Dual-Spectrum All-Sky camera Cloud Classifier by means of Computer Vision Models

All-Sky camera Cloud Type Classifier by means of Computer Vision Models

Paolo Pertino
Alta Scuola Politecnica
Politecnico di Milano
Milano, Italy

Leonardo Pavarino
Alta Scuola Politecnica
Politecnico di Torino
Torino, Italy

Simone Lomolino
Politecnico di Milano
Milano, Italy

Enrico Miotto
Alta Scuola Politecnica
Politecnico di Torino
Torino, Italy

Daniele Rege Cambrin, Paolo Garza
Dipartimento di Automatica e Informatica
Politecnico di Torino
Torino, Italy

Elena Collino
Ricerca sul Sistema Energetico
RSE SpA
Milano, Italy

Maciej Sakwa, Emanuele Ogliari
Dipartimento di Energia
Politecnico di Milano
Milano, Italy

***Abstract*—This paper deals with the problem of automatically classifying clouds, comparing the performance of different neural networks using infrared-only images, visible-only images, and visible plus infrared images through constructing a labeled dataset from ground-based all-sky camera images. The images are preprocessed to remove the geometrical distortion introduced by the fisheye lenses of the cameras and then fed to an EfficientNet-based model; two possible ways of combining the two types of images are explored, fusing the feature vectors at two different steps. The performances of the three approaches are then compared. The results show that the models trained on visible plus infrared images, on average, are slightly better than those trained on the infrared-only and visible-only ones.**

***Index Terms*—Clouds, Cloud type Classification, Image Classification, Visible Images, Infrared Images**

I. Introduction

The ever-increasing awareness of global warming continues to drive interest in photovoltaic energy production, with both its advantages and problems. One such difficulty is the short-term predictability of the power output, which is strongly linked with cloud coverage; in this context, this paper explores the automatic recognition of cloud types to aid in such a task. Furthermore, it can be a relevant safety feature in airports, where clouds are currently being classified by specialized individuals to inform pilots.

According to the definition given by the World Meteorological Organization (WMO) [1], ten different cloud genera can be distinguished: Cirrus (Ci), Cirrocumulus (Cc), Cirrostratus (Cs), Altocumulus (Ac), Altostratus (As), Nimbostratus (Ns), Stratocumulus (Sc), Stratus (St), Cumulus (Cu), Cumulonimbus (Cb). In the literature, many different categorizations have been proposed to adapt categories to automatic recognition from ground images. The field showed the first publication in 1995 [3], which suggested a method to distinguish between five sky/cloud types. The results were not satisfactory, with an accuracy of 39%, but were later improved upon, and the foundational method was established by introducing the idea of features, in particular texture features. The strategy was widely used until the early 2010s, with various works such as [5], [9]. The former paper works on a 1500 images dataset with an algorithm based on a k-Nearest-Neighbor architecture, using 12 predetermined features describing color and texture. The accuracy on the selected dataset gets to 97% over seven different classes but drops to 75% when all images are taken into account. The work also elaborates on the advantages of ground-based all-sky imagers (ASI), compared to satellites, for cloud classification. Furthermore, these cameras excel in covering a wide patch of sky while combining both high temporal and spatial resolution at a lower cost than dedicated satellites, showcasing a promising strategy for short-term forecasting.

The development of deep learning (DL) techniques provided the availability to boost classification performance. The first tests were done in works such as [25] and [16]. The former, both timewise and as reported information, achieved moderate results by getting to 88% accuracy on the 10 categories of the CCSN data set. The latter obtained variable results, between 80% and 97%, depending on which of the 7 classes of the TJNU-GCD [10] was being considered.

The best recent results come from [8], which achieved 93.43% accuracy on the 7 classes of the TJNU-GCD and a slightly lower 92.35% on the more complex 11 classes. Those results are possible thanks to the extensive information in the datasets and the availability of experts in their labeling, providing exceptionally high-quality training material. [2] focused on a similar problem to determine pixel by pixel if the satellite image contained a cloud or its shadow. The final objective was to determine if the ground area had been correctly imaged, but it also indirectly performed a recognition of the shape of the cloud cover. In [12], the authors implemented a multi-evidence and multi-modal fusion network. In particular, it combined ground images, and weather parameters to enhance clouds' recognition. The data were obtained by installing the camera on an already existing meteorological station. The neural network, trained on 8000 images, performed well, obtaining an 88.96% accuracy across seven sky types. Lastly,

Corresponding author: Emanuele Ogliari (emanuelegiovanni.ogliari@polimi.it)

[7] implemented the usage of radar, lidar (light detection and ranging), and microwave radiometer data to estimate cloud fraction and both liquid and ice water contents. This approach, however, is still too expensive to be implemented on a distributed large-scale network.

Our paper, in the critically important area of class selection, adopts the classification proposed by the Singapore Whole Sky IMaging CATegories Database (SWIMCAT) [4]. The subdivision has been considered simple enough to properly test the Infrared (IR) imagery. Other relevant datasets of high quality were the rest of the TJNU family, which are described in [10], [11], [13]. Another very large dataset is available at: [26], and the infrared-based at: [15], [24].

This paper introduces a novel approach to cloud classification that combines DL and ground-based ASI cameras, both in the visible and infrared spectrum. In summary, the key contributions of this work include the creation of a labeled dataset of 10,080 RGB and IR images, the training of state-of-the-art open-source Convolutional Neural Networks (CNN) on the task of cloud classification, and one of the first usages of infrared ground cameras and a preliminary comparison with visible spectrum.

II. Dataset

A. Image Acquisition

The two datasets on which DL models have been trained are composed, respectively, of RGB images and thermal infrared (IR) sky dome images. The cameras that produce the images are installed on the rooftop of the Energy Department of Politecnico di Milano.

The RGB images have been obtained by a Mobotix model Q26 Hemispheric camera pointed at the sky. This camera can capture 360° images with a resolution of 3072x2048px [17]. The acquisition rate is one image per minute. The IR images have been obtained by a Reuniwatt model Sky InSight camera, which can capture 360° thermal infrared images with a resolution of 640x480px [20]. In this case, the acquisition rate is down to one image per 30 seconds, but only one per minute is kept to achieve a one-to-one correspondence with the visible spectrum set. The resolution of the IR camera, even, though lower, shouldn't have had much influence on the results. The two cameras are located at a distance of about 5 meters, resulting in nearly identical coverage, making it possible to compare fairly the performance obtained based on RGB and IR images.

The observation period has been from the 1^{st} of October 2023 to the 14^{th} of October 2023 both for RGB and IR, in the timeframe between 7:00 AM and 7:00 PM to ensure adequate daylight to label the images and have a fair comparison between the two pipelines. The final curated dataset consists of a total of 10,080 images, with 720 RGB and 720 IR images per day.

B. Image preprocessing

The proposed methodology follows the structure depicted in Figures 1 and 2 and is also described more robustly in [18].

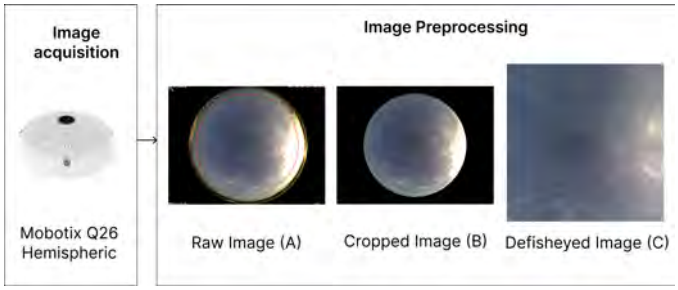


Fig. 1: The complete preprocessing pipeline for RGB images. The image is padded, cropped, and resized at the end.

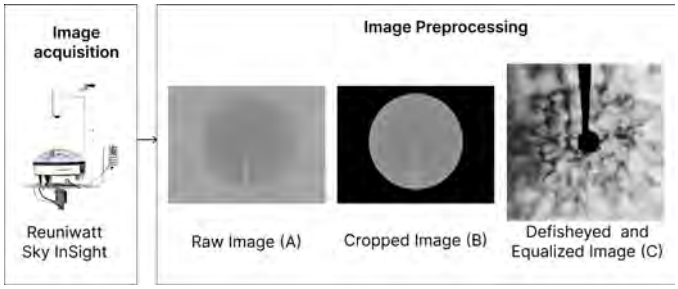


Fig. 2: The complete preprocessing pipeline for IR images. The image is padded, cropped, resized, and normalized at the end.

The raw images pass through a preprocessing stage before being labeled and fed into a detection block that trains and validates the performances obtained by the models.

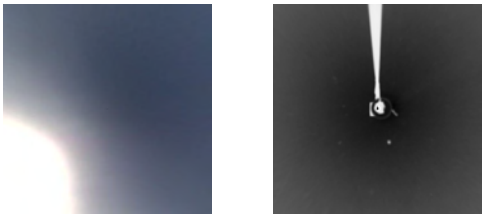
The preprocessing stage for RGB is structured as follows:

1) Circular padding: it involves zeroing out the pixel values outside the circular region that represents the sky, centered at (x, y) with a specified radius r_{crop}. The selected radius is slightly smaller than the actual radius of the original region to omit ground objects or buildings and focus on a slightly narrower section of the sky.
2) Region cropping: the extracted circular region is fit in a squared frame.
3) Frame resizing: the frame is resized to 512x512 pixels.
4) Image "defisheyeing": the geometric distortion induced by the fisheye lens of the camera is removed.

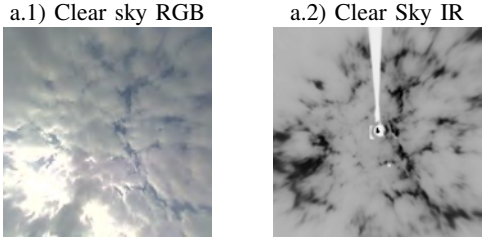
As depicted in Figure 2, for IR images the preprocessing pipeline is similar with an addition of a normalization step.

C. Image labelling

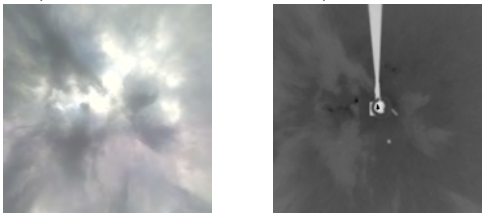
After curating the image set, the next step involves labeling the data. The choice of class types and their quantity was carefully made, tackling a tradeoff between the existing literature on cloud types and their visual characteristics, the available image data (which captures the sky and the clouds' shape and patterns in the visible spectrum), and the expertise of the annotators. Based on these considerations, the classification scheme proposed by the Singapore Whole sky Imaging CATegories Database (SWIMCAT) [4] was selected for annotating the images. The labels assigned to the images are not strictly



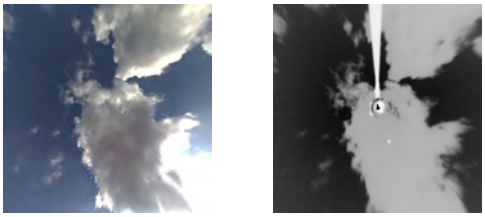
a.1) Clear sky RGB a.2) Clear Sky IR



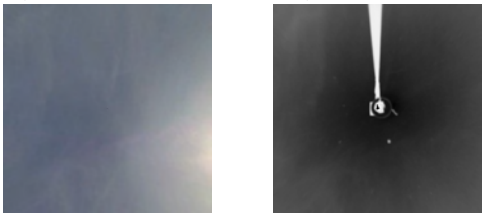
b.1) Patterned RGB b.2) Patterned IR



C.1) Thick Dark RGB c.2) Thick Dark IR



d.1) Thick White RGB d.2) Thick White IR



e.1) Veil RGB e.2) Veil IR

Fig. 3: Examples images showing the dataset and comparing the same view in RGB and IR. a) clear sky, b) patterned clouds, c) thick dark clouds, d) thick white clouds, and e) veil clouds.

mutually exclusive. Depending on the species and varieties of clouds, they may fit into multiple categories. For instance, Altocumulus Castellanus and Altocumulus Lenticularis could be classified either as Thick-White Clouds, while Altocumulus Floccus and Altocumulus Stratiformis Translucidus Undulatus exhibit a patterned appearance.

To minimize personal biases in the annotation process, three annotators were involved with a majority voting system to label each of the 10,080 RGB images. After labeling the images in the visible spectrum, the labels were transferred to the corresponding infrared images to complete the final dataset. Examples of the labelling process can be seen in Figure 3.

The dataset is partitioned using a temporal strategy, which is essential due to the sequential nature of the captured images taken one minute apart. These consecutive images often depict nearly identical scenes and are typically assigned the same class, resulting in sequences of similarly labeled images. To avoid bias in the validation and test sets, each sequence is preserved as a whole and assigned entirely to one of the three sets. This method prevents the splitting of sequences across the train, validation, and test sets, thereby maintaining the integrity of the evaluation process.

The dataset was then divided into three distinct folds, each with its own training, validation, and test sets. The test sets in each fold are unique and do not overlap with those from the other folds, allowing for comprehensive cross-validation of the model's performance across varied sky conditions. By using the temporal-based split strategy, the robustness of the evaluation is enhanced, as each test set captures different scenarios for each class.

III. Proposed Mehodology

A. Models

The task has been approached using the EfficientNet B0 model as a baseline. EfficientNet [23] is a family of CNNs designed to deliver great accuracy in image classification tasks while maintaining computational efficiency. The main innovation of EfficientNet lies in its distinctive approach to scaling CNN architectures, which balances model complexity across multiple dimensions, namely depth, width, and resolution, through a method called compound scaling [19].

The architecture of EfficientNet is built upon Mobile Inverted Bottleneck (MBConv) layers, which combines depthwise separable convolutions and inverted residual blocks, inspired by MobileNetV2 [22]. These layers are crafted to maximize efficiency while preserving strong representational capacity. Additionally, EfficientNet integrates the Squeeze-and-Excitation (SE) block [6], which helps the model prioritize essential features over less relevant ones. The SE block employs global average pooling to reduce the spatial dimensions of the feature map to a single channel, followed by two fully connected layers. These layers enable the model to learn channel-wise feature dependencies and generate attention weights that are applied to the original feature map, thereby emphasizing critical information.

The classification task was approached using a two-step methodology. In the first step, RGB images and IR images were employed separately for image classification. In the second step, both RGB and infrared (IR) images were used jointly to potentially enhance classification performance.

In the initial phase, a straightforward approach was adopted by following standard practices and leveraging transfer learning. Specifically, an EfficientNet-B0 network pre-trained on the ImageNet dataset was used as a feature extractor, with its weights kept frozen (not updated during the training procedure) to maintain the integrity of the learned representations.

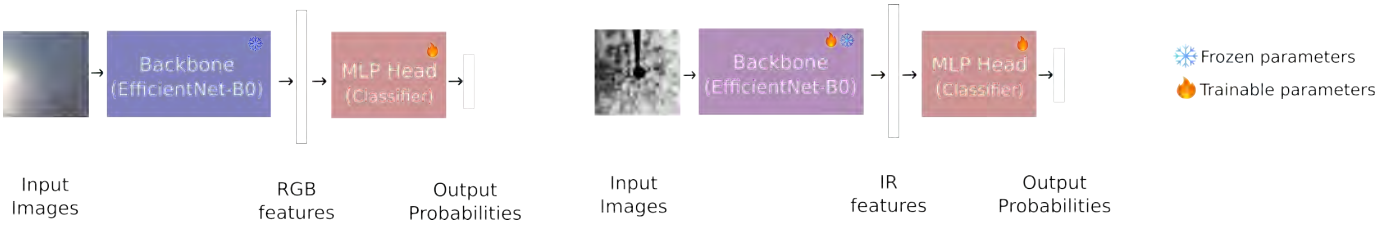


Fig. 4: RGB only model (left) and IR only model (right)

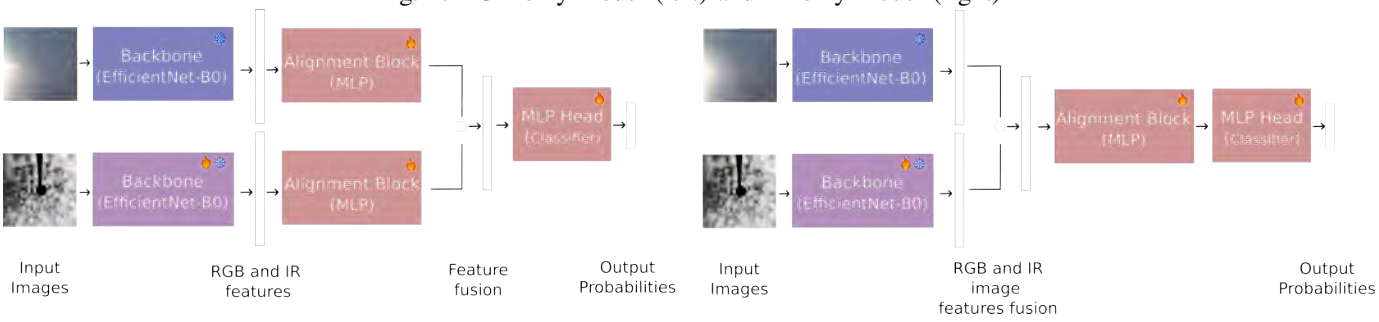


Fig. 5: RGB and IR joint models: late fusion (left), early fusion (right)

The features extracted by this backbone network were then passed to a compact feed-forward neural network functioning as a classifier. This classifier was responsible for categorizing the image features into the five distinct classes discussed in subsection II-C.

By using only RGB images, the model, depicted in Figure 4, relies on the rich color and texture information inherent in the visible spectrum. This approach establishes a baseline performance, allowing for a clear comparison when additional data modalities are introduced.

As a second step, the preprocessed IR images from the Reuniwatt Sky Insight camera were integrated into the pipeline. Adding this additional source of information, introduced the challenge of determining how to combine the two modalities for the classification task. Inspired by the work of [21], two different approaches were developed to address the multi-modality issue, as illustrated in Figure 5:

- In the first approach, shown in Figure 5 (left), the backbone network (EfficientNet-B0) is used to extract features from both the RGB and IR images. These feature vectors are then processed separately by two independent Alignment Blocks, each consisting of a multi-layer perceptron. This design allows the model to learn modality-specific representations, focusing on the unique characteristics of each modality. The ”aligned” feature vectors are then combined either through summation or concatenation. While this method preserves the independence of the modalities until the final classification step, it limits the model's ability to capture cross-modal interactions, which may reduce classification accuracy.
- In the second approach, shown in Figure 5 (right), the backbone extracts features from both image types, as in the previous case. However, here the features are fused earlier by either concatenating or adding them immediately to create a multimodal feature vector. An alignment block, consisting of a multi-layer perceptron, is then used to blend the information in this vector, producing a unified feature vector for classification. Early fusion enables interaction between the RGB and IR feature spaces, potentially yielding higher-quality features through the fusion process. Additionally, using a single alignment block is more parameter-efficient, reducing both model complexity and computational costs.

In both scenarios, two separate networks of the same type were used as backbones. For the RGB images, as in the pipeline utilizing only RGB images, the backbone weights were initialized with those pre-trained on ImageNet. For the IR images, however, the backbone was firstly separately trained from scratch to learn meaningful features specific to this data format. Using the same pre-trained network on ImageNet on both image types to extract features led to significant performance degradation because the features extracted from the IR images were irrelevant, given the weights tuned for RGB images. This mismatch introduced noise during the process of feature fusion and final classification. The classifier, along with the alignment block for the combined use of RGB and IR images, was trained accordingly.

Following common practices, data augmentation has been applied during training. The transformations applied to RGB images include: random horizontal and vertical flips (at 50% probability), random rotation (between -180 and 180 degrees), and random color jittering (brightness, contrast, and saturation factors ranging between 0.6 and 1.4 and a hue factor randomly selected between -0.05 and 0.05). For IR images, the same augmentation pipeline was used, excluding the random color jittering, as it makes less sense for grey-scale images.

The AdamW optimizer [14] was adopted with an initial learning rate of $1e^{-3}$. An early stopping callback was used,

TABLE I: Comparison of results across different configurations. *LF* stands for Late-Fusion and *EF* for Early Fusion.

	RGB Only			**IR Only**			**RGB + IR LF**			**RGB + IR EF**		
Class	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Clear sky	0.88	0.78	0.82	0.65	0.81	0.72	0.84	0.77	0.80	0.85	0.78	0.81
Patterned	0.58	0.49	0.53	0.64	0.64	0.64	0.71	0.68	0.69	0.71	0.64	0.67
Thick Dark	0.75	0.74	0.74	0.55	0.47	0.51	0.85	0.65	0.74	0.74	0.77	0.75
Thick White	0.71	0.75	0.73	0.73	0.67	0.70	0.72	0.75	0.73	0.74	0.75	0.74
Veil	0.87	0.91	0.89	0.90	0.73	0.81	0.89	0.90	0.90	0.89	0.91	0.90
Mean	0.82	0.83	0.82	0.80	0.73	0.76	0.84	0.83	0.83	0.84	0.84	0.84

monitoring validation accuracy with patience of 10 epochs, to accelerate the training process and avoid overfitting.

B. Metrics

In image classification, model performance is evaluated using various metrics that reflect different aspects of predictive quality. In this study, we employ precision, recall, and F1-score to deal with class imbalances.

Precision measures the proportion of correct positive predictions among all positive predictions made by the model. It indicates how well the model avoids incorrectly classifying an image as belonging to a certain class when it does not. It is calculated with the following equation:

$$P = \frac{Correctly\ Predicted\ Ture\ Class}{Total\ Predicted\ True\ Class} \tag{1}$$

Recall measures the proportion of actual positives that the model correctly identified. It indicates how well the model captures all relevant instances. It is calculated with the following equation:

$$R = \frac{Correctly\ Predicted\ Ture\ Class}{Total\ Actual\ True\ Class} \tag{2}$$

Finally, the F1-score is the harmonic mean of precision and recall. It provides a balance between the two metrics, thus it is particularly useful when the distribution of classes is imbalanced or when both false positives and false negatives are equally important. It is calculated with the following equation:

$$F1 = 2\frac{P \cdot R}{P + R} \tag{3}$$

In multi-class classification tasks, precision, recall, and F1-score are typically computed for each class individually. However, to evaluate the model's overall performance across all classes three averaging strategies, namely micro-averaging, macro-averaging, and weighted-averaging, are commonly used.

Micro-averaging aggregates the results from all classes by treating each single prediction equally, regardless of the class they belong to. This involves summing the true positives, false positives, and false negatives across all classes and then calculating precision, recall, and F1-score based on these combined values.

Macro-averaging, on the other hand, computes the metric (precision, recall, and consequently F1-score) for each class individually, and then takes the average of these values. In this way, each class contributes equally to the final score, regardless of the number of instances it contains.

IV. Results and discussion

After training and validation, the models were evaluated using the priorly restricted test dataset with the metrics described in subsection III-B. These scores have been calculated based on the test set which is comprised of images that the model has never seen before. To enhance the robustness of the evaluation procedure, a 3-fold cross-validation-like strategy was adopted. The metrics from the three different test sets are then averaged and reported in Table I, showing the classification reports.

As shown in Table Table I (RGB Only), the model trained exclusively on RGB images achieved an overall accuracy of 0.83. The "Clear sky" and "Veil" classes recorded the highest precision and F1 scores, with values of 0.88 and 0.87, respectively, suggesting that the model effectively distinguishes these classes due to their distinct visual features in RGB images. However, the "Patterned" and "Thick Dark" classes showed considerably lower precision and recall, with F1-scores of 0.49 and 0.74, respectively, as these images were frequently misclassified as "Thick White" clouds. This outcome is expected, as these classes inherently share visual similarities in the RGB spectrum, making them more challenging to differentiate. Specifically, several images were difficult to annotate due to the co-occurrence of different sky conditions, particularly those at the margins of temporal sequences, which reflect clouds from adjacent sequences.

When incorporating both RGB and IR images, both alignment strategies slightly improve the model's performance, particularly resulting in beneficial for patterned clouds while matching or even improving metrics for other classes. The macro-averaged metrics, which account for the imbalanced structure of the dataset, reflect this enhancement more robustly. This suggests that the features extracted from IR images, when combined with those from RGB images, are more discriminative than using only RGB features, thus helping the cloud classification task.

Comparing the results between the two proposed feature fusion strategies indicates that both methodologies are effec-

tive in capturing cross-modal interactions, with none of the two consistently outperforming the other. However, the early fusion approach is preferable due to its lower computational complexity.

V. Conclusions

This study demonstrates the use of ground-based images from ASI cameras for cloud-type classification, contrasting the traditional reliance on satellite imagery. Beyond visible spectrum images, the research also explores thermal infrared images, addressing challenges such as sun glare occlusion.

It has been determined that infrared images require additional preprocessing compared to RGB to facilitate annotation, and despite these efforts, the results indicate that further improvements are needed in the labeling.

The combined usage of RGB and IR images proved marginally better and showed great promise in the discrimination of particular cloud types.

Moreover, the sequential nature of the problem studied, where images are extracted in sequences and exhibit intrinsic temporal correlation, suggests the potential use of neural network architectures with recurrent connections to incorporate temporal information. It is noteworthy that one of the significant advantages of ground-based images is their higher temporal resolution (w.r.t. the ones coming from satellites), which opens up new possibilities for leveraging this additional feature.

Acknowledgements

We thank Alta Scuola Politecnica for assembling this amazing work group and Reuniwatt for providing the infrared all-sky imager Sky InSight™ and data used in this study.

References

[1] Cloud definition. https://cloudatlas.wmo.int/en/definition-of-a-cloud.html.

[2] Cesar Aybar, Luis Ysuhuaylas, Jhomira Loja, Karen Gonzales, Fernando Herrera, Lesly Bautista, Roy Yali, Angie Flores, Lissette Diaz, Nicole Cuenca, Wendy Espinoza, Fernando Prudencio, Valeria Llactayo, David Loaiza, Martin Sudmanns, Dirk Tiede, Gonzalo Mateo-Garcia, and Luis Gómez-Chova. Cloudsen12, a global dataset for semantic understanding of cloud and cloud shadow in sentinel-2. *Scientific Data*, 9:782, 12 2022.

[3] K A Buch, Jr and Chen-Hui Sun. Cloud classification using whole-sky imager data. 2 1995.

[4] Soumyabrata Dev, Yee Hui Lee, and Stefan Winkler. Categorization of cloud image patches using an improved texton-based approach. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 422–426, Sep. 2015.

[5] A. Heinle, Andreas Macke, and Anand Srivastav. Automatic cloud classification of whole sky images. *Atmospheric Measurement Techniques*, 3, 05 2010.

[6] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-Excitation networks. September 2017.

[7] Anthony Illingworth, Robin Hogan, E. O'Connor, Dominique Bouniol, Melissa Brooks, Julien Delanoë, D. Donovan, Jon Eastment, Nicolas Gaussiat, J. Goddard, Martial Haeffelin, Henk Klein Baltink, Oleg Krasnov, Jacques Pelon, Jean-Marcel Piriou, Alain Protat, H. Russchenberg, A. Seifert, A. Tompkins, and C.L. Wrench. Cloudnet continuous evaluation of cloud profiles in seven operational models using ground-based observations. *Bulletin of the American Meteorological Society*, 88:883–898, 06 2007.

[8] Sheng Li, Min Wang, Shuo Sun, Jia Wu, and Zhihao Zhuang. Cloud-densenet: Lightweight ground-based cloud classification method for large-scale datasets based on reconstructed densenet. *Sensors*, 23:7957, 09 2023.

[9] Lei Liu, Sun Xuejin, Feng Chen, Shijun Zhao, and Taichang Gao. Cloud classification based on structure features of infrared images. *Journal of Atmospheric and Oceanic Technology - J ATMOS OCEAN TECHNOL*, 28:410–417, 03 2011.

[10] Shuang Liu, Linlin Duan, Zhong Zhang, Xiaozhong Cao, and Tariq S. Durrani. Ground-based remote sensing cloud classification via context graph attention network. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022.

[11] Shuang Liu, Mei Li, Zhong Zhang, Xiaozhong Cao, and Tariq S. Durrani. Ground-based cloud classification using task-based graph convolutional network. *Geophysical Research Letters*, 47(5):e2020GL087338, 2020.

[12] Shuang Liu, Mei Li, Zhong Zhang, Baihua Xiao, and Tariq Durrani. Multi-evidence and multi-modal fusion network for ground-based cloud recognition. *Remote Sensing*, 12:464, 02 2020.

[13] Shuang Liu, Mei Li, Zhong Zhang, Baihua Xiao, and Tariq S. Durrani. Multi-evidence and multi-modal fusion network for ground-based cloud recognition. *Remote Sensing*, 12(3):464, 2020.

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[15] Qixiang Luo, Zeming Zhou, Yong Meng, Qian Li, and Miaoying Li. Ground-based cloud-type recognition using manifold kernel sparse coding and dictionary learning. *Advances in Meteorology*, 2018(1):9684206, 2018.

[16] Qi Lv, Qian Li, Kai Chen, Yao Lu, and Liwen Wang. Classification of ground-based cloud images by contrastive self-supervised learning. *Remote Sensing*, 14(22), 2022.

[17] Mobotix. Q26 hemispheric. https://www.mobotix.com/en/products/outdoor-cameras/q26-hemispheric.

[18] Paolo Pertino, Leonardo Pavarino, Simone Lomolino, Enrico Miotto, Daniele Rege Cambrin, Paolo Garza, and Emanuele Ogliari. Ground-based contrail detection by means of computer vision models: a comparison between visible and infrared images. 2024.

[19] Petru Potrimba. What is efficientnet? the ultimate guide, 2023.

[20] Reuniwatt. Sky insight ir camera official documentation. https://reuniwatt.com/en/products-and-services/247-all-sky-observation-sky-insight/.

[21] Elan Rosenfeld, Preetum Nakkiran, Hadi Pouransari, Oncel Tuzel, and Fartash Faghri. APE: Aligning pretrained encoders to quickly learn aligned multimodal representations. October 2022.

[22] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. January 2018.

[23] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[24] Sun Xuejin, T.-C Gao, D.-L Zhai, S.-J Zhao, and J.-G Lian. Whole sky infrared cloud measuring system based on the uncooled infrared focal plane array. *Hongwai yu Jiguang Gongcheng/Infrared and Laser Engineering*, 37:761–764, 10 2008.

[25] Jinglin Zhang, Pu Liu, Feng Zhang, and Qianqian Song. Cloudnet: Ground-based cloud classification with deep convolutional neural network. *Geophysical Research Letters*, 45(16):8665–8672, 2018.

[26] Liang Zhang, Kebin Jia, Pengyu Liu, and Chunyao Fang. Cloud recognition based on lightweight neural network. pages 1033–1042, 2020.