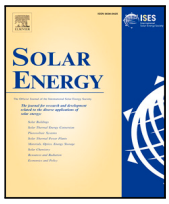
Solar Energy 300 (2025) 113822

Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener



Advancing semantic cloud segmentation in all-sky images: A semi-supervised learning approach with ceilometer-driven weak labels

David Magiera [a,*], Yann Fabel [a], Bijan Nouri [a], Niklas Blum [a], Dominik Schnaus [b], Luis F. Zarzalejo [c]

[a] German Aerospace Center (DLR), Institute of Solar Research, Calle Doctor Carracido 44, Almería, 04005, Spain
[b] Technical University of Munich, Chair of Computer Vision & Artificial Intelligence, Boltzmannstrasse 3, Garching, 85748, Germany
[c] CIEMAT Energy Department - Renewable Energy Division, Av. Complutense 40, Madrid, 28040, Spain

ARTICLE INFO

Dataset link: https://doi.org/10.5281/zenodo.14639170

Keywords:
Semi-supervised learning
Weak labels
Semantic cloud segmentation
All-sky imager
Ceilometer

ABSTRACT

Semantic segmentation of all-sky images provides high-resolution cloud coverage information useful for applications in meteorology, climatology, optical satellite downlink operations, and solar energy. While deep neural networks are highly effective for segmentation, their performance depends on large labeled datasets to learn complex visual features. To address this challenge, we introduce a semi-supervised learning approach for semantic cloud segmentation, combining advanced techniques such as ceilometer-driven weak labeling, pseudo-labeling, and consistency regularization. At the core of this approach is CloudMix, a novel data augmentation technique tailored specifically for cloud segmentation tasks. Our method begins with assigning weak labels to over 47,000 all-sky images using ceilometer data, which are combined with 616 manually labeled images to train a segmentation model. By employing pseudo-labeling and weak-to-strong consistency regularization, the model leverages both labeled and weakly labeled data effectively. The semi-supervised model surpasses a fully supervised baseline and a state-of-the-art model in pixel accuracy and mean Intersection over Union (mIoU) across validation, test and domain-shift test dataset. In particular, the detection of mid- and high-layer clouds improves significantly, with an increase in IoU of more than 7 and 9 percentage points on the test dataset. Furthermore, on the domain-shift test dataset, the semi-supervised model achieves over 20 and 27 percentage points higher mIoU than the baseline and state-of-the-art, respectively. These results underscore the robustness and generalization capabilities of the proposed method, making it a promising solution for cloud segmentation.

1. Introduction

Detecting clouds in ground-based imagery is important for several applications, including meteorology and climatology [1] and supporting optical satellite downlink operations to optical ground stations [2,3]. Solar energy is another increasingly important application of cloud detection. One of the challenges for the integration of solar energy is to manage the spatial and temporal variability of solar irradiance. Variations due to diurnal and seasonal changes can be easily accounted for and are predictable. Intra-hour and intra-minute variations in local solar irradiance are mostly caused by clouds [4,5], which are difficult to predict due to the complex dynamics of clouds. Especially large PV parks can benefit from reliable solar irradiance forecasts [6], for instance for ramp rate control [7,8], and can be optimized in terms of efficiency [9] and grid stability [10]. To assess the impact of clouds on solar irradiance, intra-hour forecasts, so-called solar nowcasts, are required. Such nowcasting systems typically rely on ground-based observations and measurements, like all-sky imagers and radiometers to obtain high temporal and spatial resolutions [11]. Using stereographic approaches based on multiple all-sky imagers, a typical spatial coverage is in the range of 10^1 to 10^2 km^2, covering the size of even very large PV parks [12,13]. In case of a large-scale network of all-sky imagers, several thousand square-kilometers can be covered [5], offering opportunities to anticipate short-term power production of entire regions. In such physics-based nowcasting systems, the underlying models are composed of a series of processing steps that describe the physical phenomena. These steps, typically include cloud detection, classification, tracking, geolocation and transmittance estimation [14]. Semantic segmentation holds significant potential for enhancing cloud tracking, geolocation and the analysis of clouds radiative effects, especially under complex multi-layer conditions [15]. Cloud coverage information can also serve as an input feature in data-driven approaches [16]. Other

* Corresponding author.
E-mail addresses: mag.david12@yahoo.de, david.magiera@dlr.de (D. Magiera).

https://doi.org/10.1016/j.solener.2025.113822
Received 13 January 2025; Received in revised form 27 June 2025; Accepted 22 July 2025
Available online 15 August 2025

data-driven approaches train specific models based on the prevailing cloud types in the sky [17].

Therefore, the detection and classification of clouds in ground-based imagery has been a subject of increasing research activity over the past two decades. Historically, thresholding-based methods were early approaches to distinguish between clear-sky and cloud pixels. Since a clear atmosphere is dominated by Rayleigh scattering the sky appears blue while clouds, dominated by Mie scattering, appear white. Different thresholding methods have been proposed to distinguish cloudy- and clear sky-pixels in all-sky images, like a fixed threshold on the red-blue ratio [18], the red-blue difference [19], or a combination of multiple thresholds [20]. Other methods include also the green color channel [21], transforming the image into other color spaces such as hue-saturation-intensity (HSI) and deciding based on saturation [22]. All these methods work under certain conditions, but their performance is significantly compromised in presence of excessive saturation of the color channels and in turbid atmospheric conditions with a high concentration of aerosols in the atmosphere, due to shifts in the red-blue-green (RGB) ratio. Hence, over the past decade, machine learning-based methods have been applied to cloud detection. First, shallow fully-connected architectures were utilized [23], followed by deep convolutional neural networks (CNN) for cloud detection [24,25]. These methods showed superior accuracy while being less-prone to high atmospheric turbidities compared to the existing thresholding methods. However, all of these methods only distinguish between clear sky and cloudy pixels, but not between different cloud layers or cloud generas as defined by the World Meteorological Organization (WMO) [26]. The first approach to differentiate between the different cloud generas on pixel-level was proposed by [27]. A simplified classification, differentiating between the three cloud-layers was proposed by [28].

A major challenge with these methods is the requirement for a large number of pixel-level annotated images. The process of annotating all-sky images at the pixel level is difficult due to ambiguities, which makes it extremely time consuming, and usually not feasible for large volumes due to economic and time constraints. This issue has been addressed by approaches like self-supervised learning [29], a form of unsupervised learning that does not require manually labeled data, but automatically generates pseudo-labels for unlabeled data based on a pretext task. [28] used self-supervised learning for semantic cloud segmentation to pretrain a model on a large set of unlabeled all-sky images, followed by fine-tuning on a smaller set of pixel-level annotated all-sky images. Semi-supervised learning is another paradigm that leverages a small amount of labeled data and a large amount of unlabeled data simultaneously to effectively train deep neural networks. In computer vision, semi-supervised learning has been successfully applied to image classification [30,31] and semantic segmentation [32]. Self-training, a form of semi-supervised learning, has been applied to semantic cloud segmentation by [33].

In this work, we adopt the categorization of clouds into three layers (low, mid and high) as defined by the WMO [26], considering their distinct optical characteristics on average. We propose a ceilometer-driven approach to assign weak labels to all-sky images on a large scale, to reduce the reliance on extensive human labeling. These weak labels, combined with a small set of labeled images, are leveraged for semi-supervised learning. Our method builds upon the weak-to-strong consistency regularization framework [31,32] and incorporates a novel data augmentation technique, CloudMix, specifically designed for semantic cloud segmentation. To our knowledge, this work introduces the first approach to integrate ceilometer measurements directly into the learning process for purely camera-based semantic cloud segmentation. While [34] demonstrated that combining ceilometer data with camera imagery improves cloud classification accuracy compared to camera-only methods, our approach offers a unique advantage. By using the ceilometer measurements for training purposes only, and not requiring them for inference, our approach can be easily applied to any site that is equipped with a suitable all-sky imager on its own.

The remainder of this work is organized as follows: Section 2 details our method for ceilometer-driven weak labeling, the resulting dataset, and labeled training, validation and test datasets. Section 3 presents our proposed semi-supervised learning approach, including our novel data augmentation technique. In Section 4, we validate our methodology by comparing the segmentation performance of a semi-supervised model trained with our approach to a fully-supervised model trained exclusively on labeled data and a state-of-the-art semantic cloud segmentation model on a validation and two test datasets. Finally, Section 5 concludes the work and provides a brief outlook.

2. Cloud image datasets

In this section we describe the data utilized for training, validation, and testing of our model. First, we present details on the hardware and image properties. Subsequently, we briefly introduce the labeled dataset employed for model training and validation. Thereafter, a comprehensive description of the generation of a novel weakly labeled dataset is presented. Finally, the test data will be described briefly.

2.1. Data acquisition

All camera and sensor data was acquired at CIEMAT's (Spanish research institute: Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas) Plataforma Solar de Almería (PSA) located in southern Spain at 37° 5′ 38″ N and 2° 21′ 32″ W. Images were taken with all-sky imagers based on off-the-shelf surveillance cameras from Mobotix (models Q25, Q26, Q71). The Mobotix Q25 and Q26 models capture images with a resolution of 4.35 megapixels, while the Mobotix Q71 model captures images with a resolution of 8.29 megapixels. However, due to computational reasons, the images are cropped and resized to a square format of 512 × 512 pixels prior to being passed to the network. Furthermore, a camera mask is used to remove static objects from the surrounding site environment and very low elevation masks ($< 10°$), as there is a lot of uncertainty in the annotation of these regions. The exposure time is set to a fixed value of 160 µs, and no solar occulting devices are installed. The cameras are configured to capture images in 30s intervals from sunrise to sunset, yielding approximately 1000 to 1600 images per day. The cloud base height measurements were acquired with a ceilometer manufactured by Lufft of type CHM15k-Nimbus.

2.2. Labeled dataset for model training

The labeled training data is taken from [28]. It comprises 770 labeled images. Clouds are categorized as either low-, mid-, and high-layer clouds adopted from the categorization by the WMO, combining the 10 main cloud genera into three cloud layers based on typical cloud base heights. The same dataset split into training and validation sets as in [28] is applied: 80% (616) training samples and 20% (154) validation samples.

2.3. Ceilometer-driven weak labeling of all-sky images

The generation of the weakly labeled dataset constituted the initial phase of the implementation of the proposed training method. The primary hypothesis is that the cloud base height measurements obtained from a ceilometer could be screened for conditions in which only a single cloud layer was present. A single-layer condition is defined as a period during which only clouds with cloud base heights from one of the three cloud layers are detected by the ceilometer. Based on the assumption that the majority of clouds captured by the all-sky imagers in close proximity to the ceilometer will be of the same cloud layer, weak labels can be assigned to the respective images on an image level.

Table 1
The thresholds for the heuristics applied to the ceilometer measurements to assign image-level weak labels to all-sky images. The height levels defined by the WMO [26] for mid-latitude regions like southern Spain are given in parentheses for reference.

Cloud-layer	min. threshold [m]	max. threshold [m]
Low-layer		2000 (2400)
Mid-layer	3000 (1800)	6000 (8000)
High-layer	8000 (6000)	

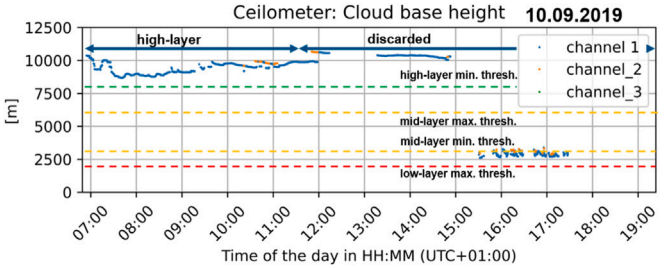


Fig. 1. Ceilometer-driven weak labeling process. The first 5 h of the day were labeled as high-layer cloud conditions by the algorithm. Afterwards the heuristic was not fulfilled anymore as clouds with cloud base heights inside the high-layer and mid-layer thresholds are detected by the sliding window operation.

The temporal extent for the heuristic to be valid is defined as a sliding window operation looking four hours into the past and four hours into the future. The thresholds for each cloud layer, specified in Table 1, were adapted from the WMO [26] with a margin of safety with the objective of limiting the prevalence of ambiguous cloud conditions in the dataset. For instance, mid-layer clouds have typical cloud base heights ranging from 1800 to 8000 meters in mid-latitude regions such as southern Spain, according to the WMO. These thresholds exhibit a significant degree of overlap with the typical maximum cloud base heights for low-layer clouds (2400 m) and the typical minimum cloud base heights for high-layer clouds (8000 m). In order to reduce the degree of overlap, the maximum cloud base height for low-layer clouds was reduced to 2000 m, the interval for mid-layer clouds was narrowed to [3000 m, 6000 meters], and the minimum for high-layer clouds was increased to 8000 m. Furthermore, at least one cloud detection must occur within a temporal extent of ten minutes for the heuristic to be considered valid to reduce the number of images with minimal cloud coverage and noise of the ceilometer, as these images contribute little useful information to the training process. The weak labeling procedure for ceilometer measurements for one day is illustrated in Fig. 1.

Moreover, only images with sun elevation angles exceeding 20° and with Linke turbidity values below 4 were considered during the weak-labeling procedure. This was done to exclude images where the distinction between cloudy- and clear sky-pixels is too difficult or subject to a high degree of uncertainty. In addition, all parts of the images with elevation angles below 30 degrees were masked. This is because the ceilometer only provides a point measurement of the cloud base height vertically above the installation site. The masking thus reduces the field of view in the images and ensures that the distance of the captured clouds in the images to the location of the ceilometer is limited. Consequently, the assumption for the weak-labeling based on the ceilometer measurements still holds. An illustrative example image and its corresponding image-level weak label are presented in Fig. 2 for each cloud layer. A qualitative assessment of weak label accuracy is provided in Appendix A.

We applied the weak labeling algorithm to automatically label images from an all-sky imager installed in close proximity to the ceilometer from July 2019 until October 2021. The procedure yielded 47595 weakly labeled images. Of these, 21341 images were weakly labeled as low-layer, 19885 images as high-layer, and 6396 images as mid-layer, which constitutes the minority class in this case. Oversampling was applied during training to neutralize the class imbalance. As captured in Fig. 3, the dataset encompasses a diverse range of atmospheric conditions containing a broad variety of sun elevation angles and atmospheric turbidities.

2.4. Test datasets

We labeled a total of 48 additional images to enable a comparison between the proposed method, its fully-supervised baseline, and a state-of-the-art model as described in [28]. Of these, 36 test images were captured using Mobotix Q25/Q26 camera models, which contain the same CMOS chip as the camera used for acquiring the training data. Additionally, 12 images were labeled from the Mobotix Q71 camera model to assess the generalization capabilities under changing camera hardware, representing a domain-shift scenario. For the remainder of this work, the 36 images will be referred to as the *test dataset*, while the 12 images will be referred to as the *domain-shift test dataset*. The images were selected from the years 2021 and 2023, with an evenly distributed selection across months to account for seasonal variations.

Fig. 4 shows the pixel-level class distributions and the image-level cloud condition distributions of the training, validation, test and domain-shift test datasets. In the training and validation datasets, a higher proportion of pixels belong to the sky and low-layer classes and a lower proportion of pixels belong to the mid- and high-layer classes compared to the test and domain-shift test datasets. Also, the two test datasets contain a larger share of all-sky images with multi-layer cloud conditions compared to the training and validation datasets. This makes the test datasets more challenging for semantic cloud segmentation. First, because mid- and high-layer clouds are typically more difficult to detect accurately than clear-sky or low-layer clouds [28]. Second, multi-layer cloud conditions increase the complexity of the scenery and make it difficult to clearly distinguish the cloud boundaries.

3. Proposed semi-supervised learning method

In the following section, we present our method for training a semantic cloud segmentation model with labeled and weakly labeled all-sky images. After a brief general overview of our training architecture, we describe the design of the deployed pseudo-labeling and consistency regularization strategies in detail.

3.1. General overview

The architecture of our semi-supervised learning method is based on a student–teacher architecture, as illustrated in Fig. 5. Both models are semantic cloud segmentation models with an identical CNN architecture. Prior to the semi-supervised learning procedure, the teacher model is trained exclusively on the labeled training dataset with fully supervised learning. Next, the student model is trained on the labeled training images and the weakly labeled images, and is guided by the teacher model's predictions on the weakly labeled images. The weights of the teacher model are kept constant during the optimization of the student model.

From a high-level perspective, the semi-supervised training of the student model can be viewed as two independent streams of image processing. In the first stream, the labeled images are processed and in the second stream, the weakly labeled images. A separate loss function is calculated for each stream. For the labeled data, model optimization is performed by calculating the standard cross-entropy loss, denoted as L_L, which is the same loss function utilized to train the teacher model beforehand. The weakly labeled images are utilized to calculate an additional consistency loss, denoted as L_C. The consistency loss is calculated based on generated pseudo-labels. Those are obtained from the predictions of the teacher model that are combined with the weak labels derived from the ceilometer data. Both streams calculate their

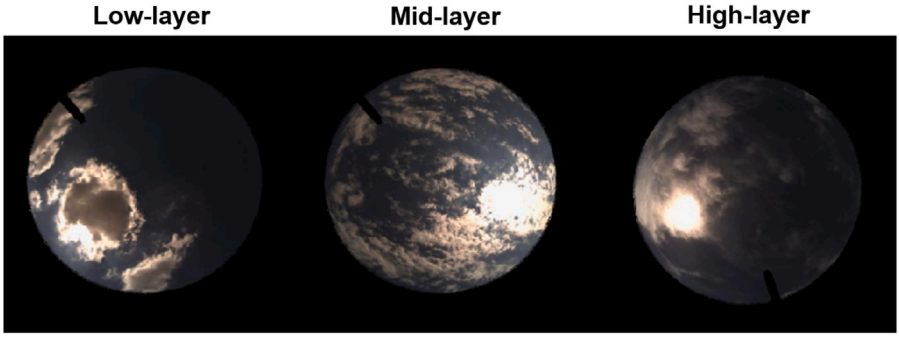


Fig. 2. One example for each image-level weak label from the generated weakly labeled dataset.

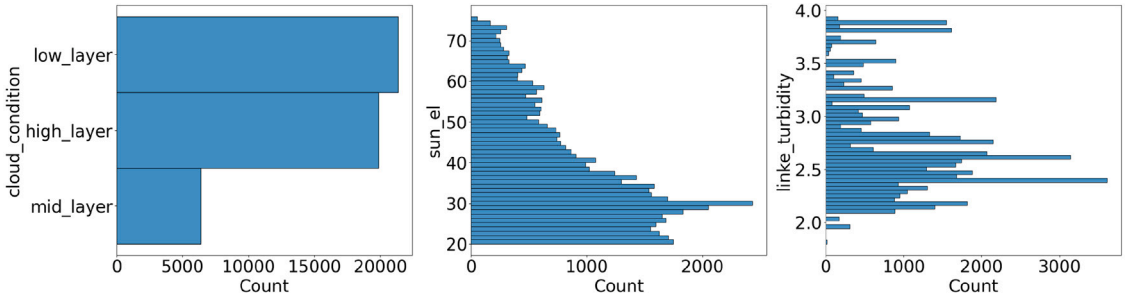


Fig. 3. Data distributions of the generated weakly labeled dataset. From left to right: By weak label, sun elevation angle and linke turbidity.

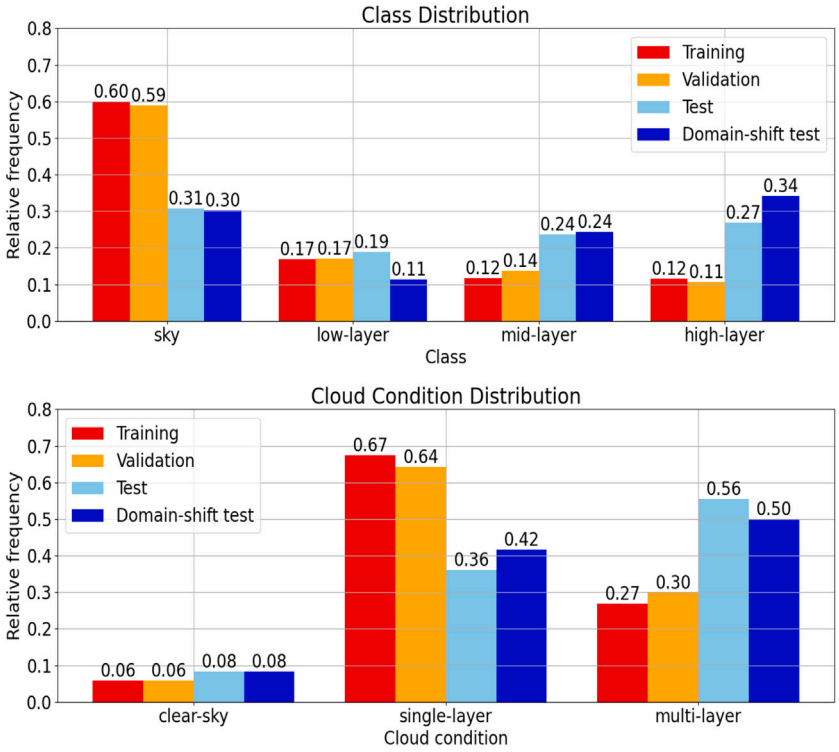


Fig. 4. Comparison of the data distributions of the validation and test datasets. Top: By class labels on the pixel-level. Bottom: By cloud conditions on the image-level.

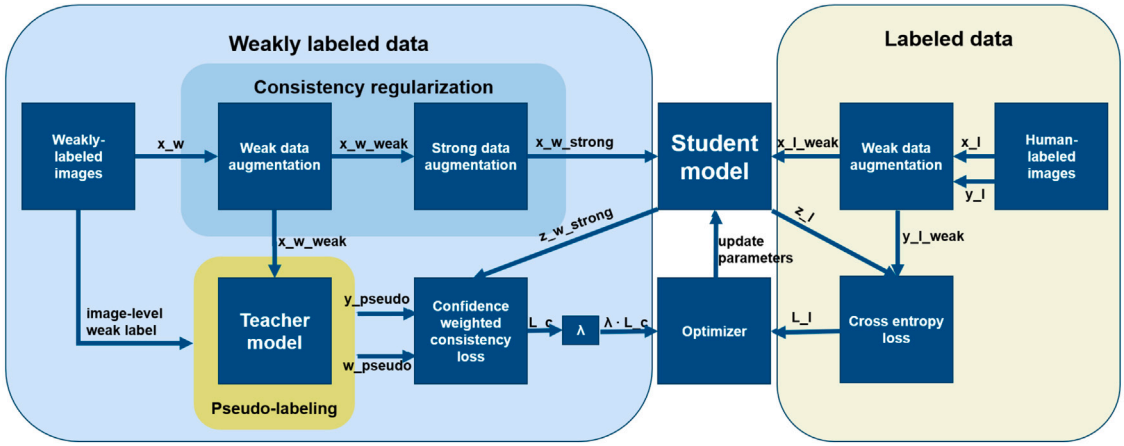


Fig. 5. A high-level overview of the training architecture of our proposed semi-supervised learning approach for semantic cloud segmentation.

respective losses independently, but are combined into a total loss function, denoted as L_{total}, for a joint model optimization, as defined in Eq. (1).

$$L_{total} = L_L + \lambda \cdot L_C \tag{1}$$

The hyper-parameter λ, referred to as the consistency weight, determines the extent to which the optimization of the labeled data loss L_L is constrained by the consistency loss L_C. In this work, we set λ with a fixed value throughout the training process for reasons of stability and simplicity.

3.2. Weak label enhanced pseudo-labeling

Pseudo-labeling is a common method in semi-supervised learning, where artificial labels, or pseudo-labels, are generated from unlabeled data or, in this case, weakly labeled data. We generate the pseudo-labels by fusing the predicted segmentation masks of the teacher model with the image-level weak label.

First, images from our weakly labeled dataset, denoted as $\mathbf{x}_w$, are passed to the teacher model to make predictions, denoted as $\mathbf{z}_w$. Next, pixel-wise classes are determined by computing the *argmax* function, as defined in Eq. (2). Then, all pixels corresponding to one of the three cloud layers are overwritten with the image-level weak label representing the cloud layer observed by the ceilometer. The remaining pixels classified as clear sky remain the same. Furthermore, pixel-wise confidences are obtained from the prediction $\mathbf{z}_w$ of the teacher model using the softmax regularization, as defined in Eq. (3), which yields a probability distribution over the possible classes for each pixel. The specific pixel-wise confidence weight, denoted as $\mathbf{w}_{pseudo}$, is derived by indexing the probability distribution with the identified class in the pseudo-label for the specific pixel, as defined in Eq. (4). These confidence weights serve to quantify the consensus between the prediction of the teacher model and the image-level weak label. In other words, they provide a measure of alignment between the labeled training images and the specific image-level weak label, determined through the cloud base height measurements of the ceilometer. This process is depicted in Fig. 6. In this particular instance, pixels predicted by the teacher model as high-layer clouds (green) do not align with the image-level weak label (mid-layer). Consequently, these pixels exhibit confidence weights approaching zero, as indicated by the blueish hue in the confidence weighted mask. These confidence weights are utilized to mitigate the influence of pixels with a high degree of uncertainty in the pseudo-label during the calculation of the consistency loss. A detailed definition of the calculation of the consistency loss will be provided towards the end of this section.

$$\mathbf{y}(\mathbf{z}) = \arg\max_{\mathbf{i}}(\mathbf{z}) \tag{2}$$

$$\sigma_{SM}(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}, i = 1...C \tag{3}$$

$$\mathbf{w}_{pseudo}(\mathbf{z}) = \sigma_{SM}(\mathbf{z})_{i=y_{pseudo}} \tag{4}$$

Pseudo-labeling alone is reported to perform poorly, especially on small amounts of labeled data. This is due to over-fitting to noise in the pseudo-labels and confirmation bias as addressed by [35]. Typically, the generated pseudo-labels are used for consistency regularization, which is also done in this approach.

3.3. Consistency regularization

Consistency regularization is a key technique in semi-supervised learning to leverage unlabeled data and weakly labeled data. The main idea of consistency regularization is to enforce the same predictions for similar perturbed views of the same input image. Our method uses a weak-to-strong consistency regularization framework popularized by [31] for semi-supervised image classification and recently transferred to the semi-supervised semantic segmentation domain by [32].

3.3.1. Weak-to-strong consistency regularization

For our case of semantic cloud segmentation, the weakly labeled images $\mathbf{x}_w$ are first transformed using weak data augmentations to obtain weakly augmented views $\mathbf{x}_{w\,weak}$. Weak augmentations in the context of semantic cloud segmentation are simple transforms as horizontal and vertical flipping, image rotations and minor resizing of the input image. These transformations increase the variability of input without changing the semantic content of the image, specifically its ground truth segmentation mask. Then, the weakly augmented images $\mathbf{x}_{wweak}$ are augmented a second time with strong data augmentations, resulting in strongly augmented views $\mathbf{x}_{wstrong}$. In the context of semantic cloud segmentation, strong data augmentations are transformations that change the image content more substantially, such as color jittering, including changes in contrast, brightness, and saturation, or Gaussian blurring. In addition to color jittering and Gaussian blurring, we developed and deployed a third strong augmentation technique specific to semantic cloud segmentation, called *CloudMix*, for even stronger perturbation on an image level.

3.3.2. CloudMix data augmentation

Semi-supervised semantic segmentation represents a more challenging problem than semi-supervised image classification, as discussed in [36]. From a semantic segmentation perspective, color jittering and Gaussian blurring are often insufficient for augmenting images for effective consistency regularization. Hence, semi-supervised semantic segmentation requires domain specific strong data augmentation.

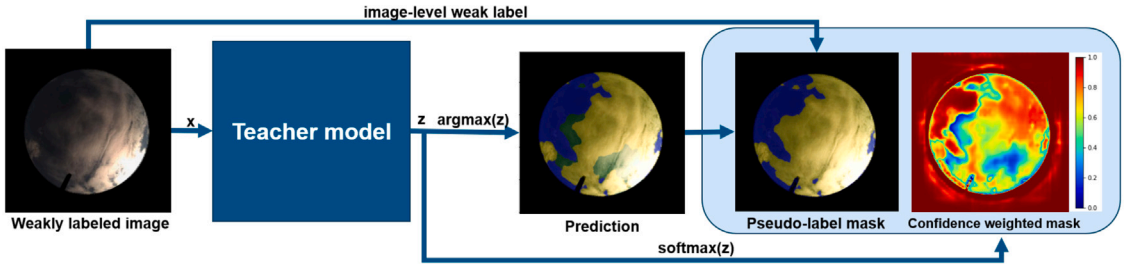


Fig. 6. The workflow of pseudo-label generation using a teacher model and image-level weak labels. (Colormap pseudo-label mask: blue: sky; red: low-layer; yellow: mid-layer; green: high-layer).

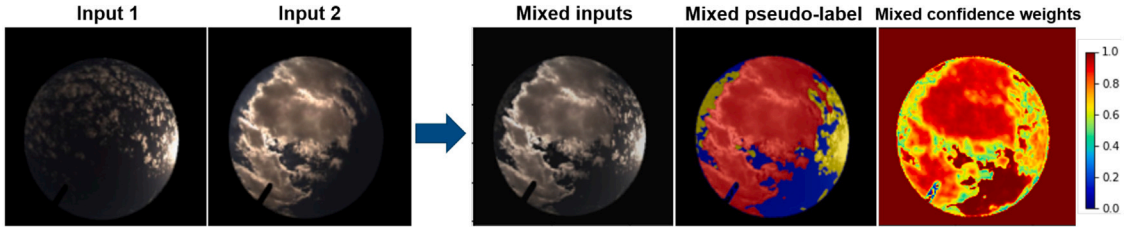


Fig. 7. Mixing of two images and their respective pseudo-label masks using the proposed CloudMix data augmentation technique. (Colormap mixed pseudo-label: blue: sky; red: low-layer; yellow: mid-layer; green: high-layer).

Table 2
Class hierachy utilized for CloudMix data augmentation.

Class hierarchy	Class label
1	clear-sky
2	high-layer
3	mid-layer
4	low-layer

A commonly used strong data augmentation technique for semantic segmentation is CutMix proposed by [37]. The CutMix augmentation technique involves cutting out parts of an image and inserting it into another image. In the same way, the respective pixel-level labels are combined into a new segmentation mask. However, the original approach, only overlays rectangular regions of arbitrary size, creating hard cuts at the boundaries and resulting in unnatural cloud sceneries.

Inspired by this, we propose a new data augmentation technique called CloudMix. In contrast to the original approach, CutMix, CloudMix respects the cloud shape observed in the images that are blended. Additionally, since our clouds layers correspond to different cloud base heights, a physically correct order can be obtained when blending image pairs. This approach ensures the preservation of the original cloud boundaries and yields a more "natural" mixed image. This is achieved by applying the hierarchy defined in Table 2. Formally the process can be defined as follows: *Two images can be mixed by adopting the pixel and its label with the higher class-hierarchy from both images for each pixel position respectively*. To illustrate, a mid-layer pixel is expected to overlay clear-sky and high-layer pixels, but would be expected to be overlaid by low-layer pixels. The CloudMix process for one image pair with low-layer and mid-layer clouds is depicted in Fig. 7. In instances where two images with an identical cloud type are to be combined, the pixels of the initial image are accorded precedence during the process.

In this study, CloudMix was applied alongside color jittering and Gaussian blurring to achieve sufficient data augmentation for weak-to-strong consistency regularization.

3.3.3. Confidence weighted consistency loss

We conclude the weak-to-strong consistency framework by defining the consistency loss function. The weakly augmented images $\mathbf{x}_{wweak}$ are utilized to generate pseudo-labels $\mathbf{y}_{pseudo}$ and the respective confidence weights $\mathbf{w}_{pseudo}$ by fusing the predictions of the teacher model with the image-level weak labels as stated by our pseudo-labeling strategy. The student model in contrast predicts on the strongly augmented images, denoted by $\mathbf{x}_{wstrong}$. The predictions of the student $\mathbf{z}_{wstrong}$, the generate pseudo-labels $\mathbf{y}_{pseudo}$, and the confidence weights $\mathbf{w}_{pseudo}$, are used to calculate a confidence weighted cross entropy loss, also denoted as consistency loss, defined in Eq. (5).

$$L_C(\mathbf{y}_{pseudo}, \mathbf{w}_{pseudo}, \mathbf{z}_{wstrong}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} w_i \cdot y_{ij} \log(\sigma_{SM}(z_{ij})) \tag{5}$$

4. Experiments

In this section, we evaluate the effectiveness of our proposed method for semantic cloud segmentation by comparing the segmentation performance of three models:

1. A Semi-supervised model trained with our method on the labeled dataset (616 images) and weakly labeled dataset (47595 images) presented in Section 2.
2. A Fully-supervised model trained exclusively on the labeled dataset (616 images) as a baseline.
3. A State-of-the-art semantic cloud segmentation model from [28], fine-tuned on the same labeled dataset (616 images). This model will be referred to as Fabel 2022 model for the rest of this section.

Segmentation is evaluated on semantic segmentation metrics, such as pixel accuracy and Intersection over Union (IoU), as described in Section 4.3. In addition to the validation dataset (154 images), we also evaluate the models on the test dataset (36 images) and domain-shift test dataset (12 images), presented in Section 2.

4.1. Model training and selection

All models were trained on a single Nvidia RTX A5000 GPU with 16 GB GPU RAM on a Dell Precision 7560 laptop and implemented in PyTorch Lightning [38] a lightweight wrapper for PyTorch [39].

Table 3
Hyperparameter selection for the training of deep cloud segmentation models.

Hyper-parameter	Fully-supervised	Semi-supervised	Fabel 2022
Input size	512 × 512	512 × 512	512 × 512
Arch., backbone	DeepLabv3+, ResNet50	DeepLabv3+, ResNet50	U-Net, ResNet34
Initialization	ImageNet	ImageNet	Self-supervised
Epochs	40	100	2 × 20
Batch size	4	8, 56 (weakly labeled)	4
Training samples	616	616, 47595 (weakly labeled)	616
Optimizer	AdamW	AdamW	Adam
Learning rate	1e−4	5e−4	1e−3, 1e−4
Scheduler	OneCycleLR	OneCycleLR	OneCycleLR
Normalization mean	[0.1662, 0.1688, 0.1571]	[0.1662, 0.1688, 0.1571]	[0.1739, 0.1696, 0.1715]
Normalization std	[0.1811, 0.1732, 0.1536]	[0.1811, 0.1732, 0.1536]	[0.1376, 0.1297, 0.1175]
Consistency weight λ	–	2	–
Training time	20 min	7 h	–
Data augmentations	Weak	Weak and strong	Horizontal and vertical flipping

4.1.1. Hyperparameter optimization and model selection

During the development stage of our models, hyperparameters such as batch sizes and the optimal number of epochs were tuned experimentally. Suitable learning rates were determined using the learning rate finder proposed by [40] and were dynamically scheduled using the one-cycle policy as described in [41]. The model weights were updated using the AdamW [42] optimizer, with a default weight decay of 1×10^{-2}. The models utilized for the final evaluation were selected based on the model checkpoints with the best mean Intersection over Union (mIoU) score on the validation dataset.

4.1.2. Fully-supervised baseline model

The model architecture is a convolutional neural network (CNN) with encoder–decoder structure based on the DeepLabv3+ [43] architecture with a ResNet50 [44] encoder, which is initialized with ImageNet [45] weights before training. Training is conducted for 40 epochs with a batch size of 4 and a learning rate of 1×10^{-4}. Data augmentation includes flipping, rotating, and random cropping, as detailed in Table 4. Training is repeated three times to account for random fluctuations. Each training run takes approximately 20 min.

4.1.3. Semi-supervised model

The model architecture and initialization are identical to those employed for the fully-supervised baseline model, thereby facilitating a fair comparison between the two models. The semi-supervised model is trained for 100 epochs with a batch size of 8 and a learning rate of 5×10^{-4}. The training process can be extended over a greater number of epochs in comparison to the training of the fully-supervised counterpart. This is due to the substantial quantity of weakly labeled images, which prevents the model from overfitting to the labeled training images. The same weak augmentations (see Table 4) are applied as for the baseline model, while additionally strong augmentations are used. In addition to color jittering and Gaussian blurring, we apply our novel CloudMix augmentation (see Table 5). Each batch contains 7 weakly labeled samples and 1 human-labeled sample, ensuring consistent guidance from human-labeled data during optimization. A greater proportion of weakly labeled samples is included in each batch to enhance consistency loss computation, improving gradient estimation and stabilizing training. To accommodate even more weakly labeled samples per optimization step, gradient accumulation [46] is set to 8, resulting in an effective batch size of 64, comprising 56 weakly labeled and 8 human-labeled images. The consistency weight (λ) is fixed at 2 for simplicity. The fully-supervised baseline model is utilized as the teacher model for pseudo-label generation during the training of the semi-supervised model. Training takes approximately 7 h (see Table 3).

Table 4
Utilized weak data augmentation techniques.

Augmentation	Intensity	Probability
RandomResizedCrop	+/−10%	0.5
RandomRotation	[0°, 360°]	1.0
RandomHorizontalFlip		0.5
RandomVerticalFlip		0.5

Table 5
Utilized strong data augmentation techniques.

Augmentation	Intensity	Probability
ColorJitter	+/−10% brightness, contrast, saturation	0.8
GaussianBlur	$\sigma \in (0.75, 1.25)$	0.5
CloudMix		1.0

4.2. Fabel 2022 model

The architecture is based on a U-Net [47] with a ResNet34 [44] encoder. The model was pre-trained on a clustering-based pretext task [29] with self-supervised learning on 286477 all-sky images and fine-tuned on the labeled dataset of 616 images, presented in Section 2. It should be noted that the model training was not conducted as part of this work; rather, the model was utilized as a pre-trained model. For further implementation and training details, please refer to the original work of [28].

4.3. Metrics

The overall semantic segmentation performance is evaluated using pixel accuracy and mean Intersection over Union (mIoU). Pixel accuracy calculates the number of correctly predicted pixels divided by the

Table 6
Accuracy and mIoU on the validation and test datasets.

Metric	Validation dataset			Test dataset		
	Fabel 2022	Baseline	SSL	Fabel 2022	Baseline	SSL
Accuracy	84.28	88.20	**88.67**	68.27	67.63	**70.92**
mIoU	74.71	79.80	**80.52**	53.07	52.80	**56.84**

number of all pixels and is defined as

$$pixelAccuracy = \frac{\sum_{c=1}^{C} TP_c}{N} \tag{6}$$

where C denotes the number of classes, TP_c the number of true positives for the respective class c and N the total number of pixels. The mIoU assesses the overlapping area of predicted and groundtruth pixels by their union, defined as

$$mIoU = \frac{1}{N} \sum_{c=1}^{C} IoU_c \cdot w_c \tag{7}$$

where IoU_c denotes the Intersection over Union (IoU) for the class c and $w_c = TP_c + FN_c$ the support of the respective class calculated as the sum of true positives and false negatives for the respective class. The IoU_c for a specific class is defined as

$$IoU_c = \frac{TP_c}{TP_c + FP_c + FN_c} \tag{8}$$

where TP_c are the true positives, FP_c the false positives and FN_c the false negatives for class c. Furthermore, we also evaluate the precision and recall for each class. The precision indicates the proportion of true positives predictions among all positive predictions, defined as

$$precision_c = \frac{TP_c}{TP_c + FP_c} \tag{9}$$

The recall quantifies the proportion of true positives to the total ground truth positives, defined as

$$recall_c = \frac{TP_c}{TP_c + FN_c} \tag{10}$$

The border part of the all-sky images, indicating masked image areas (see Section 2.1), is neglected for the calculation of our metrics as this would distort the results.

4.4. Results

We compare the semantic segmentation results of the three models on the validation and test datasets. First, we examine the results for the multi-layer segmentation, followed by an examination of the binary segmentation results.

4.4.1. Multi-layer segmentation

The semi-supervised model was the best model on the validation set with an accuracy of 88.67% and mIoU of 80.52% as presented in Table 6. It slightly outperformed the fully-supervised baseline, with less than one percentage point difference in both metrics. In contrast, the gap to the Fabel 2022 model was more pronounced, showing a difference of 4.39% in accuracy and 5.81% in mIoU.

Similarly, the semi-supervised model was found to perform best on the test dataset, achieving an accuracy of 70.92% and an mIoU of 56.84%. It surpassed the Fabel 2022 model by 2.6% in accuracy and 3.7% in mIoU, and showed even slightly greater improvements over the fully-supervised model. Notably, while the fully-supervised baseline achieves higher accuracy and mIoU than the Fabel 2022 model on the validation dataset, it is outperformed by the Fabel 2022 model on the test dataset. This could be caused by overfitting of the fully-supervised model to the specific cloud conditions in the validation dataset, as the model was selected based on the best mIoU score on this specific dataset. Nevertheless, it seems to generalize poorly to unseen data, which explains the lower mIoU on the test dataset and highlights the downsides of fully-supervised learning.

Table 7
Classwise IoU, precision and recall on the test dataset.

Metric	Class	Fabel 2022	Baseline	SSL
IoU	Clear-sky	79.34	79.56	**80.13**
IoU	Low-layer	**55.17**	52.78	49.32
IoU	Mid-layer	34.24	29.58	**37.41**
IoU	High-layer	38.22	42.75	**52.63**
Recall	Clear-sky	94.63	**94.96**	89.28
Recall	Low-layer	**77.50**	75.27	64.33
Recall	Mid-layer	55.09	46.98	**57.08**
Recall	High-layer	43.34	49.31	**66.77**
Precision	Clear-sky	83.08	83.07	**88.67**
Precision	Low-layer	65.69	63.85	**67.87**
Precision	Mid-layer	47.50	44.39	**52.05**
Precision	High-layer	**76.36**	76.26	71.31

The significant drop in accuracy and mIoU, by over 17 and 23 percentage points respectively, between the validation and test datasets can be attributed to two main factors. First, as discussed in Section 2.4 and shown in Fig. 4, the relative frequency of the challenging mid- and high-layer cloud classes is substantially higher in the test set compared to the validation set. Second, the test dataset contains a significant number of images with multi-layer cloud conditions, which are particularly challenging for the models to predict accurately.

In terms of classwise IoU, the semi-supervised model achieved the best results on the test dataset across all classes except the low-layer class as shown in Table 7. The most significant improvements were observed for the mid- and high-layer classes, previously identified as the most challenging to predict by [28]. In contrast, the Fabel 2022 and baseline models demonstrated higher recall for the clear-sky and low-layer classes but exhibited lower precision compared to the semi-supervised model, indicating that while they identified more true positives for these classes, they also produced more false positives. This difference may suggest that the semi-supervised model effectively mitigates bias toward the majority classes in the labeled training dataset (clear-sky and low-layer), a common challenge in imbalanced data scenarios, as discussed by [48]. Despite these advances, mid-layer clouds remain the most difficult to predict due to optical similarities with low- and high-layer clouds in many cases. While weak labeling and semi-supervised learning techniques have shown some improvement, challenges remain.

Next, we analyze the confusion matrices of the three models on the test dataset, as shown in Fig. 8. The majority of misclassifications for all models occured between adjacent cloud layers, such as between low-layer and mid-layer clouds or between mid-layer and high-layer clouds. In addition, some high-layer clouds were misclassified as clear sky and vice versa by all models. The semi-supervised model improved in the detection of mid-layer clouds, showing less confusion with low-layer clouds, but at the cost of more misclassification of low-layer clouds as mid-layer clouds, compared to the baseline and Fabel 2022 models. Overall, more high-layer cloud pixels were correctly predicted by the semi-supervised model, as less confusion occurred with the mid-layer and clear-sky classes. However, clear sky pixels are also more likely to be misclassified as high-level clouds, compared to the baseline and Fabel 2022 models. This strengthens the argument, that the semi-supervised model may be less biased towards the majority classes in the training data than the models fine-tuned exclusively on labeled all-sky images, as mentioned in the previous paragraph. For detailed results on the validation dataset, including the corresponding confusion matrices and classwise metrics, please refer to Appendix B.

Examining segmentation examples, as shown in Fig. 9, provides insight into the qualitative improvements in cloud segmentation and highlights the types of scenes that remain particularly challenging. In

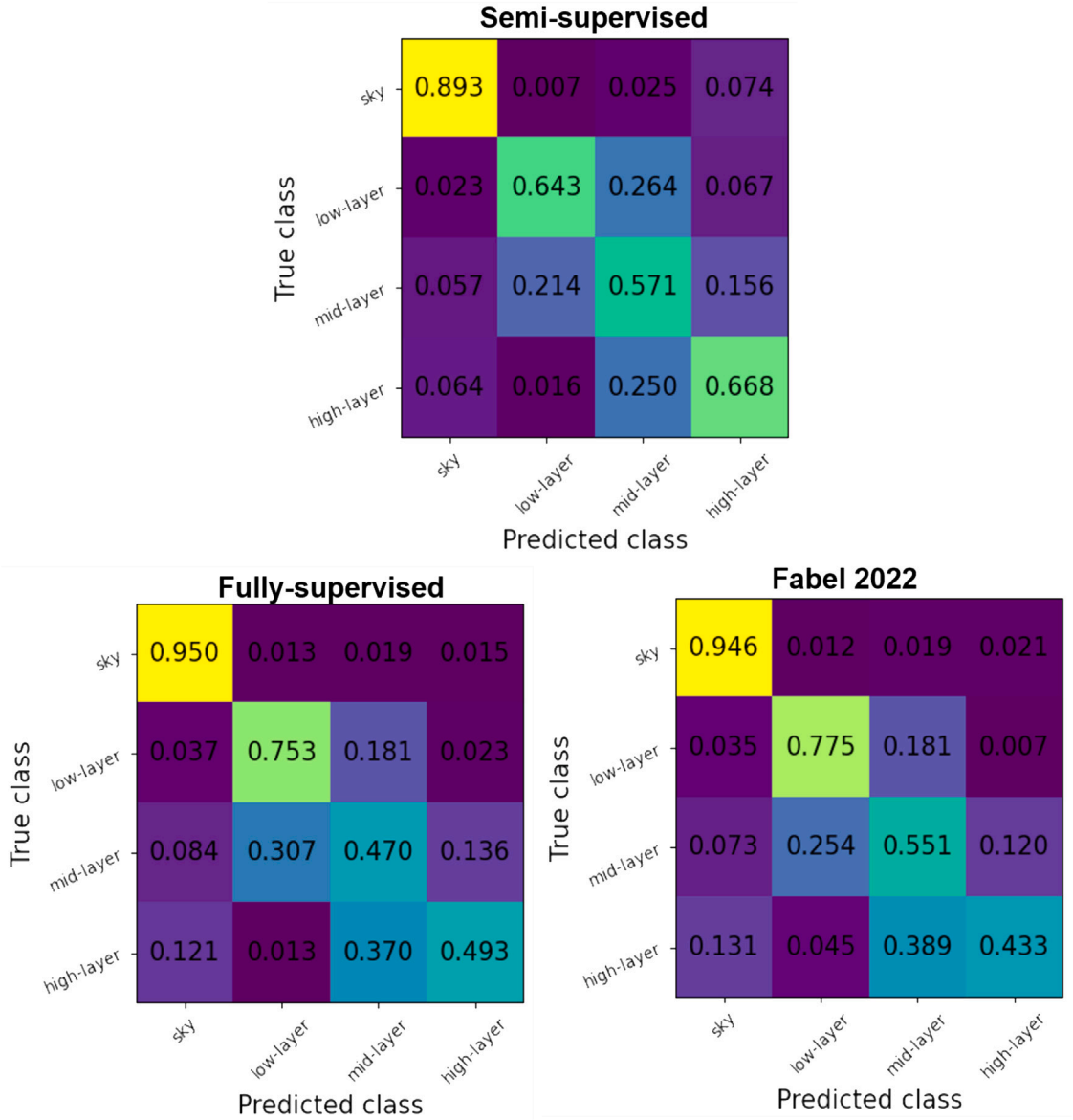


Fig. 8. Confusion matrices on the test dataset.

Table 8
Accuracy and mIoU of binary segmentation on the validation and test datasets.

Metric	Validation Dataset			Test Dataset		
	Fabel 2022	Baseline	SSL	Fabel 2022	Baseline	SSL
Accuracy	94.49	94.44	**94.63**	92.30	92.24	**93.07**
mIoU	89.74	89.55	**89.82**	86.19	86.22	**87.34**

example (a), the semi-supervised model largely succeeded in correctly classifying fine cirrus clouds in the high-layer, whereas the Fabel 2022 and baseline models misclassified these as mid-layer or even low-layer clouds. However, the predictions do not fully align with the ground truth, particularly in regions where the clouds are barely visible, posing significant challenges for both prediction and annotation. Similarly, example (b) illustrates a case where mid-layer clouds were correctly identified only by the semi-supervised model. Examples (c) and (d) highlight the complexity of scenes with multiple cloud layers in a single all-sky image. Particularly in example (d), where stratus-like overcasts coexist with low-layer clouds, distinguishing between cloud layers remains a major challenge due to reduced illumination and contrast.

4.4.2. Binary segmentation

Next, we compare the three models for binary segmentation on the validation and test datasets. As presented in Table 8, the semi-supervised model demonstrates minor improvements over both the Fabel 2022 and baseline models on both datasets. On the validation dataset, the differences in accuracy and mIoU were minimal, with margins of less than one percentage point. On the test dataset, the semi-supervised model achieved approximately one percentage point higher accuracy and mIoU than the Fabel 2022 and baseline models, indicating that the proposed method is also advantageous for binary segmentation. It is important to note that all three models perform well on the binary segmentation task with pixel accuracy over 94% on the validation dataset and over 92% on the test dataset. Most confusions come from thin high-layer clouds with clear-sky, which are often not easy to distinguish even for human experts.

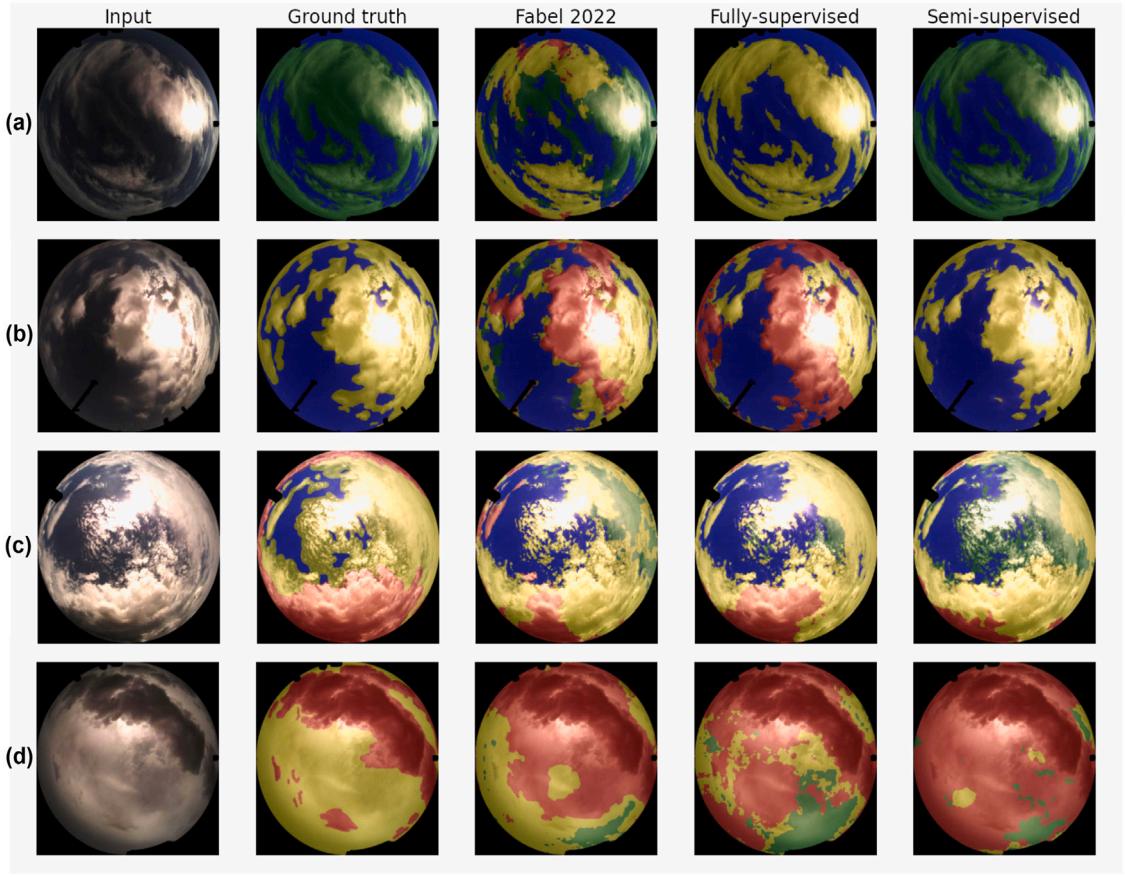


Fig. 9. Examples of all-sky images, ground-truth masks and the predictions of the evaluated models of the test dataset with the corresponding prediction accuracy and mIoU. Examples (a) and (b) represent cloud conditions, where our method leads to improvements. Examples (c) and (d) represent the complexity of scenery with multiple cloud layers in a single all-sky image. The images have been adjusted in brightness and contrast for readability. (Colormap ground truth masks and predictions: blue: sky; red: low-layer; yellow: mid-layer; green: high-layer).

Table 9
Accuracy and mIoU on the 12 all-sky images from the domain-shift test dataset.

Metric	Fabel 2022	Baseline	SSL
Accuracy	52.29	62.11	**77.70**
mIoU	37.82	45.24	**65.30**

4.4.3. Multi-layer segmentation under domain-shift

Finally, we compare the segmentation quality of the three models under domain-shift conditions. To this end, accuracy and mIoU were evaluated on the 12 all-sky images of the domain-shift test dataset, which contains images captured using a different camera model (Mobotix Q71) than the training data (Mobotix Q25/Q26). As shown in Table 9, the semi-supervised model significantly outperformed both the baseline and the Fabel 2022 models in mIoU, with improvements of over 20 and 27 percentage points, respectively. While the performance of the semi-supervised model remained comparable to its results on the in-domain test dataset, the Fabel 2022 and baseline models experienced substantial degradation under domain-shift conditions. This suggests that our method can lead to improved generalization capabilities when faced with changes in camera hardware. However, further investigation is required to draw definitive conclusions, as a dataset of 12 images is limited in covering very versatile cloud conditions.

5. Conclusions

In this work, we presented a novel semi-supervised learning approach for semantic cloud segmentation, developed to address the challenge imposed by limited labeled data. To this end, we incorporated advanced training techniques, including ceilometer-driven weak labeling of all-sky images, pseudo-labeling, and weak-to-strong consistency regularization. As part of this work, over 47,000 all-sky images were assigned with image-level weak labels using ceilometer data. Furthermore, we introduced CloudMix, a data augmentation technique specifically tailored for semantic cloud segmentation, which mixes all-sky image pairs and their corresponding ground truth masks. To validate our method, we compared the segmentation results of our novel semi-supervised model against a baseline model trained in a fully-supervised manner and a state-of-the-art model from the literature. These comparisons were conducted on a validation dataset comprising 154 all-sky images, a test dataset containing 36 images, and a domain-shift test dataset from a different camera model containing 12 images. Consistently, our semi-supervised model demonstrated superior performance across all datasets, particularly in terms of pixel accuracy and IoU. Notably, the detection of underrepresented cloud types improved, such as mid- and high-layer clouds. Most importantly, the semi-supervised model maintained stable segmentation quality under domain shift conditions, whereas the performance of the other models declined significantly. These results highlight the enhanced robustness and generalization capabilities of our method, making it a more suitable option for real-world deployment across diverse imaging devices and environments.

However, differentiating between cloud types remains a significant challenge. Particularly mid-layer clouds are often confused with low- or high-layer clouds, due optical similarities in many cases. This visual ambiguity not only impacts classification performance but also complicates the creation of accurate ground-truth annotations. Similar

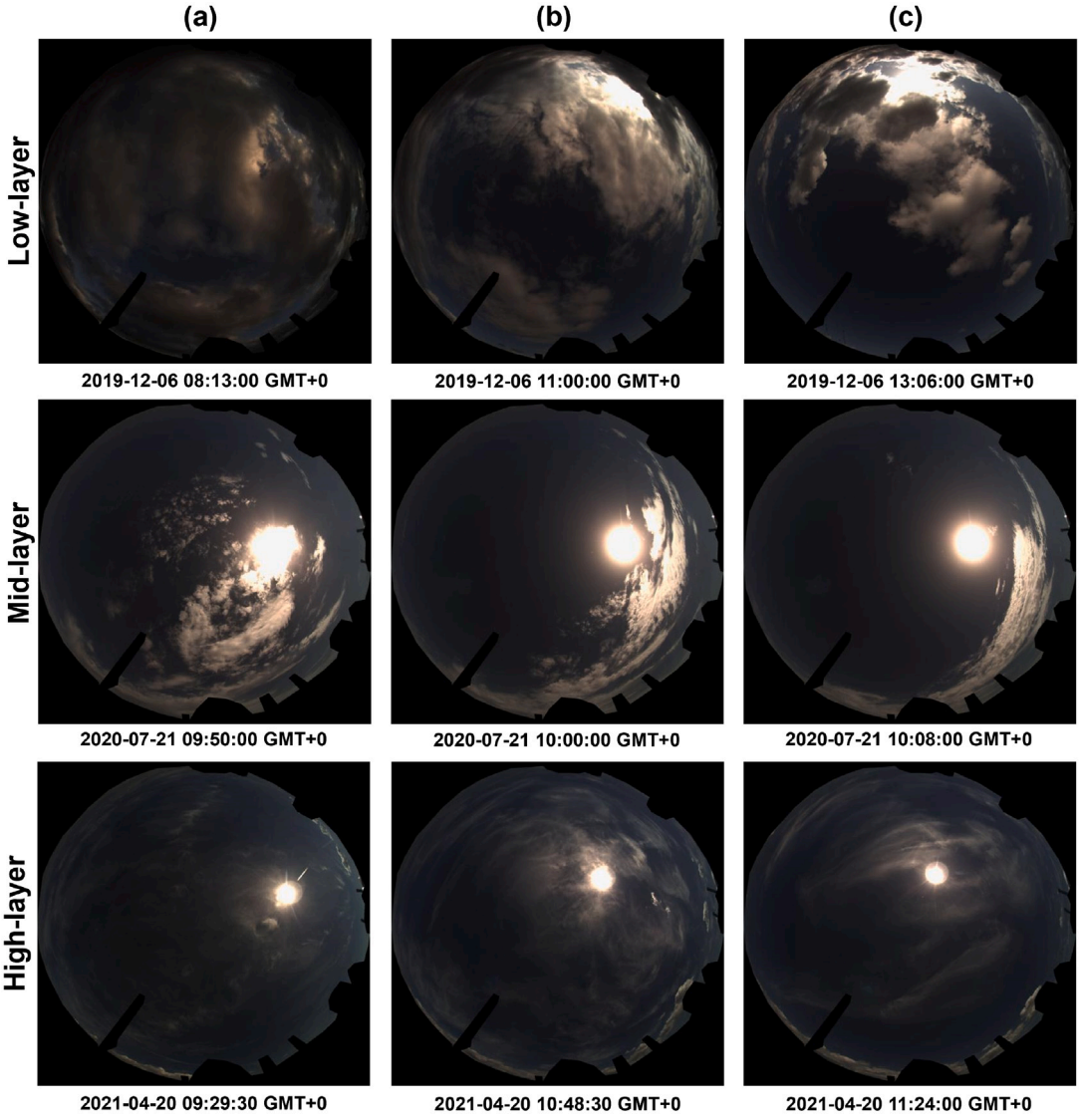


Fig. A.1. Examples of weak label generation time intervals demonstrating the effectiveness of ceilometer-driven weak labeling. (a) shows images at the beginning of weakly labeled time intervals. (b) shows images taken between the beginning and end of the weakly labeled time intervals. (c) shows images at the end of the weakly labeled time intervals.

difficulties can be observed in complex multi-layer conditions. This is especially the case in overcast scenarios where reduced illumination and missing cloud boundaries further hinder both segmentation accuracy and annotation reliability. To address these challenges, further research on ground-based cloud detection is necessary. For instance, incorporating additional information about cloud motion could assist to distinguish different layers due to varying motion patterns. Also, applying semi-supervised learning to larger datasets from multiple cameras and observation sites could further improve generalization capabilities. Moreover, this study did not examine the impact of CloudMix independently because the primary focus was to demonstrate the effectiveness of the complete semi-supervised framework, which includes CloudMix. However, future research could explore the extent to which different data augmentation techniques, such as CloudMix, add value to semantic cloud segmentation in isolation.

CRediT authorship contribution statement

David Magiera: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yann Fabel:** Writing – review & editing, Supervision, Software, Data curation, Conceptualization. **Bijan Nouri:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Niklas Blum:** Writing – review & editing, Software, Conceptualization. **Dominik Schnaus:** Writing – review & editing, Methodology, Conceptualization. **Luis F. Zarzalejo:** Writing – review & editing, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the German Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection through the AuSeSol project (grant agreement no. 67KI21007A), based on a decision by the German Bundestag.

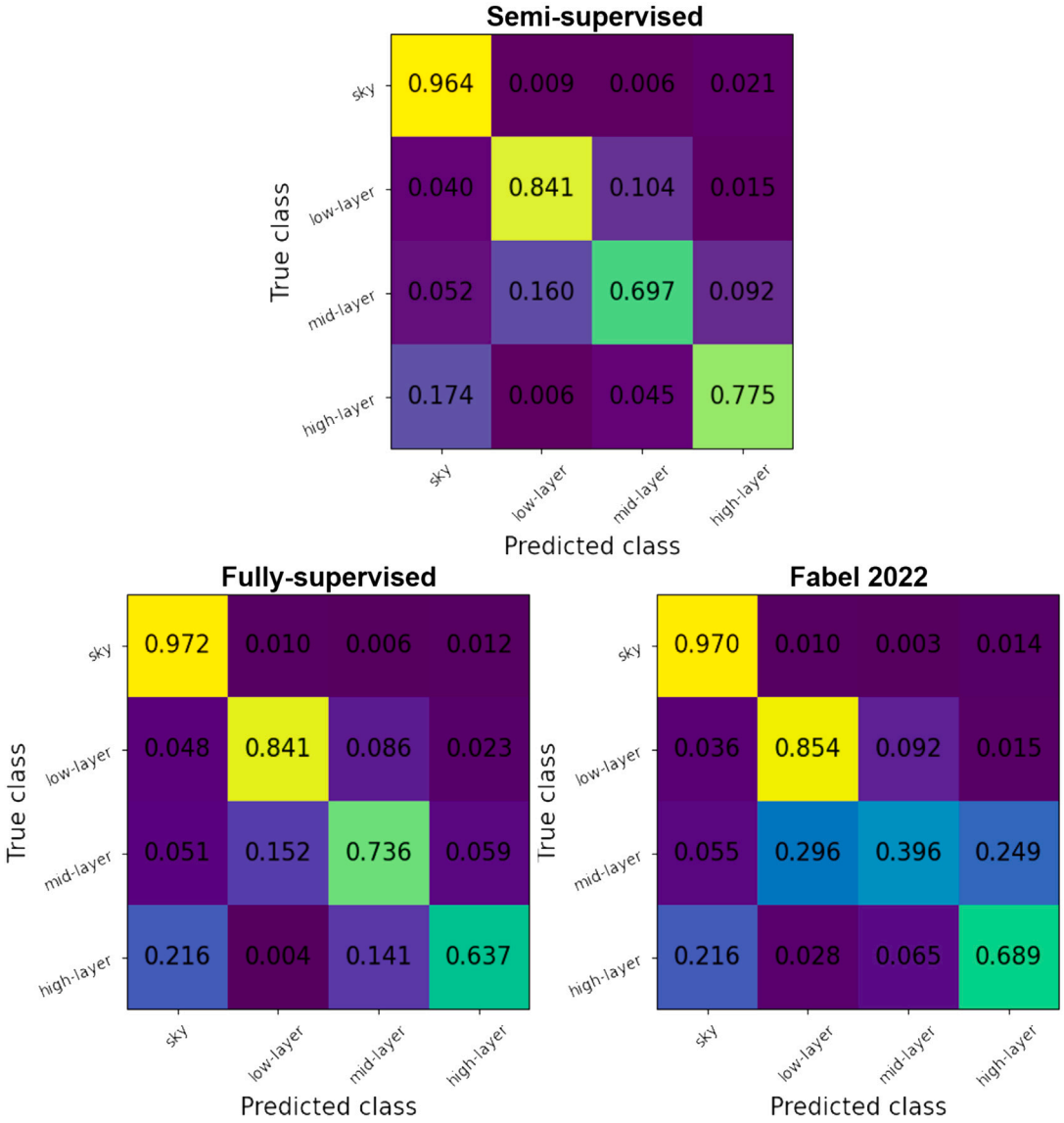


Fig. B.1. Confusion matrices for the semi-supervised, the fully-supervised baseline and the Fabel 2022 models on the validation dataset.

Appendix A. Qualitative verification of weak label generation

While a comprehensive quantitative comparison with manually annotated labels would counteract the goal of reducing annotation effort, we conducted a qualitative verification to assess the reliability of the weak labels. Therefore, we randomly selected one weakly labeled time interval for each loud layer (low, mid and high) and manually inspected representative images (first, middle and last) to verify the consistency and plausibility of the assigned labels, as shown in Fig. A.1.

The images in the first row are weakly labeled as low-layer and primarily depict thick cumulus clouds, which typically occur at lower altitudes. In contrast, the clouds in the third row are thin cirrus clouds, which are prevalent at high altitudes above 6000 m [26]. The clouds in the second row appear thicker than the high-layer clouds but not as dark as the low-layer clouds. This is a typical optical appearance for mid-layer clouds. Visibly, the images in each row primarily capture clouds from the assigned layer. This suggests that our assumption that weak labels can be assigned based on ceilometer measurements of cloud base height is mostly correct. Still the weak labels are not perfect, because clouds that do not travel above the zenith are not captured due to the point measurement of the ceilometer. This explains why low-layer clouds appear on the edges of the images in the third row. However, since these clouds are so far away, they only occupy a small portion of the images and can be mostly disregarded by masking all parts of the images with elevation angles below 30 degrees, as discussed in Section 2.3.

Appendix B. Detailed results on the validation dataset

Classwise metrics on the validation dataset are shown in Table B.1. The semi-supervised model achieves slightly lower IoU (less than 1%) for the clear sky, low-layer and mid-layer classes compared to the fully-supervised baseline and Fabel 2022 models, but improves by over 7% and 13% in IoU for the high-layer class.

The confusion matrices on the validation dataset are shown in Fig. B.1. For the semi-supervised model, there is less confusion of high-layer clouds with clear sky and mid-layer clouds compared to the fully-supervised baseline and Fabel 2022 models.

Data availability

The used datasets in this work will be made publicly available under the following https://doi.org/10.5281/zenodo.14639170.

Table B.1
Classwise IoU, precision and recall for the semi-supervised, the fully-supervised baseline and the Fabel 2022 models on the validation dataset.

Metric	Class	Fabel 2022	Baseline	SSL
IoU	Clear-sky	**91.38**	91.25	91.37
IoU	Low-layer	66.15	**72.61**	72.25
IoU	Mid-layer	33.59	**59.14**	58.52
IoU	High-layer	48.54	54.15	**61.68**
Recall	Clear-sky	97.04	**97.15**	96.36
Recall	Low-layer	**85.40**	84.12	84.06
Recall	Mid-layer	39.64	**73.17**	69.66
Recall	High-layer	68.89	63.70	**77.53**
Precision	Clear-sky	94.00	93.76	**94.61**
Precision	Low-layer	74.59	**84.15**	83.73
Precision	Mid-layer	68.76	75.05	**78.55**
Precision	High-layer	62.17	**78.32**	75.11

References

[1] C.J. Hahn, W.B. Rossow, S.G. Warren, ISCCP cloud properties associated with standard cloud types identified in individual surface observations, J. Clim. 14 (2001) 11–28, http://dx.doi.org/10.1175/1520-0442(2001)014<0011:ICPAWS>2.0.CO;2.

[2] D. Giggenbach, M.T. Knopp, C. Fuchs, Link budget calculation in optical LEO satellite downlinks with on/off-keying and large signal divergence: A simplified methodology, Int. J. Satell. Commun. Netw. 41 (5) (2023) 460–476, http://dx.doi.org/10.1002/sat.1478, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/sat.1478.

[3] M.T. Knopp, D. Giggenbach, M. Giggenbach, B. Nouri, Assessment of ciemat's plataforma solar de almeria as a ground station site for optical LEO satellite downlinks, in: International Conference on Space Optics, Antibes Juan-les-Pins, France, 2024, pp. 21–25.

[4] R.H. Inman, H.T. Pedro, C.F. Coimbra, Solar forecasting methods for renewable energy integration, Prog. Energy Combust. Sci. 39 (6) (2013) 535–576, http://dx.doi.org/10.1016/j.pecs.2013.06.002, URL https://www.sciencedirect.com/science/article/pii/S0360128513000294.

[5] N.B. Blum, S. Wilbert, B. Nouri, J. Stührenberg, J.E. Lezaca Galeano, T. Schmidt, D. Heinemann, T. Vogt, A. Kazantzidis, R. Pitz-Paal, Analyzing spatial variations of cloud attenuation by a network of all-sky imagers, Remote. Sens. 14 (22) (2022) http://dx.doi.org/10.3390/rs14225685, URL https://www.mdpi.com/2072-4292/14/22/5685.

[6] P. Kuhn, B. Nouri, S. Wilbert, C. Prahl, N. Kozonek, T. Schmidt, Z. Yasser, L. Santigosa, L. Zarzalejo, A. Meyer, L. Vuilleumier, D. Heinemann, P. Blanc, R. Pitz-Paal, Validation of an all-sky imager-based nowcasting system for industrial PV plants, Prog. Photovolt., Res. Appl. 26 (2017) http://dx.doi.org/10.1002/pip.2968.

[7] H. Wen, Y. Du, X. Chen, E. Lim, H. Wen, L. Jiang, W. Xiang, Deep learning based multistep solar forecasting for PV ramp-rate control using sky images, IEEE Trans. Ind. Informatics 17 (2) (2021) 1397–1406, http://dx.doi.org/10.1109/TII.2020.2987916.

[8] J. Schaible, B. Nouri, L. Höpken, T. Kotzab, M. Loevenich, N. Blum, A. Hammer, J. Stührenberg, K. Jäger, C. Becker, S. Wilbert, Application of nowcasting to reduce the impact of irradiance ramps on PV power plants, EPJ Photovolt. 15 (2024) 15, http://dx.doi.org/10.1051/epjpv/2024009.

[9] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F.J. de Pison, F. Antonanzas-Torres, Review of photovoltaic power forecasting, Sol. Energy 136 (2016) 78–111.

[10] J. Marcos, L. Marroyo, E. Lorenzo, D. Alvira, E. Izco, Storage requirements for PV power ramp-rate control, Sol. Energy 86 (10) (2011) 2677–2684.

[11] D. Yang, W. Wang, C.A. Gueymard, T. Hong, J. Kleissl, J. Huang, M.J. Perez, R. Perez, J.M. Bright, X. Xia, D. van der Meer, I.M. Peters, A review of solar forecasting, its dependence on atmospheric sciences and implications for grid integration: Towards carbon neutrality, Renew. Sustain. Energy Rev. 161 (2022) 112348, http://dx.doi.org/10.1016/j.rser.2022.112348, URL https://www.sciencedirect.com/science/article/pii/S1364032122002593.

[12] Z. Peng, D. Yu, D. Huang, J. Heiser, S. Yoo, P. Kalb, 3D cloud detection and tracking system for solar forecast using multiple sky imagers, Sol. Energy 118 (2015) 496–519, http://dx.doi.org/10.1016/j.solener.2015.05.037, URL https://www.sciencedirect.com/science/article/pii/S0038092X15002972.

[13] B. Nouri, P. Kuhn, S. Wilbert, N. Hanrieder, C. Prahl, L. Zarzalejo, A. Kazantzidis, P. Blanc, R. Pitz-Paal, Cloud height and tracking accuracy of three all sky imager systems for individual clouds, Sol. Energy 177 (2019) 213–228, http://dx.doi.org/10.1016/j.solener.2018.10.079, URL https://www.sciencedirect.com/science/article/pii/S0038092X18310570.

[14] B. Nouri, S. Wilbert, N. Blum, Y. Fabel, E. Lorenz, A. Hammer, T. Schmidt, L.F. Zarzalejo, R. Pitz-Paal, Probabilistic solar nowcasting based on all-sky imagers, Sol. Energy 253 (2023) 285–307, http://dx.doi.org/10.1016/j.solener.2023.01.060, URL https://www.sciencedirect.com/science/article/pii/S0038092X23000683.

[15] B. Nouri, S. Wilbert, L. Segura, P. Kuhn, N. Hanrieder, A. Kazantzidis, T. Schmidt, L. Zarzalejo, P. Blanc, R. Pitz-Paal, Determination of cloud transmittance for all sky imager based solar nowcasting, Sol. Energy 181 (2019) 251–263, http://dx.doi.org/10.1016/j.solener.2019.02.004, URL https://www.sciencedirect.com/science/article/pii/S0038092X19301306.

[16] C.-L. Fu, H.-Y. Cheng, Predicting solar irradiance with all-sky image features via regression, Sol. Energy 97 (2013) 537–550, http://dx.doi.org/10.1016/j.solener.2013.09.016, URL https://www.sciencedirect.com/science/article/pii/S0038092X13003770.

[17] H.-Y. Cheng, C.-C. Yu, Multi-model solar irradiance prediction based on automatic cloud classification, Energy 91 (2015) 579–587, http://dx.doi.org/10.1016/j.energy.2015.08.075, URL https://www.sciencedirect.com/science/article/pii/S0360544215011512.

[18] C. Long, J. Sabburg, J. Calbó, D. Pages, Retrieving cloud characteristics from ground-based daytime color all-sky images, J. Atmos. Ocean. Technol. J ATMOS OCEAN TECHNOL 23 (2006) http://dx.doi.org/10.1175/JTECH1875.1.

[19] A. Heinle, A. Macke, A. Srivastav, Automatic cloud classification of whole sky images, Atmos. Meas. Tech. 3 (3) (2010) 557–567, http://dx.doi.org/10.5194/amt-3-557-2010, URL https://amt.copernicus.org/articles/3/557/2010/.

[20] Q. Li, W. Lyu, J. Yang, A hybrid thresholding algorithm for cloud detection on ground-based color images, J. Atmos. Ocean. Technol. 28 (2011) 1286–1296, http://dx.doi.org/10.1175/JTECH-D-11-00009.1.

[21] A. Kazantzidis, P. Tzoumanikas, A. Bais, S. Fotopoulos, G. Economou, Cloud detection and classification with the use of whole-sky ground-based images, Atmos. Res. 113 (2012) 80–88, http://dx.doi.org/10.1016/j.atmosres.2012.05.005, URL https://www.sciencedirect.com/science/article/pii/S0169809512001342.

[22] V. Jayadevan, J. Rodriguez, A. Cronin, A new contrast-enhancing feature for cloud detection in ground-based sky images, J. Atmos. Ocean. Technol. 32 (2015) 209–219, http://dx.doi.org/10.1175/JTECH-D-14-00053.1.

[23] A. Taravat, F. Del Frate, C. Cornaro, S. Vergari, Neural networks and support vector machine algorithms for automatic cloud classification of whole-sky ground-based images, IEEE Geosci. Remote. Sens. Lett. 12 (3) (2015) 666–670, http://dx.doi.org/10.1109/LGRS.2014.2356616.

[24] W. Xie, D. Liu, M. Yang, S. Chen, B. Wang, Z. Wang, Y. Xia, Y. Liu, Y. Wang, C. Zhang, SegCloud: a novel cloud image segmentation model using a deep convolutional neural network for ground-based all-sky-view camera observation, Atmospheric Meas. Tech. 13 (4) (2020) 1953–1961, http://dx.doi.org/10.5194/amt-13-1953-2020, URL https://amt.copernicus.org/articles/13/1953/2020/.

[25] M. Hasenbalg, P. Kuhn, S. Wilbert, B. Nouri, A. Kazantzidis, Benchmarking of six cloud segmentation algorithms for ground-based all-sky imagers, Sol. Energy 201 (2020) 596–614, http://dx.doi.org/10.1016/j.solener.2020.02.042, URL https://www.sciencedirect.com/science/article/pii/S0038092X2030147X.

[26] W.M. Organization, International cloud atlas, 2017, URL https://cloudatlas.wmo.int/en/clouds-definitions.html. (Accessed 18 December 2024).

[27] L. Ye, Z. Cao, Y. Xiao, Z. Yang, Supervised fine-grained cloud detection and recognition in whole-sky images, IEEE Trans. Geosci. Remote Sens. 57 (10) (2019) 7972–7985, http://dx.doi.org/10.1109/TGRS.2019.2917612.

[28] Y. Fabel, B. Nouri, S. Wilbert, N. Blum, R. Triebel, M. Hasenbalg, P. Kuhn, L.F. Zarzalejo, R. Pitz-Paal, Applying self-supervised learning for semantic cloud segmentation of all-sky images, Atmospheric Meas. Tech. 15 (3) (2022) 797–809, http://dx.doi.org/10.5194/amt-15-797-2022, URL https://amt.copernicus.org/articles/15/797/2022/.

[29] M. Caron, P. Bojanowski, A. Joulin, M. Douze, Deep clustering for unsupervised learning of visual features, 2019, arXiv:1807.05520.

[30] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, 2018, arXiv:1703.01780.

[31] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E.D. Cubuk, A. Kurakin, H. Zhang, C. Raffel, FixMatch: Simplifying semi-supervised learning with consistency and confidence, 2020, arXiv:2001.07685.

[32] L. Yang, L. Qi, L. Feng, W. Zhang, Y. Shi, Revisiting weak-to-strong consistency in semi-supervised semantic segmentation, 2023, arXiv:2208.09910.

[33] L. Ye, Y. Wang, Z. Cao, Z. Yang, H. Min, A self training mechanism with scanty and incompletely annotated samples for learning-based cloud detection in whole sky images, Earth Space Sci. 9 (6) (2022) http://dx.doi.org/10.1029/2022EA002220, e2022EA002220. e2022EA002220 2022EA002220.

[34] J. Huertas-Tato, F.J. Rodríguez-Benítez, C. Arbizu-Barrena, R. Aler-Mur, I. Galvan-Leon, D. Pozo-Vázquez, Automatic cloud-type classification based on the combined use of a sky camera and a ceilometer, J. Geophys. Res.: Atmos. 122 (20) (2017) 11,045–11,061, http://dx.doi.org/10.1002/2017JD027131.

[35] L. Lu, M. Yin, L. Fu, F. Yang, Uncertainty-aware pseudo-label and consistency for semi-supervised medical image segmentation, Biomed. Signal Process. Control. 79 (2023) 104203, URL https://api.semanticscholar.org/CorpusID:252474564.

[36] G. French, S. Laine, T. Aila, M. Mackiewicz, G. Finlayson, Semi-supervised semantic segmentation needs strong, varied perturbations, 2020, arXiv:1906.01916.

[37] S. Yun, D. Han, S.J. Oh, S. Chun, J. Choe, Y. Yoo, CutMix: Regularization strategy to train strong classifiers with localizable features, 2019, arXiv:1905.04899.
[38] W. Falcon, The PyTorch Lightning team, PyTorch Lightning, 2019, http://dx.doi.org/10.5281/zenodo.3828935, URL https://github.com/Lightning-AI/lightning.
[39] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library, 2019, arXiv:1912.01703.
[40] L.N. Smith, Cyclical learning rates for training neural networks, 2017, arXiv:1506.01186.
[41] L.N. Smith, Super-convergence: Very fast training of neural networks using large learning rates, 2018, arXiv preprint arXiv:1708.07120.
[42] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2017, arXiv preprint arXiv:1711.05101.
[43] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018, arXiv:1802.02611.
[44] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015, arXiv:1512.03385.
[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255, http://dx.doi.org/10.1109/CVPR.2009.5206848.
[46] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, K. He, Accurate, large minibatch SGD: Training ImageNet in 1 hour, 2017, arXiv preprint arXiv:1706.02677.
[47] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, 2015, arXiv:1505.04597.
[48] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.