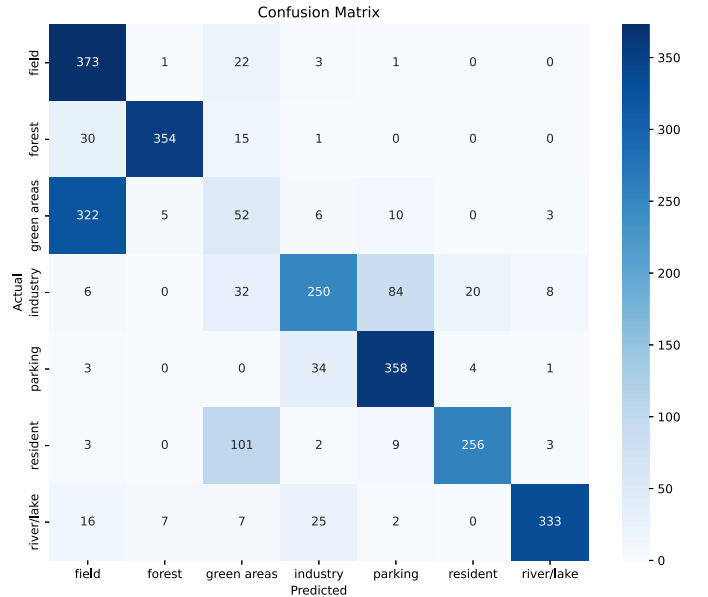


Fig. 7. Confusion matrix of RSSCN7 dataset classification.

instances of the *green areas* class, while exhibiting high accuracy for the remaining classes (see Table IX and Fig. 7).

Despite the overall lower accuracy of the MiniCPM-V model on the RSSCN7 dataset compared to other approaches, the performance across most classes remains relatively strong. The exceptionally low performance on the *green areas* class suggests that the model struggles with distinguishing this class from others, potentially due to high visual similarity with classes like *field* and *forest*. This confusion may arise from overlapping features and textures that make it challenging for the model to accurately differentiate between these categories. Overall, while the MiniCPM-V model shows promise in handling several classes within the RSSCN7 dataset, targeted improvements are necessary to enhance its performance on more ambiguous or visually similar categories.

D. Results for the Merged Dataset

The evaluation of the MiniCPM-V model on the merged dataset, which integrates the MAI, RSICD, and RSSCN7 datasets, provides a comprehensive assessment of the model's ability to generalize across diverse satellite image distributions and class heterogeneities (see Table X and Fig. 4). The merged dataset presents a more complex classification task due to the varied characteristics and label distributions inherited from the individual datasets.

Overall, the model achieved a moderate accuracy of 0.4349 on the merged dataset, which is lower compared to its performance on the individual datasets. This reduction in accuracy can be attributed to the increased diversity and complexity of the combined data, which challenges the model's generalization capabilities. Despite this, certain classes within the merged dataset exhibit high accuracy scores, indicating that the model can effectively recognize specific types of satellite imagery when sufficient distinguishing features are present.

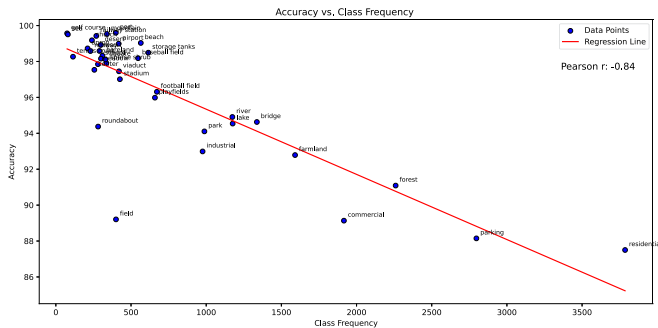


Fig. 8. Merged dataset accuracy scores-class frequency graphics.

For instance, classes such as *Airport* (Accuracy: 0.9899), *Bridge* (0.9463), *Mountain* (0.9953), and *Port* (0.996) demonstrate exceptionally high accuracy, precision, and recall values. These results suggest that the model performs well on classes with distinct and consistent visual patterns. Conversely, classes like *Apron* (Accuracy: 0.9873), *Bare Land* (0.9832), and *Commercial* (0.8913) show lower performance metrics, highlighting areas where the model struggles, possibly due to overlapping features with other classes or insufficient training samples.

The F1 scores across classes reveal a similar trend, with higher scores for well-defined classes and lower scores for more ambiguous or less represented classes. For example, the *Residential* class achieved an F1 score of 0.7417, while the *Sparse Shrub* class only reached 0.0435. These disparities indicate that while the model is capable of handling certain categories effectively, it faces challenges with classes that have subtle distinctions or limited representation in the merged dataset.

Fig. 8 illustrates the relationship between class frequency and accuracy scores, demonstrating a positive correlation where classes with higher frequency tend to achieve better accuracy.

The merged dataset results also reflect the inherent difficulties of multilabel classification, where the model must predict multiple classes simultaneously for each image. The complexity of this task increases with the diversity of the merged dataset, as the model must navigate a broader range of visual patterns and class overlaps. Despite these challenges, the MiniCPM-V model demonstrates robust performance on several key classes, underscoring its potential for applications requiring comprehensive satellite image recognition across varied environments.

In summary, the results for the merged dataset highlight both the strengths and limitations of the MiniCPM-V model. While it excels in accurately classifying certain distinct and well-represented classes, its performance diminishes in more complex and less frequent categories. These findings emphasize the need for further model refinement and potentially additional training data to enhance its generalization capabilities across diverse and heterogeneous satellite image datasets (see Fig. 10).

VI. DISCUSSION

Our study focuses on object recognition in satellite imagery using the MiniCPM-V VLM, rather than object counting or few-shot classification. While NWPU-MOC [48] aims to count

objects across multiple categories, our approach is designed to identify and classify objects within satellite images, leveraging a VLM that processes both textual and visual data. Similarly, while HSL-MINet [49] tackles few-shot classification using a multiview framework, our work does not rely on few-shot settings but rather evaluates the MiniCPM-V model across multiple large-scale remote sensing datasets (MAI, RSICD, RSSCN7, and a merged dataset).

Moreover, unlike previous works that primarily utilize CNN-based architectures or density map regression techniques, MiniCPM-V employs transformer-based vision models (Vision Transformers) along with a language-driven recognition process, allowing the model to integrate textual cues with visual patterns. Our findings highlight that LLMs can effectively assist in multilabel satellite image recognition, achieving competitive performance in object identification tasks.

While NWPU-MOC [48] and HSL-MINet [49] provide valuable insights into remote sensing object counting and few-shot learning, our research contributes to the field by demonstrating the potential of VLMs for remote sensing object recognition, offering a new perspective on leveraging large-scale multimodal learning in satellite image analysis. In addition, our approach differs from transformer-based multimodal learning frameworks such as the synchronized class token fusion (SCT Fusion) architecture [50], which fuses multiple input modalities to enhance classification accuracy. Unlike these methods, MiniCPM-V operates with a single modality (visual data) while leveraging a textual component to improve object recognition, making it adaptable to datasets lacking multimodal annotations. Similarly, our method contrasts with multitask fine-grained feature extraction [51], where each class is learned separately through a series of binary classification tasks. While this method enhances classification granularity, MiniCPM-V efficiently processes all class labels in a unified framework, reducing computational complexity while maintaining strong classification performance. Compared to HiReNet [52], which models hierarchical relationships to improve few-shot learning, our approach does not specifically target low-data scenarios. Instead, it is evaluated across large-scale datasets, demonstrating its robustness in recognizing diverse satellite imagery patterns. These comparisons underscore how MiniCPM-V presents a viable alternative to existing methods, offering a lightweight yet effective solution for multilabel classification without requiring hierarchical modeling, multitask decomposition, or multimodal fusion.

The evaluation of the MiniCPM-V model across the MAI, RSICD, RSSCN7, and merged datasets reveals a nuanced performance landscape. The model demonstrated strong accuracy and balanced precision and recall on the RSSCN7 and RSICD datasets, indicating its effectiveness in environments with distinct and well-represented classes. Particularly, the high recall on the RSICD dataset (0.5836) underscores MiniCPM-V's capability to identify relevant instances effectively. However, the model struggled with the MAI dataset, achieving a low accuracy of 0.0701, which can be attributed to the dataset's fully multilabeled nature, increasing the complexity of the classification task. Similarly, the merged dataset, which combines the challenges of all individual datasets, resulted in a moderate accuracy of

0.4349. This decline highlights the difficulties the model faces in generalizing across diverse data distributions and handling class heterogeneity.

Recent studies in remote sensing classification have explored network fusion approaches, combining CNN-based architectures with transformer models to enhance classification performance. For instance, a super-resolution framework integrating CNN architectures has shown significant improvements in remote sensing image classification by refining feature resolution and enhancing textural details [53]. Similarly, a network-level fusion of self-attention and ViTs has been introduced for land use and land cover classification, demonstrating that hybrid models leveraging both local and global feature extraction can yield superior results [54]. Unlike these approaches, which focus on fusing different visual network architectures, MiniCPM-V integrates language-driven object recognition into the classification pipeline. Rather than merging feature extraction networks, it employs a vision-language approach, allowing the model to process textual information alongside image features, which is particularly advantageous for datasets, where semantic relationships between objects provide additional classification cues. While network fusion methods enhance pixel-based feature learning, our approach introduces a different paradigm by leveraging multimodal alignment to improve recognition capabilities.

The discrepancies in performance across different datasets suggest that while MiniCPM-V excels in scenarios with clear and consistent visual patterns, it requires further refinement to manage multilabel classifications and underrepresented classes effectively. The positive correlation between class frequency and accuracy in the merged dataset indicates that the model benefits from ample training data, which enhances its ability to generalize for more prevalent classes. Conversely, classes with limited representation or subtle distinctions pose significant challenges, leading to lower performance metrics. This aligns with findings from multitask fine-grained feature mining approaches, which demonstrate that class-specific learning can improve underrepresented category recognition [51]. However, unlike such approaches, MiniCPM-V does not require a dedicated training mechanism for rare categories, instead relying on large-scale learning to infer semantic relationships between object classes.

Comparative analysis with existing models shows that MiniCPM-V achieves competitive recall on the RSICD dataset but falls short in accuracy on the RSSCN7 dataset compared to models like GLNet. This suggests that while MiniCPM-V has strengths in certain contexts, there is room for improvement in handling specific classes and ensuring consistent performance across all categories. Future work could explore hybrid approaches that integrate hierarchical structures for underrepresented classes or modality-aware training pipelines, taking inspiration from state-of-the-art remote sensing classification techniques. The results presented in this study highlight the viability of VLMs in satellite image recognition and open avenues for further optimizations that balance efficiency and accuracy across multilabel classification tasks.

VII. CONCLUSION

This study presents a comprehensive evaluation of the MiniCPM-V model's performance in satellite image recognition across multiple datasets. The results demonstrate that MiniCPM-V is highly effective in classifying distinct and well-represented classes within the RSSCN7 and RSICD datasets, achieving high accuracy and balanced precision and recall. However, the model's performance diminishes in more complex and heterogeneous datasets, such as MAI and the merged dataset, where multilabel classification and class diversity introduce significant challenges.

The findings highlight the necessity for further enhancements to MiniCPM-V to improve its generalization capabilities across diverse and multilabeled environments. Future work should focus on incorporating advanced techniques such as data augmentation, transfer learning, and strategies to address class imbalance. In addition, refining the model's architecture to better handle multilabel classifications and leveraging more comprehensive training data could enhance its robustness and accuracy across varied satellite image datasets.

In conclusion, while MiniCPM-V shows promising results in specific scenarios, achieving consistent and high performance across all tested datasets will require targeted improvements. Addressing the identified limitations will pave the way for MiniCPM-V to be a more versatile and reliable tool in the field of remote sensing and satellite image analysis.

APPENDIX

Image Samples from MAI Dataset				
Ground Truths	residential, bridge, park, stadium	parking lot, residential, bridge, park	commercial, parking lot, residential	river, storage tanks
Predictions	apron, commercial, parking lot, port	residential, roundabout, works	parking lot, commercial, residential	apron, storage tanks
Image Samples from MAI Dataset				
Ground Truths	commercial, parking lot, residential, bridge	residential, park, roundabout	farmland, woodland	commercial, parking lot, residential
Predictions	apron, bridge, commercial, farmland, parking lot, residential, roundabout	apron, baseball field, parking lot	farmland, woodland	residential, commercial, parking lot, roundabout

Fig. 9. Predictions for MAI dataset. Red: Refers to wrong predictions. Blue: Refers to correct predictions which are in the image but not in the ground truth labels.

Image Samples from the Merged Dataset				
Ground Truths	apron, parking, lake	residential, river, bridge, lake	baseball field, parking, residential, lake, park	airport
Predictions	runway, apron, parking	residential, bridge, river	residential, roundabout, baseball field	airport
Image Samples from MAI Dataset				
Ground Truths	beach	church	desert	mountain
Predictions	bridge, beach, sea	church, parking, square	desert	mountain

Fig. 10. Predictions for the merged dataset. Red: Refers to wrong predictions. Blue: Refers to correct predictions which are in the image but not in the ground truth labels.

ACKNOWLEDGMENT

The authors would like to thank for the high performance computing resources provided by the Information Technology Research Center of Vilnius University. The authors also would like to thank Google LLC for supplying computational infrastructure for this study via Google Colab platform.

REFERENCES

- [1] S. N. K. B. Amit and Y. Aoki, "Disaster detection from aerial imagery with convolutional neural network," in *Proc. Int. Electron. Symp. Knowl. Creation Intell. Comput. (Institute IES-KCIC)*, 2017, pp. 239–245.
- [2] M. T. Nabi, S. Ali, Z. Mahmood, M. A. Khan, and S. Alsenan, "A self-supervised deep-driven model for automatic weather classification from remote sensing images," *Int. J. Remote Sens.*, vol. 47, pp. 1–26, 2024, doi: [10.1080/01431161.2024.2431184](https://doi.org/10.1080/01431161.2024.2431184).
- [3] F. Ofli et al., "Combining human computing and machine learning to make sense of big (aerial) data for disaster response," *Big Data*, vol. 4, no. 1, pp. 47–59, 2016.
- [4] S. K. Bypina and K. S. Rajan, "Semi-automatic extraction of large and moderate buildings from very high-resolution satellite imagery using active contour model," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 1885–1888.
- [5] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," in *Proc. 12th Int. Conf. Learn. Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=1tZbq88f27>
- [6] H. Touvron et al., "Llama: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.
- [7] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. K. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., vol. 36, Red Hook, NY: Curran Associates, Inc., 2023, pp. 34892–34916. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf
- [8] J. Li et al., "Otter: A multi-modal model with in-context instruction tuning," 2023, *arXiv:2305.03726*.
- [9] W. Su et al., "PandaGPT: One model to instruction-follow them all," 2023, *arXiv:2305.16355*.
- [10] M. Pritt and G. Chern, "Satellite image classification with deep learning," in *Proc. IEEE Appl. Imagery Pattern Recognit. Workshop*, 2017, pp. 1–7.
- [11] A. Wang, P. Tian, and S. Wang, "High resolution satellite imagery segmentation based on adaptively integrated multiple features," *MIPPR 2007: Autom. Target Recognit. Image Anal.; Multispectral Image Acquisition*, vol. 6786, pp. 812–818, 2007.
- [12] M. Wieland and M. Pittore, "Performance evaluation of machine learning algorithms for urban pattern recognition from multi-spectral satellite images," *Remote Sens.*, vol. 6, no. 4, pp. 2912–2939, 2014.
- [13] J. Xue, Y. Leung, and T. Fung, "An unmixing-based Bayesian model for spatio-temporal satellite image fusion in heterogeneous landscapes," *Remote Sens.*, vol. 11, no. 3, 2019, Art. no. 324. [Online]. Available: <https://www.mdpi.com/2072-4292/11/3/324>
- [14] Y. Yao et al., "MiniCPM-V: A GPT-4v level MLLM on your phone," 2024, *arXiv:2408.01800*.
- [15] D. Alexey, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [16] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
- [17] O. Contributors, "Opencompass: A universal evaluation platform for foundation models," <https://github.com/open-compass/opencompass>, 2023.
- [18] C. Fu et al., "MME: A comprehensive evaluation benchmark for multi-modal large language models," 2023, *arXiv:2306.13394*.
- [19] Y. Liu et al., "MMbench: Is your multi-modal model an all-around player?," in *Proc. Eur. Conf. Comput. Vision*, 2024, pp. 216–233.
- [20] X. Yue et al., "MMM-U: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 9556–9567.
- [21] P. Lu et al., "MathVista: Evaluating mathematical reasoning of foundation models in visual contexts," in *Proc. Int. Conf. Learn. Representations*, 2024, pp. 1–17.
- [22] Y. Liu et al., "On the hidden mystery of OCR in large multimodal models," 2023, *arXiv:2305.07895*.
- [23] M. Mathew, D. Karatzas, and C. V. Jawahar, "DoCVQA: A dataset for VQA on document images," in *Proc. Winter Conf. Appl. Comput. Vis.*, 2021, pp. 2200–2209.
- [24] A. Singh et al., "Towards VQA models that can read," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8317–8326.
- [25] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, "Object hallucination in image captioning," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 4035–4045. [Online]. Available: <https://aclanthology.org/D18-1437/>
- [26] T. Yu et al., "RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 13807–13816.
- [27] H. Lu et al., "Deepseek-VL: Towards real-world vision-language understanding," 2024, *arXiv:2403.05525*.
- [28] Y. Li et al., "Mini-gemini: Mining the potential of multi-modality vision language models," 2024, *arXiv:2403.18814*. [Online]. Available: <https://arxiv.org/abs/2403.18814>
- [29] A. Young, "Yi: Open foundation models by 01.ai," 2024.
- [30] J. Bai et al., "Qwen-VL: A frontier large vision-language model with versatile abilities," 2023, *arXiv:2308.12966*.
- [31] M. I. Abdin et al., "Phi-3 technical report: A highly capable language model locally on your phone," Microsoft, Tech. Rep. MSR-TR-2024-12, Aug. 2024. [Online]. Available: <https://www.microsoft.com/en-us/research/publication/phi-3-technical-report-a-highly-capable-language-model-locally-on-your-phone/>
- [32] XTuner Contributors, "XTuner: A toolkit for efficiently fine-tuning LLM," 2023 Accessed: Jan. 30, 2025. [Online]. Available: <https://github.com/InternLM/xtuner>
- [33] W. Wang et al., "CogVLM: Visual expert for pretrained language models," in *Proc. 38th Annu. Conf. Neural Inf. Process. Syst.*, 2024. [Online]. Available: <https://openreview.net/forum?id=6dYBP3BIwX>
- [34] M. He et al., "Efficient multimodal learning from data-centric perspective," 2024, *arXiv:2402.11530*.
- [35] B. Li et al., "Llava-next: Stronger llms supercharge multimodal capabilities in the wild," 2024. [Online]. Available: <https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/>

- [36] H. Laurençon, L. Tronchon, M. Cord, and V. Sanh, “What matters when building vision-language models?,” in *Proc. 38th Annu. Conf. Neural Inf. Process. Syst.*, 2024. [Online]. Available: <https://openreview.net/forum?id=dtvJF1Vy2i>
- [37] P. Tong et al., “Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs,” *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 87310–87356, 2024.
- [38] H. Liu et al., “LLAVA-next: Improved reasoning, OCR, and world knowledge,” 2024. [Online]. Available: <https://llava-vl.github.io/blog/2024-01-30-llava-next/>
- [39] Y. Hua, L. Mou, J. Lin, K. Heidler, and X. X. Zhu, “Aerial scene understanding in the wild: Multi-scene recognition via prototype-based memory networks,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 177, pp. 89–102, 2021.
- [40] X. Lu, B. Wang, X. Zheng, and X. Li, “Exploring models and data for remote sensing image caption generation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2017.
- [41] Q. Zou, L. Ni, T. Zhang, and Q. Wang, “Deep learning based feature selection for remote sensing scene classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.
- [42] F. Liu et al., “Remoteclip: A vision language foundation model for remote sensing,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5622216.
- [43] Z. Zhang, T. Zhao, Y. Guo, and J. Yin, “RS5M and GEORSCLIP: A large scale vision language dataset and a large vision-language model for remote sensing,” *IEEE Trans. Geosci. Remote Sens.*, 2024.
- [44] Z. Yuan et al., “Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4404119.
- [45] T. Huang, “Efficient remote sensing with harmonized transfer learning and modality alignment,” 2024, *arXiv:2404.18253*.
- [46] H. Sun, Y. Lin, Q. Zou, S. Song, J. Fang, and H. Yu, “Convolutional neural networks based remote sensing scene classification under clear and cloudy environments,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 713–720.
- [47] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4724–4732.
- [48] J. Gao, L. Zhao, and X. Li, “NWPU-MOC: A benchmark for fine-grained multi-category object counting in aerial images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5606614.
- [49] Y. Jia, J. Gao, W. Huang, Y. Yuan, and Q. Wang, “Exploring hard samples in multiview for few-shot remote sensing scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5615714.
- [50] D. S. Hoffmann, K. N. Clasen, and B. Demir, “Transformer-based multi-modal learning for multi-label remote sensing image classification,” in *Proc. IGARSS IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 4891–4894.
- [51] J. Guo, H. Sun, J. Han, B. Song, Y. Chi, and B. Song, “Multitask fine-grained feature mining for multilabel remote sensing image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5632717.
- [52] F. Tian et al., “HireNet: Hierarchical-relation network for few-shot remote sensing image scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5603710.
- [53] H. M. Albarakati et al., “A unified super-resolution framework of remote-sensing satellite images classification based on information fusion of novel deep convolutional neural network architectures,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 14421–14436, 2024.
- [54] S. Rubab et al., “A novel network-level fusion architecture of proposed self-attention and vision transformer models for land use and land cover classification from remote sensing images,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 13135–13148, 2024.



Kürşat Kömürcü received the B.Sc. degree in electronics and communication engineering with Yildiz Technical University, Esenler, Türkiye, in 2023. Since September 2023, he has been working toward the M.Sc. degree in informatics with the Institute of Computer Science, Vilnius University, Vilnius, Lithuania. His research interests are focused on computer vision and deep learning.



Linas Petkevičius (Member, IEEE) received the Ph.D. degree in informatics with the Institute of Computer Science, Vilnius University, Vilnius, Lithuania, in 2020. Since 2022, he has been the Head of the Software Engineering department with the Institute of Computer Science. His research interests are focused on computer vision and deep learning, as well as statistical inference and outlier detection.