The presented work was submitted to the

Institute of Integrated Photonics

Semi-supervised Learning for Probabilistic Cloud Detection in Ground-based Imagery

Master's Thesis in Electrical Engineering

written at

RWTH Aachen University

by

David Magiera

submitted on

03.06.2024

Examiner

Univ.-Prof. Jeremy Witzens, PhD

Supervisor

Yann Fabel, M.Sc.

Acknowledgements

This thesis was conducted at the Institute of Solar Research of the German Aerospace Center (Deutsches Zentrum für Luft- und Raumfahrt e.V., DLR) within the department Qualification. The work was carried out at the DLR office in Almería, Spain.

First of all, I would like to thank my supervisor Yann Fabel for giving me the opportunity for conducting my Master's thesis at DLR. His valuable guidance, support and patience were essential for the completion of this thesis. The experience at DLR has been truly transformative.

I would also like to thank Prof. Dr. Witzens for his interest in my research and for giving me the opportunity conduct an external Master's thesis at DLR. I would also like to thank Dorothée Pawelzick for her support in administrational matters.

Furthermore, I would like to thank Alexander Schnerring for sharing his comfortable home in Almería and giving me a place to stay whenever I wanted to go to the office.

My sincere gratitude goes to my family, who have supported me through all these years. They have always trusted in my abilities without putting pressure on me. Without their support the completion of my studies would not have been possible. Finally, I would like to thank Camila for her endless patience, love and emotional support at times when I could not even imagine completing my Master's degree.

Abstract

One of the most significant challenges of our era is the transition to clean and renewable energy sources. The most abundant energy source on Earth is the sun, yet it is not consistently available due to the variability in solar irradiance. The most significant impact on local fluctuations in solar irradiance is the presence of clouds. Nowcasting systems provide intra-hour forecasts to anticipate this short-term variability. These forecasts are beneficial for grid control, the reduction of ramps in large-scale photovoltaic parks, and the operation of concentrating solar power (CSP) plants.

A common approach to nowcasting is to use ground-based sky images from All-Sky Imagers to detect clouds, which are then further tracked in a physical nowcasting system to estimate the future irradiance. The quality of these nowcasts depends highly on the quality of the cloud detection, which is commonly performed on the pixel level. Recent years have demonstrated that deep learning-based methods outperform all existing conventional approaches on these kinds of tasks. The challenge of deep learning-based methods is that they require a significant amount of high-quality, human-annotated all-sky images with annotations at the pixel level, which are in practice costly to obtain. Furthermore, the uncertainty in the predictions of systems based on artificial intelligence is often difficult to predict.

This thesis proposes an approach based on semi-supervised learning that fuses measurements from a ceilometer, which is a LiDAR sensor used to measure the cloud base height, into the learning process of a camera-based cloud detection model to improve the detection of three different cloud layers. A dataset of all-sky images with over 47000 weakly annotated images is created based on heuristics. This dataset is employed in conjunction with 770 human annotated all-sky images to train a cloud detection model, utilising semi-supervised learning techniques such as pseudo-labeling and consistency regularisation. Furthermore, a probabilistic calibration method is applied as a post-processing step to calibrate the uncertainty estimates of the developed cloud segmentation model for predictions on unseen all-sky images.

Evaluation on a benchmark of 36 all-sky images from three cameras shows the effectiveness of the two methods. The semi-supervised learning method demonstrated superior accuracy and IoU, respectively, by 2.4% and 3.5% compared to the current state-of-the-art method for semantic cloud segmentation. Calibration error metrics, such as ECE and MCE, are reduced significantly from 0.176 to 0.037 and from 0.276 to 0.111 through the probabilistic calibration, indicating a notable improvement in uncertainty estimation.

Contents

Acronyms

ASI All Sky Imager

ASPP Atrous Spatial Pyramid Pooling

BNN Bayesian Neural Network

CBH cloud base height

CIEMAT Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

CMOS Complementary Metal-Oxide-Semiconductor

CNN Convolutional Neural Network

DNI Direct Normal Irradiance

MS COCO Microsoft Common Objects in Context

CSP Concentrating Solar Power

DDPM Denoising Diffusion Probabilistic Model

ECE Expected Calibration Error

FCN Fully Convolutional Network

HDR High Dynamic Range

HSI hue, saturation, intensity

IoU Intersection over Union

LiDAR Light Detection and Ranging

LTS Local Temperature Scaling

MCE Maximimum Calibration Error

PSA Plataforma Solar de Almería

PV Photovoltaic

RGB red, green, blue

SOTA state-of-the-art

SVM Support Vector Machine

ToF Time of Flight

WMO World Meteorological Organization

1 Introduction

1.1 Motivation

Detecting clouds in ground-based imagery is important for many applications, including meteorology and climatology [1] and supporting optical satellite downlink operations to optical ground stations [2].

Another increasingly important application of cloud detection is related to solar energy, one of the main drivers for a sustainable and clean future in terms of energy production. One of the challenges for solar energy is 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 [3], which are difficult to predict.

Especially large Photovoltaic (PV) parks can benefit from reliable solar irradiance predictions, for instance for the ramp rate control [4], and can be optimized in terms of efficiency [5] and grid stability [6]. PV, another technology of solar energy, is also effected by the variability of solar irradiance. Concentrating Solar Power (CSP) consist of large mirror structures that focus solar radiation onto a receiver to heat transfer fluids. Spatial variations of solar irradiance on the solar field are challenging, because they can impose damaging thermal stresses on the materials that must be minimized while maximizing energy yield using complex control strategies [7].

So-called nowcasting, based on ground-based observations using all-sky imagers, provides solar irradiance predictions with high spatial- and temporal-resolutions with prediction horizons up to 20 minutes. Highly resolved nowcasts with coverage of several km^2 can be reached using using multiple all-sky imagers with stereographic approaches [8]. Coverage of several thousand km^2 can be reached with large scale all-sky imager networks [3]. Physical nowcasting models use a chain of processing steps beginning from cloud detection, cloud tracking and cloud geolocation [9]. As the cloud detection is the first processing step, undetected errors in the cloud detection will be propagated through the whole chain into the solar irradiance prediction and possibly even into the decision making in the target application. Thus cloud detection needs to be accurate in its predictions and reliable in its uncertainty estimates.

1.2 Objectives and Challenges

This thesis aims to further develop deep learning-based cloud detection in ground-based imagery using all-sky imagers. The World Meteorological Organization (WMO) categorizes clouds into the following three layers: *Low-layer*, *mid-layer* and *high-layer* [10]. These categories are used for cloud detection in this work. Specifically, each pixel within a sky-image is to be classified into either one of the three cloud layers or to be classified as *clear-sky*. In computer vision terms this task is defined as semantic segmentation. Thus cloud detection will be further referred to as *semantic cloud segmentation* throughout this thesis.

The categorization of clouds into three layers is a strong simplification considering ten different cloud generas, with several species each, as defined by the WMO [10] and illustrated in figure 1.1. From a solar irradiance prediction standpoint this categorization can still be beneficial. *Low-layer* clouds are mostly dense water clouds with strong dampening effects on the solar irradiance. *High-layer* clouds on the other hand consist out of ice particles only, appear often hazy, and have a higher transmittance. Clouds in the *mid-layer* are a mixture of both, resulting usually in a transmittance in-between the transmittance of clouds in the other two layers [11]. Also, different cloud layers move often into different directions due to different wind directions in the atmosphere. Thus this categorization into different cloud layers has the potential to improve the cloud tracking in physical nowcasting systems.

Even for human experts, distinguishing these three cloud layers can be challenging for various reasons: Complex multi-layer conditions with varying lighting conditions, and the lack of sharp boundaries and effects as oversaturation make semantic cloud segmentation particularly difficult. Also, aerosols and atmospheric turbidity are easily confused with thin clouds in the *high-layer*. Thus apart from the semantic cloud segmentation a pixel-wise confidence estimation is required, providing reliable uncertainty estimates for each prediction. The development of a method to provide these pixel-wise confidence estimates is the first main objective of this thesis and has not been studied for multi class semantic cloud segmentation in the literature so far.

The second challenge for deep learning-based semantic cloud segmentation is the need for a significant amount of annotated images at the pixel level that represent the ground truth, as discussed in [12]. The process of annotating all-sky images at the pixel level is difficult due to ambiguities as previously mentioned, which makes it extremely time consuming, and usually not feasible for large volumes due to economic and time constraints, particularly in research. A previous study [13] demonstrated that unlabeled all-sky images can be leveraged with self-supervised learning to reduce the need for human-annotated images and that effective training of semantic cloud segmentation models with limited human annotation is possible. In the recent years impressive successes have been

demonstrated on a variety of computer vision tasks, by training deep learning models with a limited number of human-annotated samples and a large amount of unlabeled and weakly annotated samples using semi-supervised learning techniques [14][15][16]. Hence, the second primary objective of this thesis is the development of a semi-supervised learning method for the purpose of leveraging weakly annotated all-sky images based on additional sensor measurements for semantic cloud segmentation.

The remainder of the thesis is organized as follows. In chapter 2, the state-of-the-art techniques in terms of ground-based cloud detection as well as for semi-supervised learning, weak supervision, and uncertainty estimation in neural networks are presented. In chapter 3, the utilized sensors and datasets are presented. The developed methods are presented in chapter 4. The experimental results are presented and discussed in detail in chapter 5, while the conclusion and an outlook on future research are provided in chapter 6.

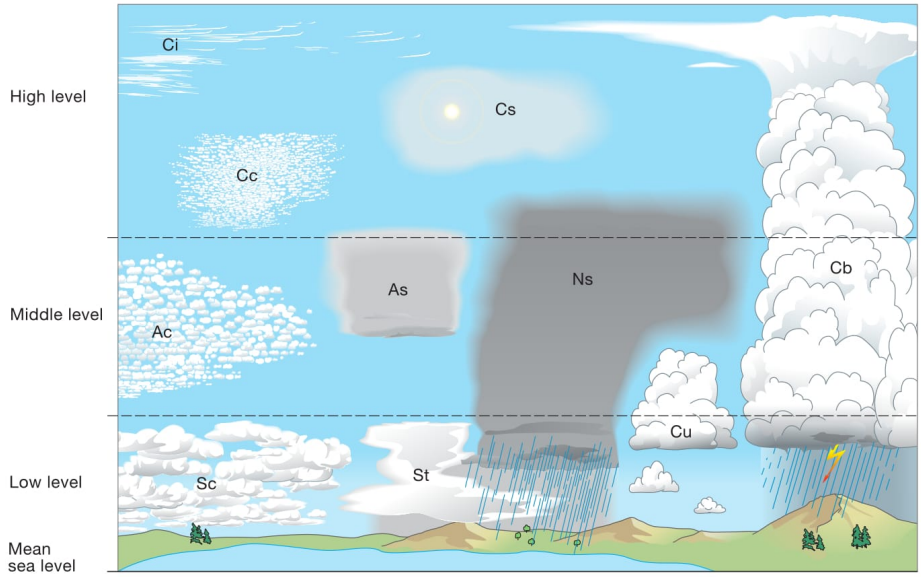


Figure 1.1: Illustration of the main cloud generas defined by the WMO [10]. The ten cloud generas: Cumulus (Cu), Stratus (St), Stratocumulus (Sc), Cumulonimbus (Cb), Altocumulus (Ac), Altostratus (As), Nimbostratus (Ns), Cirrus (Ci), Cirrocumulus (CC), Cirrostratus (Cs) are grouped by their base height into low-layer, mid-layer and high-layer.

2 Related work

The chapter first presents various methodologies for cloud detection in ground-based imagery. This is followed by an analysis of the current state-of-the-art techniques in computer vision for deep learning-based semantic segmentation, semi-supervised learning and uncertainty estimation.

2.1 Cloud detection from ground-based imagery

Cloud detection in ground-based imagery has been a topic of active research for the past two decades. Historically, thresholding-based methods were initial attempts to distinguish between clear-sky and cloud pixels, leveraging the observation that during daytime, the sky appears blue and clouds appear white. Methods based on thresholding of the red-blue ratio of the color channels with fixed thresholds [17], and hybrid thresholds [18] and based on the difference of the red and blue channels [19] have been proposed. Other methods included also the green color channel [20], transforming the image into other color spaces such as as hue, saturation, intensity (HSI) and deciding based on saturation [21]. 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, green, blue (RGB) ratio.

Over the past decade, deep learning-based methods have been applied to cloud detection. Initially, shallow fully-connected architectures were utilized [22]. This was followed by SegCloud [23], which was the first reported success in using a deep Convolutional Neural Network (CNN) for cloud detection. This method showed superior performance compared to the existing thresholding methods. All of these methods have in common to differentiate between sky and cloud pixels, but not between different cloud layers or cloud generas as defined by the WMO. The first approach to differentiate between the different cloud generas was proposed by [24] and the first approach to differentiate between the three cloud-layers was proposed by [13].

A deep learning-based method for probabilistic cloud detection differentiating between clear-sky and cloud pixels was recently proposed by [25]. However, the literature lacks

a comprehensive approach to probabilistic cloud detection for multiple cloud generas or cloud layers in ground-based imagery.

2.2 Deep learning for semantic segmentation

Semantic segmentation is a computer vision task that assigns class labels to pixels in images. It is a highly active research topic due to its relevance to a vast array of target domains, including biomedical image segmentation, automated driving and robotics.

The foundation for deep learning-based semantic segmentation was established by [12] with the introduction of Fully Convolutional Networks (FCNs), which replaced fully-connected layers with convolutional layers. This allowed for end-to-end training for pixel-wise predictions. In the context of deep learning, convolutions are small matrices, often of shape 3x3, which move with a predefined stride over the image and act as filters to detect shapes such as edges, textures, and patterns [26]. The introduction of residual skip connections by [27] made the effective training of deep architectures possible, as it addressed the vanishing gradient problem in deep neural networks and led to the invention of the widely used ResNet architecture.

The next significant advancement in deep learning-based semantic segmentation was the U-Net proposed by [28]. This involved the introduction of an encoder-decoder architecture with symmetric skip connections that link corresponding layers in the encoder and decoder parts. This U-shaped architecture allows for the efficient abstraction of context information into high-dimensional feature vectors through the encoder and precise localization of these features back into the image dimensions through the decoder. This approach permitted the implementation of efficient high-resolution semantic segmentation.

One of the current state-of-the-art (SOTA) architectures for semantic segmentation is DeepLabv3+, proposed by [29], which incorporates several innovative techniques for semantic segmentation, such as an encoder-decoder architecture, as illustrated in figure 2.1. Atrous convolutions, also known as dilated convolutions, permit the creation of larger receptive fields without the necessity to introduce more parameters or compromising the image resolution. Atrous Spatial Pyramid Pooling (ASPP) are modules that apply convolutions with varying dilation rates in parallel, capturing image context at multiple scales. ASPP enhances the ability of the network to extract both local and global contextual information, which is highly beneficial for tasks such as semantic cloud segmentation.

One of the most significant challenges for deep learning-based semantic segmentation is the necessity for substantial volume of human-annotated ground truth images. The

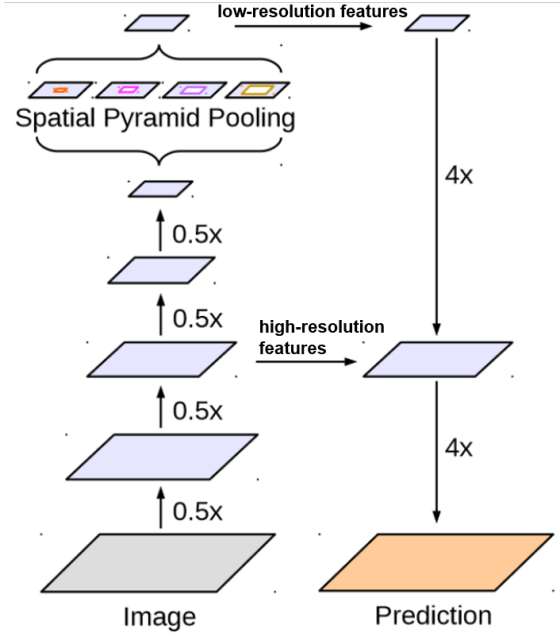


Figure 2.1: Illustration of the DeepLabv3+ encoder-decoder architecture. Credits: [29]

process of annotating images for semantic segmentation is particularly labor-intensive due to the requirement of pixel-level annotation for end-to-end training. Training on a limited number of samples often results in the model memorizing the training data rather than learning abstract, generalizable features, which leads to poor generalization capabilities on unseen data during inference. This phenomenon is also described as overfitting [30]. Several strategies have been developed to mitigate the need for large volumes of human-annotated images and still train deep neural networks without overfitting. Two of these methods are semi-supervised learning and weakly-supervised learning, which will be discussed in detail in the following section.

2.3 Semi-supervised learning

This section provides a definition of semi-supervised learning, and then presents a brief overview of the four assumptions that all semi-supervised techniques are based upon. It then goes on to discuss two specific techniques, pseudo-labeling and consistency regularization, which are used by the semi-supervised learning approach developed in this thesis. Finally, the chapter is concluded with an overview of weakly supervised learning, which is the second pillar of the proposed semi-supervised learning method.

2.4 Definition of semi-supervised learning

In machine learning, four distinct approaches exist for learning from data: supervised learning, unsupervised learning, self-supervised learning, and semi-supervised learning.

In supervised learning, the model is trained on a labeled dataset, which means that each training sample is paired with an output label. The goal of supervised learning is to learn a mapping from inputs to outputs that can be applied to new, unseen data as described in [31]. When the labels are continuous, the task is called regression. When the labels are discrete, the task is called classification.

In contrast, unsupervised learning functions without any labels. The goal is to find hidden patterns or intrinsic structures in the input data. Common applications include clustering, dimensionality reduction, and anomaly detection [32]. A special case of unsupervised learning is self-supervised learning. In self-supervised learning, the data itself provides the supervisory signal by generating artificial labels during so called pretext tasks. This is often achieved by predicting portions of the input from other portions or by creating target labels through clustering [33].

Semi-supervised learning is a hybrid form of supervised- and unsupervised learning, typically utilizing a small amount of labeled data and a large amount of unlabeled data. The objective is to leverage the vast amount of unlabeled data together with limited amount of labeled data to learn more accurate decision boundaries [34]. The necessary condition for semi-supervised learning to work is that the unlabeled samples utilized must be relevant to the task the model is trained to perform. Formally, this can be expressed as follows: the distribution of the input data, denoted as $p(x)$, must contain information about the posterior distribution $p(y|x)$, where y is the label to be predicted [35]. This condition gives rise to four assumptions that all semi-supervised learning methods are based on.

2.4.1 The four assumptions of semi-supervised learning

The four assumptions underlying semi-supervised learning, as defined in [35], are as follows: *cluster assumption*, *smoothness assumption*, *low-density assumption*, *manifold assumption*, which are defined in [35] as follows.

Cluster assumption

The *cluster assumption* states that data points belonging to the same cluster, i.e. data points that are more similar to each other are likely to share the same label. This can be visually illustrated by stating that data points that are close to each other in figure 2.2, should belong to the same class.

Smoothness assumption

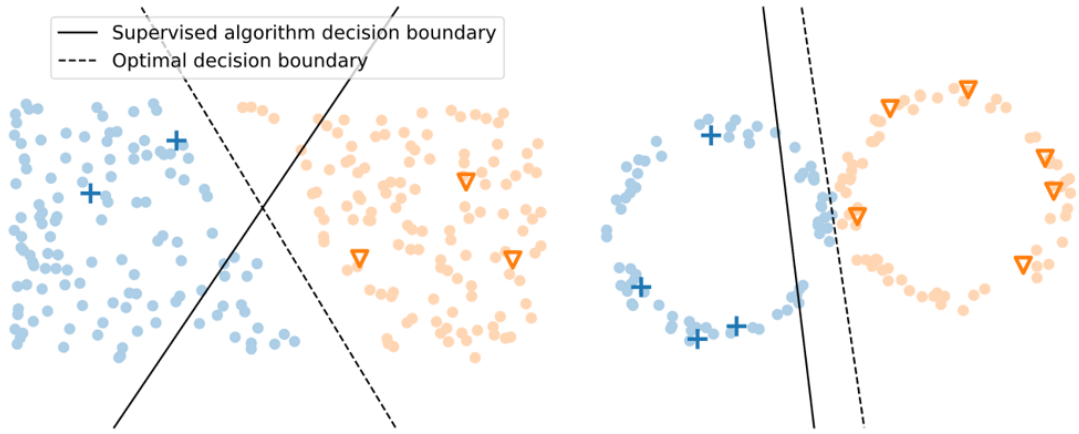
The *smoothness assumption*, also known as the *continuity assumption*, states that when two data points are proximate in the input space, they should be also be proximate in the output space.

Low-density assumption

The *low-density assumption* states that the decision boundary of a classifier should traverse low-density regions of input data distribution. As illustrated in figure 2.2 (a), semi-supervised learning that complies with the low-density and smoothness assumptions can inform a superior decision boundary than supervised learning with a limited number of labeled data points.

Manifold assumption

The *manifold assumption* states that the high-dimensional input data of the same class lie together on a lower-dimensional manifold than the input space. This implies that the high-dimensional input data can be projected into a lower-dimensional space, where it is more easily separable. Semi-supervised algorithms are designed to learn the projection into the lower-dimensional space, which is also known as *dimensionality reduction*. This involves mapping the input data into relevant feature representations, as illustrated in figure 2.2 (b).



(a) Smoothness and low-density assumptions. **(b)** Manifold assumption.

Figure 2.2: Illustration of the smoothness, low-density and manifold assumptions. The blue crosses and the orange triangles represent labeled data, while the points resemble unlabeled data. Credits: [34]

2.4.2 Semi-supervised learning techniques

Pseudo-labeling

Pseudo-labeling, also known as *self-training*, is a technique employed in semi-supervised learning to generate artificial labels on unlabeled data using the predictions as labels. Pseudo-labeling is built on the cluster assumption, which assumes that data points in the same cluster share the same label [36]. In most cases, the artificial labels are selected based on the confidence in the model's predictions, under the assumption that high confidence predictions are correct and can provide useful training signals [37]. The generated pseudo-labels are subsequently treated as if they were true labels, thereby enabling the unlabeled data to contribute to the training process. The fundamental idea is to leverage the model's predictions to enhance its learning capacity by expanding the labeled dataset. Self-training has been demonstrated to be an effective methodology for enhancing semantic segmentation [38].

Consistency regularization

Another semi-supervised learning technique is consistency regularization. This technique encourages a machine learning model to produce consistent outputs when the input data is perturbed in various ways. It leverages the smoothness assumption by assuming that various perturbations of the input data should have the same label. Perturbations in a semantic segmentation context can include transformations, also known as *data*

augmentation, such as geometrical transformations such as flipping, rotating, or scaling. Additionally, transformations in the color space and Gaussian blurring are frequently employed in the literature [39]. It is crucial to acknowledge that the data augmentations must to be selected in a manner that ensures the smoothness assumption remains valid under these transformations. In other word the perturbed images need to be realistic, as stated by [40], which is highly dependent on the specific domain.

A framework that combines both pseudo-labeling and consistency regularization with impressive results is the weak-to-strong consistency framework, proposed by [15]. Images are first weakly augmented using affine transformations such as flipping, rotating and cropping, and then fed through the model to generate pseudo-labels using confidence thresholds. In a second step, the weakly augmented images are then strongly augmented and fed again through the model. The objective of this process is to to match the prediction on the strongly-augmented images to the pseudo-labels on the weakly augmented images, as illustrated in figure 2.3. This framework could achieve impressive results with 88.61% accuracy with only 4 labeled samples per class on the CIFAR-10 [41] benchmark for image classification.

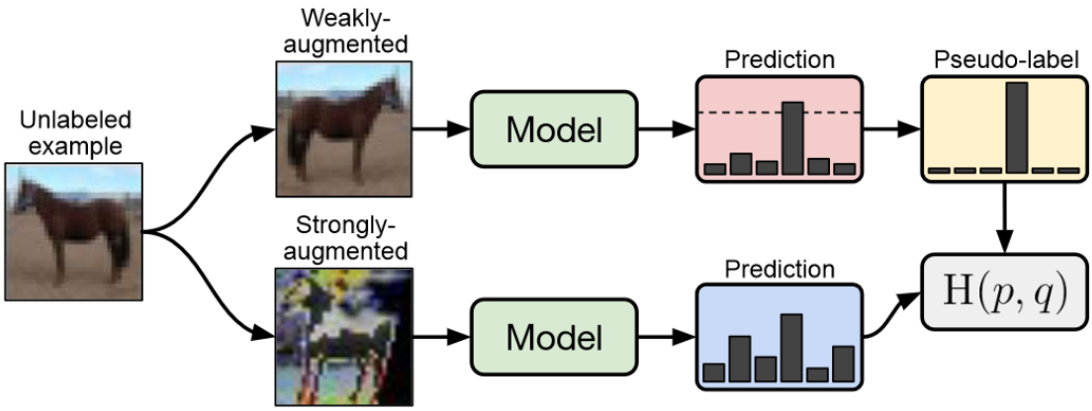


Figure 2.3: Illustration of the weak-to-strong consistency regularization framework popularized by FixMatch [15]. Credits: [15]

Pseudo-labeling-based methodologies are often challenging to implement in real-world problems due to the potential for inaccuracy in the pseudo-labels, which are further reinforced through the self-training process. Another challenge is confirmation bias as the confidence values for minority classes may be lower than those for majority classes. This can result in an even more pronounced class imbalance in the selected pseudo-labels due to confidence thresholding, as discussed in detail by [42].

2.4.3 Weakly supervised learning

Weakly supervised learning is an umbrella term for a variety of methodologies to learn under imperfect supervision. This imperfection can take several forms, such as incomplete, inaccurate, or inexact labels [43]. *Incomplete supervision* describes the situation when only a small subset of the data is labeled and the rest is unlabeled. This can be seen as a form of semi-supervised learning, as described above. *Inaccurate supervision* describes a situation in which the provided labels are noisy, i.e., they do not always accurately reflect the ground truth. This can occur due to errors made by the annotator or in ambiguous situations. *Inexact supervision* describes a situation, in which only coarse-grained supervision is provided. To illustrate, consider the context of semantic segmentation. In this domain, labels can be provided at the image level, rather than at the pixel level, which is often more feasible in a large volume. In practice, these three forms of supervision frequently occur simultaneously, as described in detail by [43]. One prominent example of weakly supervised learning in computer vision is CLIP, proposed by [44]. CLIP employs weak supervision on a large scale to learn robust image representations from 400 million image-text pairs.

2.5 Uncertainty estimation in neural networks

Real-world decision making systems, require classification networks not only to be accurate, but also to indicate when they are likely to be incorrect. This means that a network should provide a *calibrated confidence* measure in addition to its prediction, which should reflect the ground truth correctness likelihood [45]. A common problem with modern deep learning-based methods is that they tend to fail silently on unseen data. This implies, that despite predicting with high confidence, predictions can be catastrophically wrong. Such errors can have drastic consequences in domains, such as medical imaging [46] or autonomous driving [47], if undetected. This phenomenon is particularly pronounced when the input data distribution undergoes a shift, which is also referred to as *domain shift*. In the context of semantic cloud segmentation, domain shift can occur due to changes in camera hardware or a relocation of the observation site, which differs from the site on which the model was trained on. This motivates a variety of approaches to obtain calibrated confidences, which are discussed in the remainder of this section.

Bayesian Neural Networks (BNNs) [48] are a class of neural networks that incorporate Bayesian inference by directly modeling probability distributions. Exact Bayesian inference is computationally expensive and infeasible for large scale datasets and complex architectures. Consequentially, a number of methods have evolved to approximate Bayesian modeling, including Dropout [49], Monte Carlo Dropout [50], and Deep Ensembles [51]. Dropout is implemented in most architectures and the latter two require multiple for-

ward passes for each prediction during inference, increasing the computational cost as a consequence. Recently, Denoising Diffusion Probabilistic Model (DDPM), originally introduced for the purpose of image generation [52], have been demonstrated to be applicable for semantic segmentation for implicit ensembling to obtain confidence estimates [53]. Due to the numerous diffusion steps required for each prediction, training and inference for diffusion models is lengthy and computationally expensive. One solution to this problem is to decrease the image resolution, which is not ideal for semantic cloud segmentation, as it may result in the loss of important details.

The study about calibration in neural networks [45] introduced *temperature scaling* in the context of probabilistic model calibration. Temperature calibration is a simplified version of *platt scaling* [54], introduced in the context of Support Vector Machines (SVMs). Temperature calibration is performed as a postprocessing step after training to estimate a single parameter T to calibrate the outputs, of a previously trained network. During inference almost no additional computational costs arise, as it works with a single scalar-matrix multiplication. Temperature scaling is, as stated by [45], the simplest, fastest, and most straightforward calibration method, and surprisingly often the most effective.

The method of Local Temperature Scaling (LTS), which is based on temperature scaling, was proposed by [55] for the specific purpose of semantic segmentation tasks. It employs a small weight-calibration network that estimates a calibration map based on both the input and predictions. This calibration map is used to scale the confidence of the predictions for each pixel and is presented in detail in section 4.2.

3 Datasets of All-Sky Images

A reliable database of sufficient volume and diversity with respect to the conditions of the target domain is of fundamental importance for the development of reliable data-driven methods, such as deep learning-based semantic cloud segmentation models. All data utilized to implement the methods presented in chapter 4 and to obtain the experimental results presented in chapter 5 are presented in the following sections of this chapter. In total, two datasets were created as part of this thesis. Furthermore, an additional dataset of all-sky images that has was created previously is presented.

In the first section a brief overview of the observation site and the sensors used for the data acquisition is presented. The utilized and created datasets for this work are then presented in the following sections. First, the human-annotated training dataset is introduced. Then, a detailed description of the creation of a weakly labeled dataset using heuristics on the measurements of a sensor, apart from All Sky Imagers (ASIs), is provided. The final section of this chapter presents a benchmark dataset with images from multiple ASIs, that has been created as part of this thesis.

3.1 Observation site and utilized hardware

3.1.1 Plataforma Solar de Almería

All image data and meteorological measurements for the datasets presented in the following sections were acquired with the sensor infrastructure of the Plataforma Solar de Almería (PSA). The PSA is located in the desert of Tabernas in south-eastern Spain and is property of the Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), a public research center of the Spanish government. It is regarded as the world's largest and most comprehensive research and development facility dedicated to CSP [56].

The PSA has a network of meteorological measurement stations equipped with ASIs and other sensors distributed across the entire area, as illustrated in the aerial view of the PSA in figure 3.1. The all-sky images for the human-annotated dataset presented in section 3.2 were acquired at the meteorological measurement station *Kontas*. The all-sky

images and additional sensor data used to generate weak labels for the weakly annotated dataset presented in section 3.3 were acquired at the meteorological measurement station *Metas*. The all-sky images for the benchmark dataset presented in section 3.4 are from five ASIs from the meteorological measurement stations *Kontas*, *Metas* and *PVot*. The location of each meteorological test stand is marked in the aerial view of the PSA in figure 3.1.

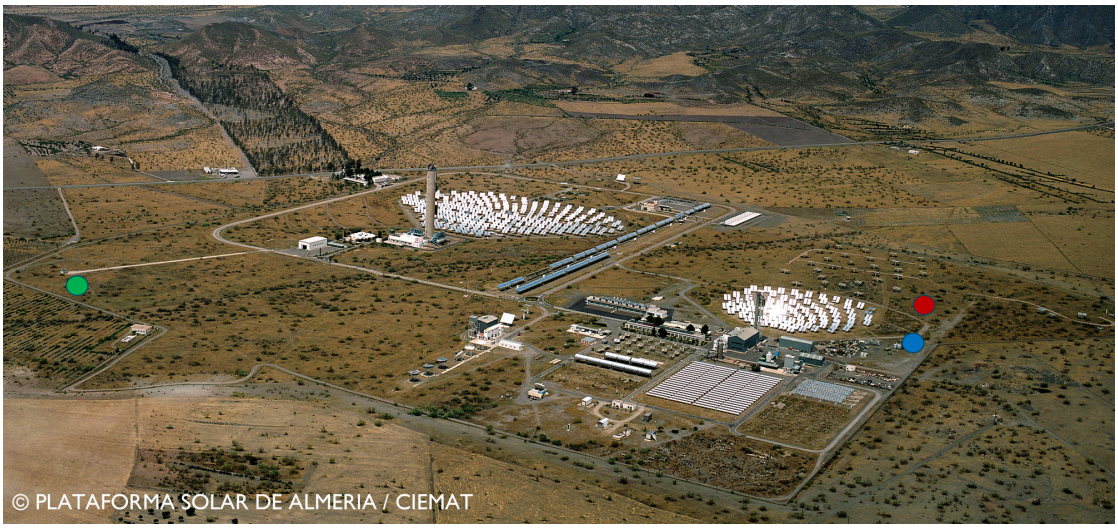


Figure 3.1: Aerial view of the PSA. The locations of the meteorological measurements stations *Kontas*, *Metas*, *PVot* are marked in red, green and blue respectively. Credits: CIEMAT

3.1.2 All-Sky Imager

An All Sky Imager ASI is a specialized camera system designed to capture wide-angle images of the entire sky [57]. These images are commonly used in atmospheric and astronomical studies, where they are employed to monitor cloud cover, measure sky luminance, observe auroras, and track other atmospheric phenomena.

In this thesis images from four different ASI models are utilized: Mobotix Q25, Mobotix Q26, Mobotix Q71, and AXIS. The images utilized for the human annotated training dataset and the weakly annotated dataset, presented in the following two sections, which were employed for the training and calibration of the cloud segmentation models in sections 5.2 and 5.3, respectively, were acquired with a Mobotix Q25. This device is shown in figure 3.2. The images utilized for the created benchmark dataset were acquired with all the ASIs. Further details about respective image characteristic are given in section 3.4, where the benchmark dataset is introduced. A Q71 model from Mobotix costs around 1300€ in 2024.

Figure 3.2: Picture of a Mobotix Q25 ASI installed at PSA.

3.1.3 Ceilometer

The ceilometer is a Light Detection and Ranging (LiDAR) sensor, which is an optical sensor used to measure the height of cloud layers above the ground. It works by emitting a laser beam vertically into the sky and then detecting the reflection of that beam off cloud particles or other airborne particles. The time it takes for the beam to return, also known as the Time of Flight (ToF), allows the ceilometer to calculate the distance to the cloud base, often denoted as the cloud base height (CBH). Due to its working principle, ceilometer provide a point measurement in the vertical direction. Ceilometers are commonly used in meteorology for weather forecasting.

The ceilometer installed at PSA is located at the *Metas* meteorological measurement station, which is marked in yellow in figure 3.1. The model is a CHM15k-Nimbus manufactured by Lufft. The new price, depending on the equipment of the device, is between 35.000€ and 40.000€ net in the year 2024. The utilized ceilometer installed at the PSA is shown in figure 3.3.

Figure 3.3: The ceilometer installed at the meteorological measurement station *Metas* at PSA.

3.2 Human-annotated training dataset

The human-annotated training dataset utilized for this thesis comprises 770 images labeled on at the pixel level. All images were captured by the ASI installed at the meteorological measurement station *Kontas*, marked in red in figure 3.1, in the year 2017. The dataset was initially created by [58] as a dataset for binary semantic cloud segmentation, differentiating between sky and cloud pixels. Eventually, the dataset was extended and the cloud labels were refined in the work of [13], who annotated each cloud pixel as low-layer, mid-layer, or high-layer cloud. This set the foundation for training deep learning-based semantic cloud segmentation models that could differentiate between these three cloud layers.

The dataset encompasses a diverse array of meteorological conditions, including diverse cloud conditions, varying sun elevation angles, and different atmospheric turbidity intensities. The respective data distributions are illustrated in figure 3.4. Despite significant human annotation effort, a dataset of 770 images is still small for deep learning-based semantic segmentation. For comparison the Microsoft Common Objects in Context (MS COCO) dataset [59], a large scale object detection, segmentation, key-point detection, and captioning dataset, contains over 164.000 on a pixel-level annotated samples. For further details regarding the the human-annotated training dataset, please refer to [13].

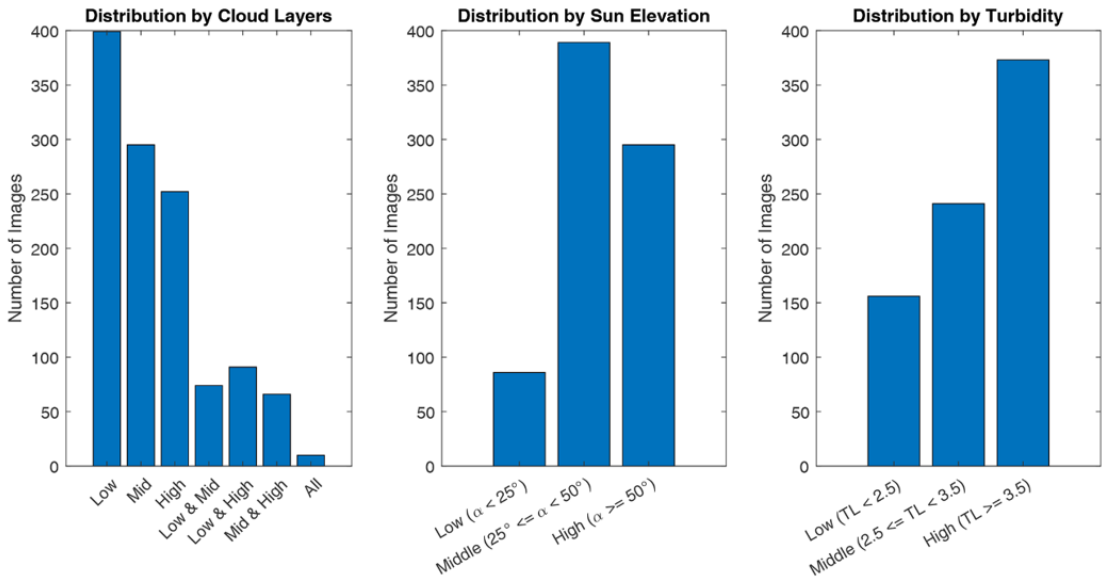


Figure 3.4: Data distributions of the human-annotated training dataset. Credits [13]

3.2.1 Training and validation split

The human-annotated dataset is divided in the same manner as in [13], with a fixed split of 80% training samples and 20% validation samples to ensure for a fair comparison between the SOTA cloud detection model from [13] and the models developed in this work. Consequently, 616 labeled training samples are employed to optimize the models for the benchmark in section 5.2. The remaining 154 samples are utilized to validate the training progress of the models and, additionally, to perform the probabilistic calibration presented in section 4.2, with the objective of ultimately obtaining a probabilistic cloud segmentation model.

3.3 Creation of a weakly labeled dataset

As previously stated, the human-annotated dataset comprising 770 samples is a relatively small dataset for training deep learning models with fully supervised learning on a challenging computer vision task, such as multi-layer semantic cloud segmentation.

As demonstrated by [13], a substantial unlabeled dataset comprising all-sky images can be employed for self-supervised learning with the objective of learning more effective feature representations. This ultimately leads to the generation of a more accurate model in terms of semantic cloud segmentation with a limited amount of human annotation. This motivates the approach of this work, to employ again unlabeled all-sky images to support the learning. This time the unlabeled data is not left completely unlabeled, as it undergoes weak labeling at the image level. This labeling is achieved through the application of heuristic to the measurements of a ceilometer.

3.3.1 Assigning image-level weak labels using heuristics

The assumption is that ceilometer measurements can be used to assign weak labels on an image-level to all-sky images. This is based on the following reasoning:

If the ceilometer detects clouds of only one of the three cloud layers for a long enough time, the ASIs close to the ceilometer will only capture clouds of that specific layer with a sufficiently high certainty. This allows the assignment of an image-level weak label of that specific detected cloud layer to the images.

The temporal extent for the heuristic to be valid is defined as a period of four hours in the past and four hours in the future. The applied thresholds for each cloud-layer are specified in table 3.1. These threshold were set with a margin compared to the definitions

of the WMO listed in table 3.2, with the objective of limiting the prevalence of ambiguous cloud conditions. The assigned weak labels are used for pseudo-labeling in the proposed semi-supervised learning approach, which is explained in detail in 4.3.

Cloud-layer	min. threshold [m]	max. threshold [m]
Low-layer	0	2000
Mid-layer	3000	6000
High-layer	8000	

Table 3.1: The thresholds for the heuristics applied to the ceilometer measurements to assign image-level weak labels to all-sky images.

Cloud-layer	min. threshold [m]	max. threshold [m]
Low-layer	0	2400
Mid-layer	1800	8000
High-layer	6000	

Table 3.2: Height levels for the three cloud layers defined by the WMO for mid-latitude regions, like Southern Spain. [10].

Furthermore, at least one cloud detection must occur within a temporal extent of ten minutes for the heuristic to be considered valid. This second rule is applied with the intention of reducing the number of images with minimal cloud coverage, as these images contribute little useful information to the training process.

Figure 3.5 illustrates the ceilometer measurements for a single day, accompanied by the ASI image keogram for the same day. The keogram is generated by slicing the images over a single day at a fixed horizontal position, followed by concatenation along the horizontal axis in chronological order. This provides a general overview of the clouds observed by the ASI for the day in question. The ceilometer measurements indicate that for the day, only clouds in the high-layer range passed the PSA with a high degree of certainty, as only clouds with a CBH in the high-layer range were detected.

3.3.2 Additional filters and masking by elevation angles

Additional filters are applied to the all-sky images to ensure the data quality is high in the weakly-annotated images. As the weakly-annotated images will be used for pseudo-labeling during semi-supervised learning, it is necessary to filter them in a way that minimizes the domain shift between human-annotated images and weakly-annotated images and that excludes too difficult conditions. In the creation of the human-annotated training dataset only images with a sun elevation angle above 10 degrees were considered [13]. The minimum sun elevation angle for the weakly annotated images is set to 20

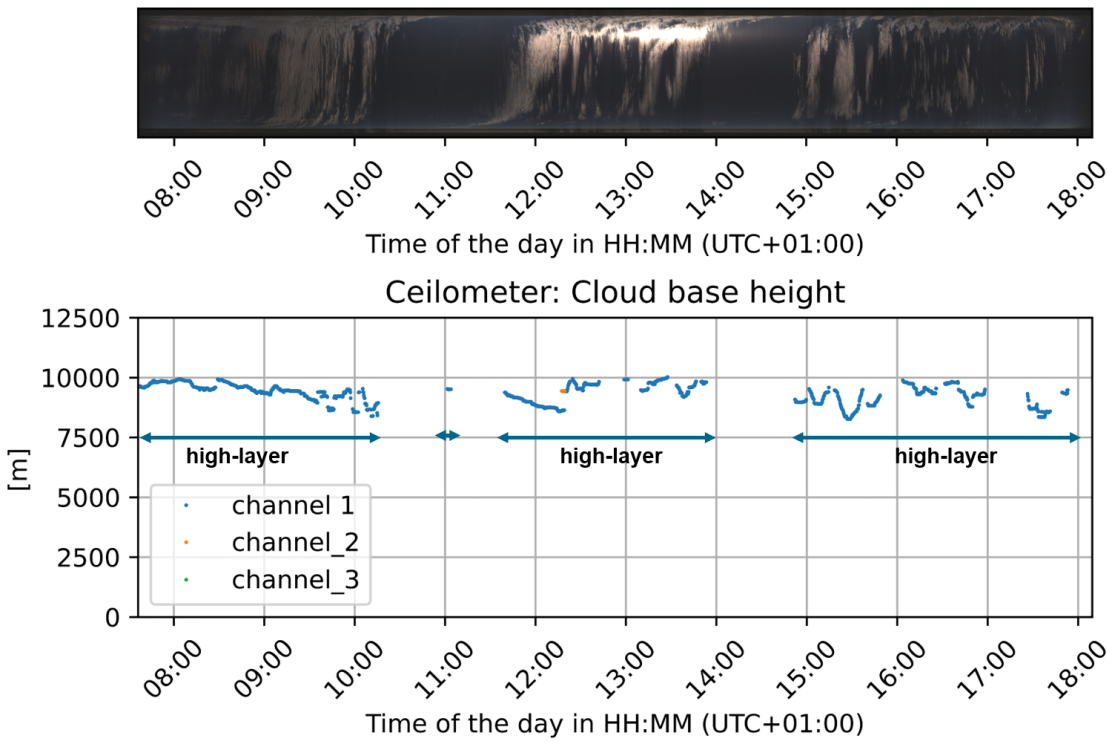


Figure 3.5: Ceilometer measurements for a day of example together with the ASI keogram for the same day.

degrees to provide a margin of safety. Furthermore, images recorded under high atmospheric turbidity values are excluded from consideration, as they could be confused with clouds. Consequently, only images with an atmospheric turbidity value below 4.0 will be considered during the weak-labeling process. The applied thresholds are listed in table 3.3.

Measurement	threshold
Min. sun elevation angle	20°
Max. atmospheric turbidity	4.0

Table 3.3: Thresholds for the selection of all-sky images for the weak-annotation process.

In addition to the presented filters, all parts of the images with elevation angles below 30 degrees are masked. This is because the ceilometer only provides measurements of the CBH vertically above the measurement station, as previously discussed. This 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. Figure 3.6 illustrates the weak-labeling approach by providing one example for each possible weak-label, where also the limited field of view can be observed.

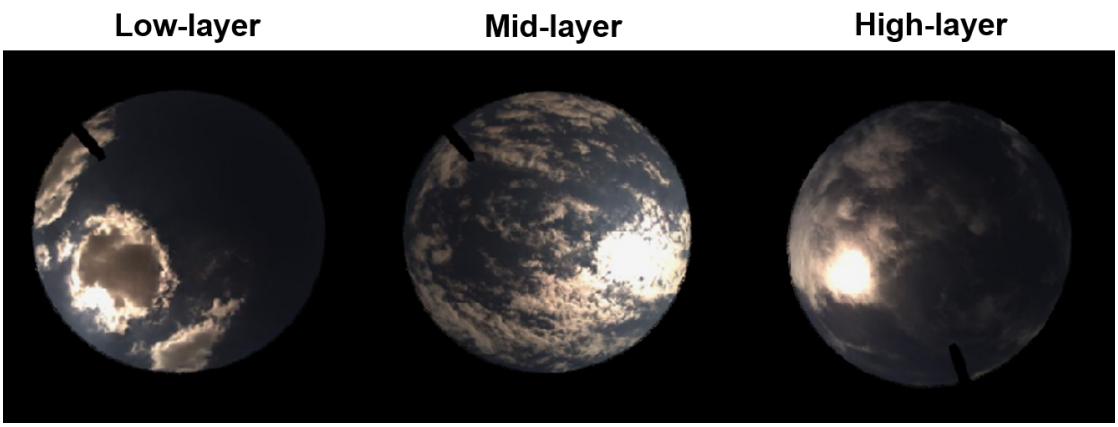


Figure 3.6: One example for each image-level weak label from the generated weakly-annotated dataset.

3.3.3 Data distributions of the generated dataset

The heuristics are applied to the ASI at the meteorological measurement station *Metas*, where the ceilometer is also installed. The images from July 2019 until the end of October 2021 are utilized. For the duration the camera model installed at *Metas* is the

same Mobotix Q25 model that was installed in the year 2017 at *Kontas*. The images from this ASI were utilized for the human-annotated dataset presented in the previous section. This should ensure a minimal domain shift between the images of the human-annotated dataset and the weakly-annotated images from this chapter and facilitate semi-supervised learning techniques, such as pseudo-labeling.

Applying the heuristics to the selected all-sky images for the specified time span of over two years of measurements yields 47595 unique weakly-annotated images. Of these, 21341 samples are weakly-annotated as low-layer, 19885 samples are weakly-annotated as high-layer, and 6396 samples are weakly-annotated as mid-layer, which constitutes the minority class in this case. To achieve balance in the dataset, the minority classes of high-layer and mid-layer are over-sampled to the quantity of low-layer, in order to obtain a balanced dataset. The distributions of the data in the generated weakly-annotated dataset, created using heuristics on ceilometer measurements, are shown in figure 3.7. As illustrated in the distributions for the sun elevation and atmospheric turbidity, the dataset encompasses a diverse range of atmospheric conditions.

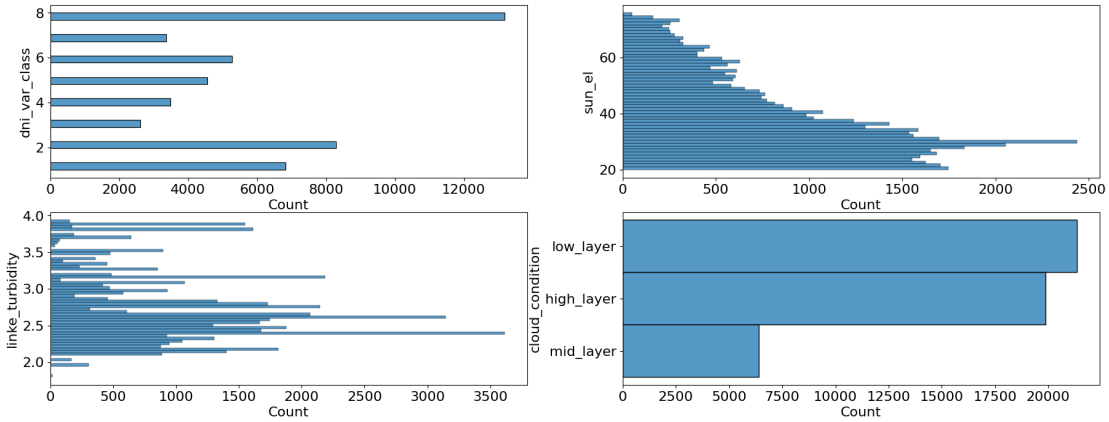


Figure 3.7: Data distribution of the weakly labeled dataset.

3.3.4 Assigning an image-level weak label based on DNI classes

The procedure described in [60] is employed to assign a Direct Normal Irradiance (DNI) class to each image, with values ranging from 1 to 8. These correspond to DNI measurements taken over the preceding 15 minutes. Classes 1 to 3 are indicative of clear sky conditions with minimal to moderate variability in DNI. Class 4 exhibits a higher average DNI, accompanied by heightened temperal variability. For classes 5 and 6 the average DNI is significantly lower while class 7 represents almost opaque overcast skies, exhibiting only few ramps. Class 8 describes fully overcast conditions with constantly low DNI. This information is leveraged to assign "overcast" as a second image-level weak label to all images with DNI class 8 in the dataset. The distribution of DNI classes is

illustrated in figure 3.7 on the top left.

3.4 Creation of a human-annotated benchmark dataset

The third dataset utilized in this thesis is the benchmark dataset. It contains 12 images of five ASIs installed at PSA, respectively, resulting in a total of 60 human-annotated images. Three of the utilized ASIs are Mobotix Q25 and Q26 models. The images of the Q25 and the Q26 are very similar in respect to image characteristics. Consequently, the domain shift between images of the two models is expected to be relatively minimal. The fourth ASI is a more recent Mobotix Q71 model with a distinct Complementary Metal-Oxide-Semiconductor (CMOS) sensor, resulting in a notable domain shift. The fifth ASIs is an AXIS model, which generates High Dynamic Range (HDR) images that exhibit significant differences in image characteristics compared to the Q25/Q26 models. An overview of the utilized camera hardware is given in table 3.4. All images were annotated at the pixel level as part of this thesis using a MATLAB graphical user interface.

The benchmark dataset was created to serve as a distinct test set, which is employed for the evaluation of the developed methods in chapter 5. In contrast to the validation dataset from section 3.2, hyperparameters are not optimized with respect to the benchmark dataset, thus ensuring an unbiased comparison between the various approaches explored in this work. Moreover, the effectiveness of the semantic cloud segmentation can be evaluated in respect to domain shifts, specifically in response to changes in camera hardware. 40% of the selected images are complex multi-layer conditions, i.e. images with at least two cloud layers present. One example image for each ASI utilized is shown in figure 3.8.

camera name	**year**	**camera model**	**exp. time [μs]**
Kontas	2021	Mobotix Q25	160
Metas	2023	Mobotix Q26	160
PVot Sky	2023	Mobotix Q26	160
PVot Q71	2023	Mobotix Q71	160
PVot AXIS	2023	AXIS	HDR

Table 3.4: The 5 all-sky-imagers used to create the benchmark dataset.

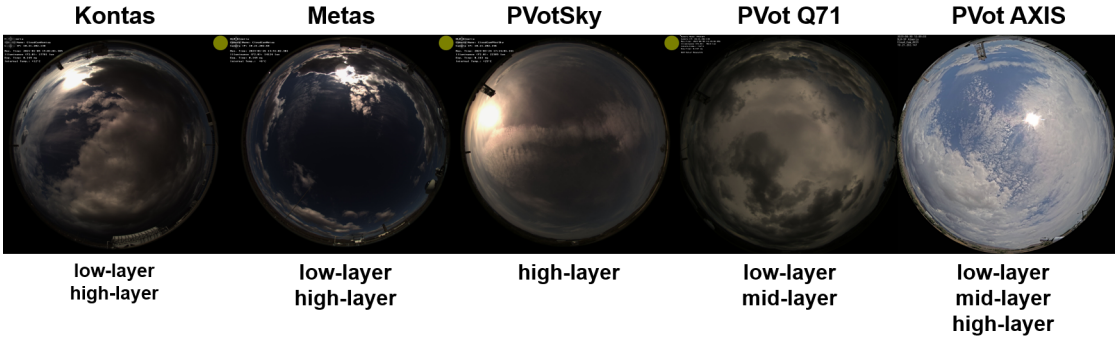


Figure 3.8: Example images from the created benchmark dataset with complex for all ASIs

4 Methods

This chapter presents three methods for semantic cloud segmentation, with each method building on the previous one. The first section presents a method for training a cloud segmentation model with fully supervised learning using only the human-annotated dataset with pixel-level labels introduced in section 3.2. This model trained with this method is used as the baseline model for comparison in the experiments in the following chapter 5. The second section presents a method for the probabilistic calibration of an already trained cloud segmentation model based on *local temperature scaling* [55]. The final section of this chapter presents a method developed as part of this thesis. This method is a semi-supervised approach to training a cloud segmentation model using both human-annotated images and weakly-annotated images with image-level weak labels at the same time. As part of the semi-supervised learning process, a new data augmentation technique called *CloudMix* is proposed. This technique was developed as part of this thesis.

4.1 Fully supervised training of a cloud segmentation model

The most straightforward approach to training a deep learning-based semantic cloud segmentation model is to use purely human-annotated data with class labels at the pixel level and fully supervised learning.

The formal description of the process is as follows. For each iteration step, a batch is utilized, which is a fixed number of images $\mathbf{x}$ and its corresponding ground truth segmentation masks $\mathbf{y}$, which are usually drawn in random order from the dataset. The model makes a prediction on the input images $\mathbf{x}$ and the raw outputs are called logits $\mathbf{z}$. The prediction vector, denoted by $\mathbf{z}$, is then utilized in conjunction with the one-hot encoded ground truth segmentation masks, $\mathbf{y}$, to calculate the cross-entropy loss, L, as defined in Equation 1.

$$L(\mathbf{y}, \mathbf{z}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \log(\sigma_{SM}(z_{ij})) \tag{4.1}$$

N is the number of pixels in the batch and C denotes the number of classes, which are five in this case with one-class respectively for sky, low-layer clouds, mid-layer clouds,

high-layer clouds and one dummy class reserved for the masked part of the images. σ_{SM}, defined in equation 4.2, is called softmax function or also known as normalized exponential function and is used to normalize the logits $\mathbf{z}$ into a probability distribution. This is achieved by scaling each logit into the interval (0,1) and by ensuring that the sum of $\sigma_{SM}(z)$ for all possible classes sums to 1 for each pixel. i denotes the class index for which the probability is calculated.

$$\sigma_{SM}(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{4.2}$$

Once the loss L for the current batch is calculated, the parameters of the model are then updated by an optimizer towards minimizing the loss L using gradient descent, which concludes the iteration. The process for a single iteration is shown in figure 4.1.

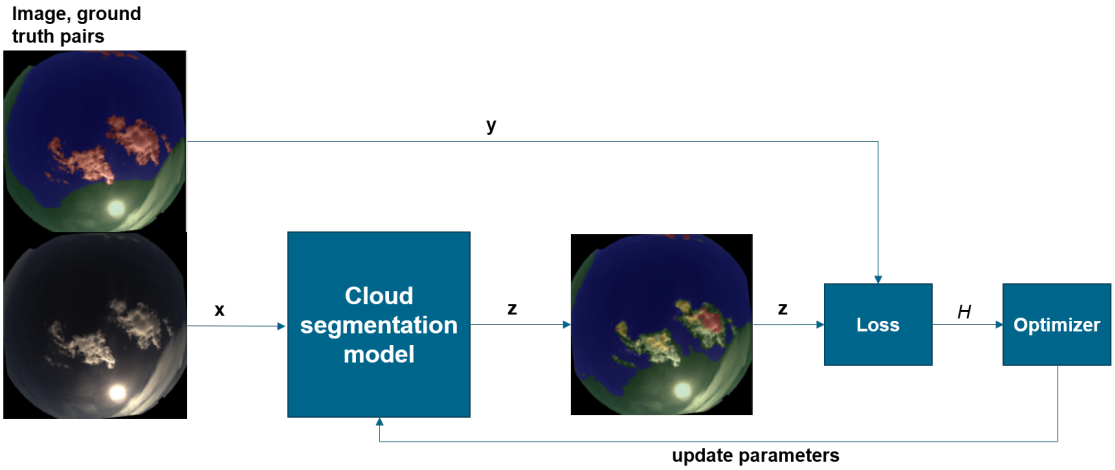


Figure 4.1: One iteration of fully-supervised learning for semantic cloud segmentation.

Once all samples from the dataset have been drawn from the training dataset, an epoch is considered complete. At the conclusion of each epoch, the training progress is evaluated by examining the loss and several metrics of interest on additional validation data. When the validation loss begins to increase for several consecutive epochs, the training is terminated, as further minimization of the loss with regard to the training data will result in overfitting to the training data. This is the rationale behind the necessity to partition the available annotated data into training and validation data, as done in 3.2.

The cloud segmentation model trained with this approach is called fully-supervised model and serves as a baseline for comparison to the methods presented in the following two sections in the experiments of the following chapter 5.

4.2 Probabilistic interpretation and calibration of a cloud segmentation model

This section adopts a probabilistic perspective. It first presents a few formal definitions, after which a method for calibrating a pretrained semantic cloud segmentation model is presented.

4.2.1 Probabilistic interpretation

The raw outputs (logits) $\mathbf{z}$ of a semantic segmentation network for a set of input images $\mathbf{x}$ can be transformed into a probability distribution by calculating $\sigma_{SM}(\mathbf{z})$ the softmax regularization function defined in equation 4.2. The confidence map is then defined as

$$P(x) = \sigma_{SM}(\mathbf{z}(\mathbf{x})) \tag{4.3}$$

where P is a multidimensional matrix with a C-dimensional probability vector for each pixel, where C is the number of possible classes.

The goal would now be that the estimated confidence maps resemble the true ground truth correctness likelihoods. As an example, if all predictions on unseen data with a confidence of 0.8 are evaluated together, the frequency of correct classifications should be at 0.8 for a well perfectly calibrated network.

A significant challenge in practice is that deep semantic segmentation networks tend to become highly overconfident, particularly once overfitting occurs [45]. The cross-entropy loss, as defined in equation 4.1, encourages the network to become overconfident, as the loss becomes zero when all pixels are predicted correctly with a confidence of 1.0 in the training data. This encourages the network to maximize the number of correctly classified samples while minimizing the uncertainty, also known as entropy minimization, in its predictions.

The following section presents a method for addressing this issue by calibrating a pretrained cloud segmentation model, such as the fully-supervised model described in section 4.1.

4.2.2 Probabilistic calibration

The challenge of calibrating the confidence estimates in deep semantic segmentation networks is addressed in [55] with a method called "Local Temperature Scaling for Probability Calibration". A small calibration network, is trained on top of the pretrained

cloud segmentation model to estimate a calibration map T based on the images $\mathbf{x}$ and the predicted logits $\mathbf{z(x)}$ from the pretrained model. The calibration map $\mathbf{T(x, z(x))}$ is a matrix comprising one scalar T_i for each pixel, that is used to calibrate the estimated probability map by calculating:

$$P_{calibrated} = \sigma_{SM}(\frac{\mathbf{z}}{\mathbf{T}}) \tag{4.4}$$

The calibration network is trained as well by minimizing the cross entropy loss 4.1 but this time with respect to the temperature values $\mathbf{T}$, which are the only component in the equation 4.5 that can be changed by updating the weights of the calibration network. The ground truth labels $\mathbf{y}$ and the output logits $\mathbf{z}$ from the segmentation network are fixed during the calibration. The cross-entropy loss calibration can be modified by adjusting the weights of the calibration network, as the ground truth labels $\mathbf{y}$ and the output logits $\mathbf{z}$ from the segmentation network remain constant throughout the calibration process.

$$L_{calibration}(\mathbf{y}, \frac{\mathbf{z}}{\mathbf{T}}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} \log(\sigma_{SM}(\frac{z_{ij}}{T_i})) \tag{4.5}$$

This method is designed to improve the calibration of the network. The minimisation of the loss function encourages the reduction of entropy in correctly classified regions of the images and vice versa for incorrectly classified regions. The calibration network is trained to recognise patterns that can then, during inference, improve the uncertainty estimates on unseen data. The calibration process for one single iteration is illustrated in figure 4.2.

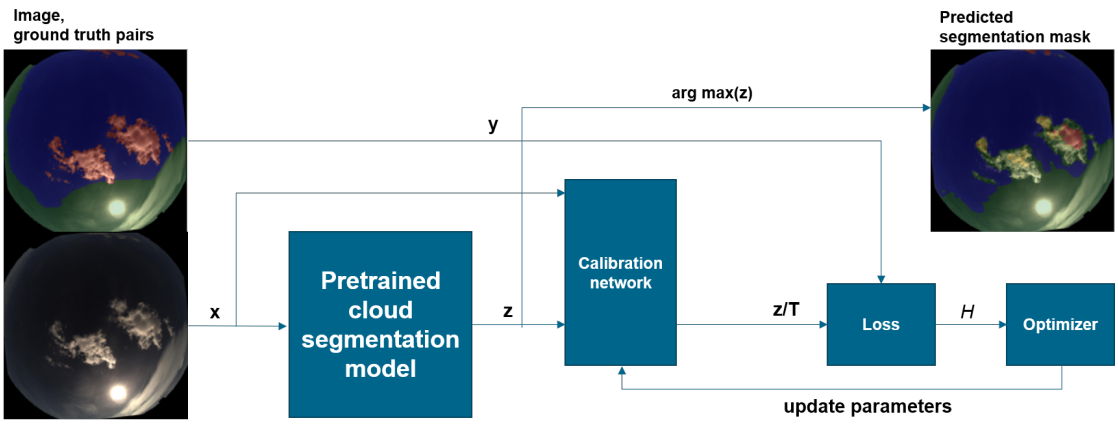


Figure 4.2: Overview of the probabilistic calibration process with local temperature scaling.

It is important to note that the calibration model must be optimized on data samples that differ from those used to optimize the cloud segmentation model. In this case, the validation set of the human-annotated dataset presented in section 3.2 is employed for

calibration. This approach eliminates the need for additional human effort to annotate new data samples.

The arg max operation, which is used to obtain the predicted segmentation map, yields the same results, whether or not the data has been calibrated, as stated in equation 4.6.

$$argmax(\mathbf{z}) = argmax(\frac{\mathbf{z}}{\mathbf{T}}) \tag{4.6}$$

The deterministic behavior of the cloud segmentation remains unaffected by the calibration. This implies that calibration is a post hoc processing step that can be performed subsequent to the training of a cloud segmentation model.

4.3 Semi-supervised learning for semantic cloud segmentation

This section presents a semi-supervised learning approach for semantic cloud segmentation developed as part of this thesis, which explores the suitability of semi-supervised learning in this context. In the context of semi-supervised learning, the image data from two independent sources is utilized simultaneously to train a semantic cloud segmentation model. The first data source is the human-annotated dataset with pixel-level ground truth labels from section 3.2. The second data source is the weakly labeled dataset with image-level weak labels presented in section 3.3. Of particular research interest is if the image-level weak-labels can be leveraged to enhance the model's ability to differentiate between the three cloud layers. This has been identified to be a challenging task [13], due to the limited availability of human-annotated samples and the difficulty of the task itself, given the certain optical similarities between the cloud layers.

First a general overview of the procedure is provided, followed by a detailed view of its sub-parts which are the pseudo-labeling and the weak-to-strong consistency regularization framework.

4.3.1 General overview

From a high-level perspective the semi-supervised learning can be observed as two independent data processing branches. In the first branch, the human-annotated data is processed and in the second branch, the weakly-annotated data is processed. For each branch a separate loss function is calculated. The human-annotated data is utilized to calculate a labeled data loss denoted as L_L. The weakly-annotated data is utilized to calculate a consistency loss denoted as L_C. The consistency loss is calculated based

on generated pseudo-labels with the predictions of a teacher model and the respective predictions of the student model, which is the model to be trained during the semi-supervised learning process. Both branches calculate their respective losses completely independently from each other but are jointly optimized by combining the two losses into a total loss function, denoted as L_{total}, which is defined as follows:

$$L_{total} = L_L + \lambda \cdot L_C \tag{4.7}$$

λ is a hyper-parameter called consistency weight and determines how strong the optimization of the labeled data loss L_L is constrained by the consistency loss L_C. The consistency weight, λ, is often scaled dynamically over the training process, with a low value initially and a gradual increase over the training process [61]. Alternatively, a fixed value can be employed as done in [15][16].

The calculation of the labeled data loss L_L is done analogous to the procedure of the fully-supervised model presented in section 4.1 by calculating the cross entropy loss 4.1 for the predictions of the model and the ground truth labels.

The calculation of the consistency loss L_C is presented in detail through the following subsections. In brief, a pretrained and during the semi-supervised learning unmodified teacher model is utilized to generate pseudo-labels and calculate confidence weights on the weakly-labeled samples with weak supervision through the image-level weak labels. The pseudo-labeling process is wrapped by a weak-to-strong consistency regularization framework to calculate the confidence-weighted consistency loss L_C on the predictions of the student model. This model is jointly optimized by the labeled and weakly labeled data. Figure 4.3 provides a high-level illustration of the learning process.

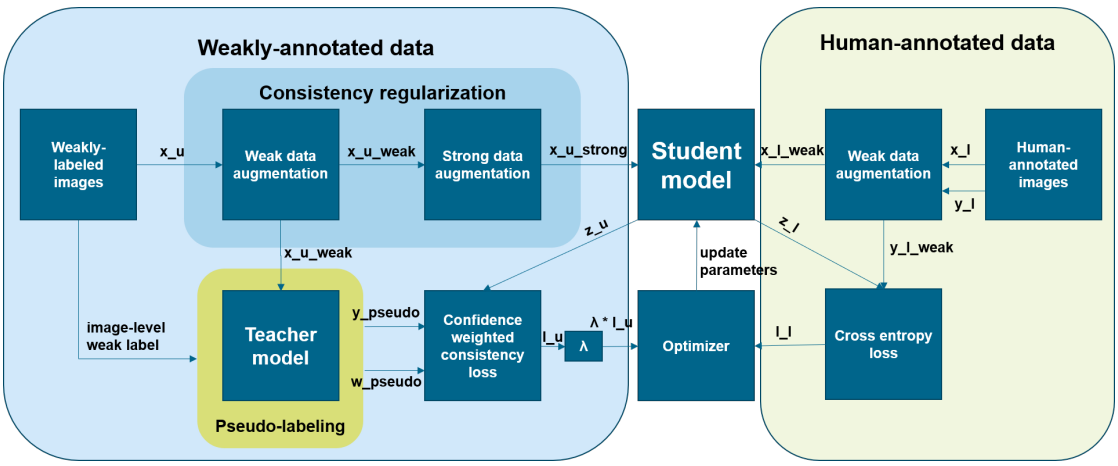


Figure 4.3: A high-level overview of the weakly supervised data flow during semi-supervised learning.

4.3.2 Pseudo-labeling

Pseudo-labeling is a prevalent methodology in semi-supervised learning, whereby artificial labels, or pseudo-labels, are generated on unlabeled data or, in this case, weakly labeled data. The process for semi-supervised semantic cloud segmentation is as follows.

The weakly labeled data $\mathbf{x}_u$ is fed to a teacher model to make predictions $\mathbf{z}_u$. The predictions $\mathbf{z}_u$ are then fused with the respective image-level weak labels by adopting all pixels classified as clear-sky and the pixels classified as one of the three cloud layers, which are overwritten by the image-level weak label. For example, as illustrated in figure 4.4, the teacher model predicted some pixels as high-layer clouds (green), while the image-level weak label for this image is mid-layer (yellow). These high-layer pixels are then overwritten by the image-level weak label as mid-layer pixels. Furthermore, confidence weights are obtained by calculating the confidence for the given class in the pseudo-label. This is achieved by calculating the *softmax* regularization from equation 4.2 $\sigma_{SM}(\mathbf{z})$ for the respective predicted class for each pixel. The generated confidence weights are used to calculate a confidence-weighted consistency loss. The pseudo-labeling of two additional example images is shown in figure 4.5.

The anticipated benefits of employing this pseudo-labeling approach in comparison to utilising a "hard" confidence threshold to discard all pixels below the threshold, as performed in previous works [15][16], are as follows:

1. All pixels are utilized.

2. Pixels with a strong agreement between teacher model and image-level weak label are weighted heavily, while pixels with a strong disagreement are weighted with a lower weight. This can be seen in the example of figure 4.4 where in regions of agreement between teacher and image-level weak label the confidence weight is higher than in regions of disagreement.

In addition to the image-level weak label indicating the prevalent cloud layer in the image, an additional image-level weak label with "overcast" is passed with certain images, as described in the last subsection of 3.3. The "overcast" weak label is employed to overwrite all pixels in the image that have not been masked with the prevalent cloud layer and set the confidence for all pixels to 1.0. This is done to eliminate erroneous predictions made by teacher in overcast situations.

Pseudo-labeling on its own is reported to perform poorly due to over-fitting to noise in the pseudo-labels and confirmation bias as addressed by [62]. Commonly the generated pseudo-labels are utilized for consistency regularization, which is also done in this approach and explained in the following section.

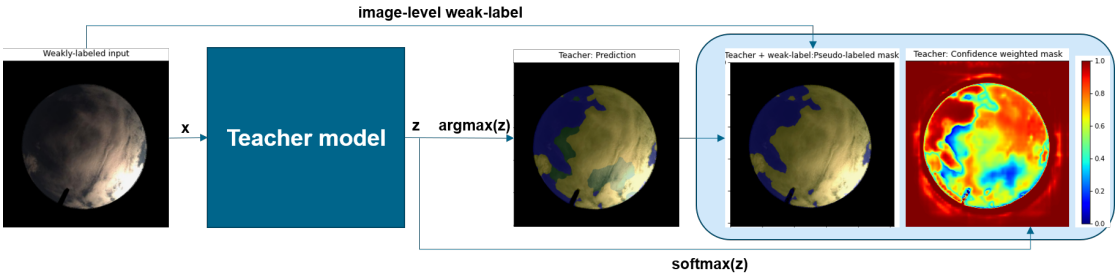


Figure 4.4: The workflow of pseudo-label generation using a teacher model and image-level weak labels.

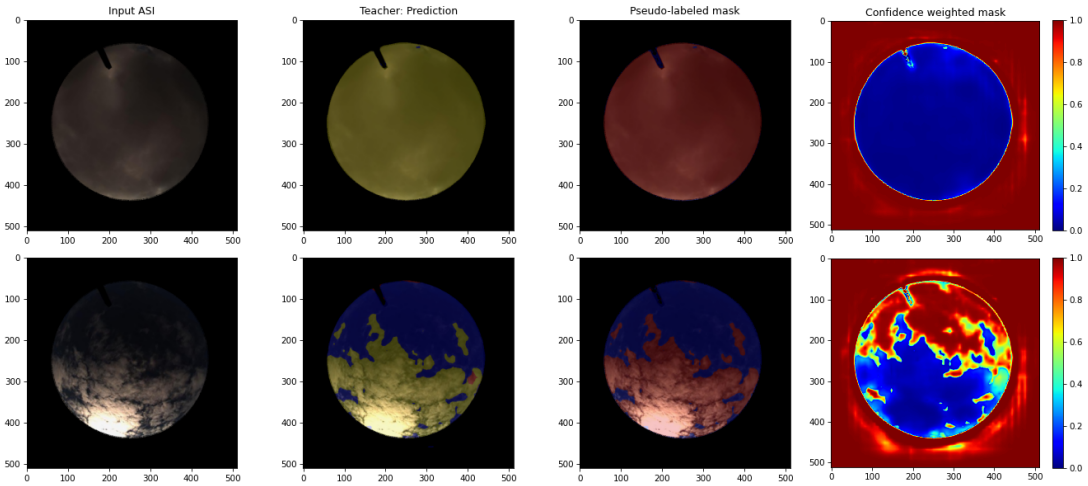


Figure 4.5: Two examples of the pseudo-labeling process.

4.3.3 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. This method uses a weak-to-strong consistency regularization framework popularized by [15] for semi-supervised image classification and recently transferred to the semi-supervised semantic segmentation domain by [16].

Weak data augmentation

Specifically for this case of semantic cloud segmentation, in a first step the weakly labeled images $\mathbf{x}_u$ are augmented using weak data augmentations to obtain weakly augmented views $\mathbf{x}_{uweak}$. 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.

Strong data augmentation

Subsequently, the weakly augmented images $\mathbf{x}_{uweak}$ are augmented a second time with strong data augmentations, resulting in strongly augmented views $\mathbf{x}_{ustrong}$. 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, as illustrated in figure 4.6.

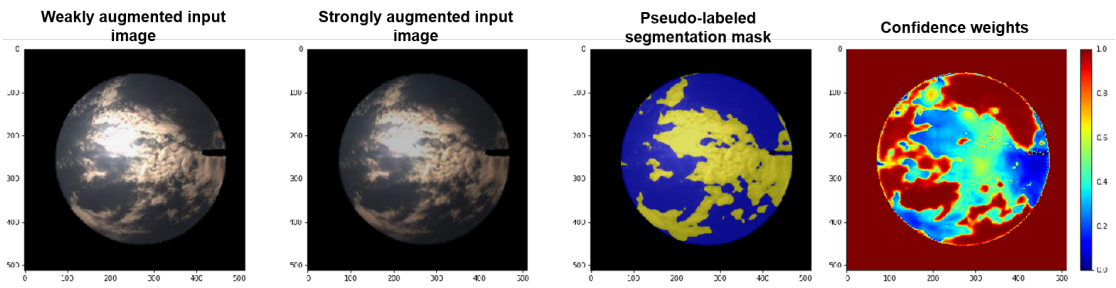


Figure 4.6: Strong data augmentation with color jitter and Gaussian blur.

In addition to the color jittering and Gaussian blurring, a third strong augmentation

technique, called *CloudMix*, is introduced for even stronger perturbation on an image level. This technique is introduced in the following.

CloudMix data augmentation

As previously discussed by [63], semi-supervised semantic segmentation represents a more challenging problem than semi-supervised image classification. This is due to the observation that the cluster assumption does not hold true for semantic segmentation and decision boundaries frequently traverse low-density regions. Semi-supervised semantic segmentation often requires domain specific strong data augmentation. This is because augmentations such as color jittering and Gaussian blurring applied on a pixel level are often insufficient for augmenting images sufficiently strong for successful consistency regularization from a semantic segmentation perspective. A commonly used strong data augmentation technique is *CutMix* proposed by [64]. The *CutMix* technique involves copying and pasting the pixel contents and the respective pixel-level labels from one image into another image for a rectangular region of arbitrary size. However, this augmentation technique may result in unnatural scenery due to the hard cuts on boundaries.

Inspired by this, a novel augmentation technique called *CloudMix* has been invented as part of this thesis. In contrast to conventional approaches that involve the delineation of discrete, rectangular regions, CloudMix respects the height information of the different cloud layers while blending image pairs. This approach ensures the preservation of the original cloud boundaries and yields a more "natural" mixed image. In practice, a pair of all-sky images can be mixed by pasting the pixels from one image into the other in case the class label for that pixel would overlay the class of the other respective pixel in reality. This is achieved by applying the hierarchy defined in table 4.1. 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.

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

Table 4.1: Class hierachy utilized for CloudMix data augmentation.

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 figure 4.7.

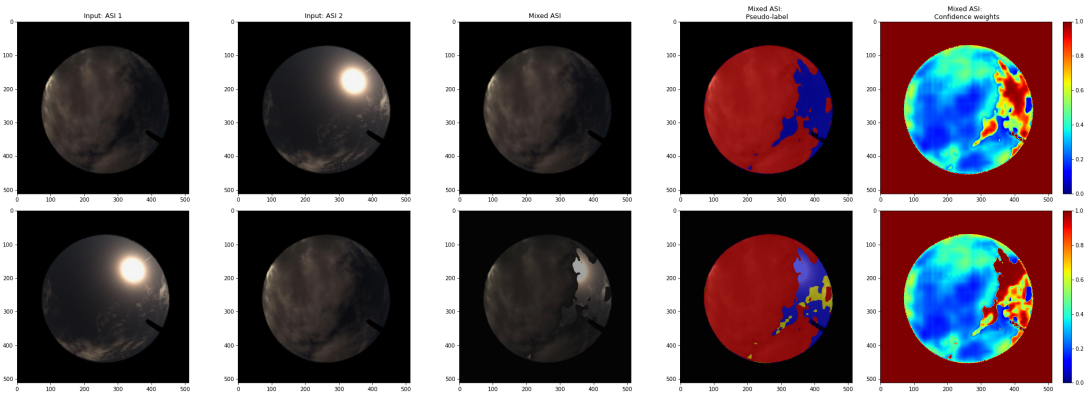


Figure 4.7: Strong data augmentation with CloudMix data augmentation.

Confidence weighted consistency loss

The weak-to-strong consistency framework can now be wrapped together by defining the consistency loss function. The weakly augmented images $\mathbf{x}_{uweak}$ 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 shown in section 4.3.2. In contrast, the student model predicts on the strongly augmented images, denoted by $\mathbf{x}_{ustrong}$. The predictions of the student $\mathbf{z}_u$, the generate pseudo-labels $\mathbf{y}_{pseudo}$, and the confidence weights $\mathbf{w}_{pseudo}$, can then be utilized to calculate a confidence weighted cross entropy loss, also denoted as consistency loss, defined in equation 4.8:

$$L_C(\mathbf{y}_{pseudo}, \mathbf{w}_{pseudo}, \mathbf{z}_u) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} w_i \cdot y_{ij} \log(\sigma_{SM}(z_{ij})) \tag{4.8}$$

This concludes one iteration of semi-supervised learning. The process is repeated analogously to the fully-supervised model until the loss for the validation data no longer improves or even begins to increase. This indicates over-fitting, and the training should be terminated at this point.

5 Experimental results

The main objectives of this thesis are to improve the current SOTA cloud segmentation in terms of its deterministic performance and to improve the probabilistic interpretability for semantic cloud segmentation. So far, a novel semi-supervised learning approach for semantic cloud segmentation using image-level weak labels based on ceilometer measurements and the corresponding weakly annotated dataset has been presented. Also, a deep learning based methodology has been presented to calibrate the confidence estimates of the pre-trained semantic cloud segmentation models in a post-processing step.

This chapter evaluates both the semi-supervised learning approach and the calibration method. First the semi-supervised model is compared to a current SOTA cloud segmentation model [13] and a fully-supervised baseline model on a deterministic benchmark based on the metrics *Accuracy*, Intersection over Union (IoU), *Precision*, *Recall*, and *confusion matrices*. Second the effectiveness of the calibration for semantic cloud segmentation is evaluated on a probabilistic benchmark, based on the metrics Expected Calibration Error (ECE), Maximimum Calibration Error (MCE), and *reliability diagrams*.

5.1 Experimental setup

5.1.1 Utilized hardware and software

All deep learning models developed in this thesis were implemented using the deep learning framework *PyTorch Lightning*. *PyTorch Lightning* is a lightweight wrapper for PyTorch designed to simplify and streamline the process of training deep learning models. *PyTorch* is an open-source machine learning library primarily developed by Facebook's AI Research lab, used for applications such as computer vision and natural language processing [65]. The implementations for all data augmentation techniques, except for the custom *CloudMix*, were imported from *Torchvision* [66]. The implementations for the utilized metrics were imported from *Torchmetrics*, a library of standardized metrics for evaluating machine learning models [67]. All models were trained on *Nvidia RTX A5000 GPU* with 16GB GPU RAM on a *Dell Precision 7560* laptop.

5.1.2 Image pre-processing

All images have to go through several pre-processing steps before they are fed into the models. First the images are masked, to remove all objects not relevant to the task from the scenery using a pre-defined camera mask. The images are then cropped to the fish-eye image section to minimize the amount of masked pixels in the image. The resulting images are then resized to a resolution of 512x512. The RGB values of the images are scaled to the range [0,1] and finally normalised to the mean and variance of the images in the dataset, by subtracting the mean μ and dividing by the standard deviation σ per colour channel. defined as

$$\mu = [0.16622005, 0.1688078, 0.15712574]$$
$$\sigma = [0.18111534, 0.1732183, 0.15360588]$$

5.2 Deterministic evaluation of the semi-supervised learning approach

In this section the developed semi-supervised learning approach is evaluated for its potential to improve semantic cloud segmentation. A model is trained with the presented semi-supervised learning approach and a separate model with identical architecture is trained with fully-supervised learning to provide a baseline for comparison. The two models are compared to a current SOTA semantic cloud segmentation model [13]. All three models are trained on the same human-annotated database. This allows a fair comparison between the three approaches of training semantic cloud segmentation models, taking into account differences in the model architectures.

First the models are evaluated on 36 human-annotated in-domain images. Second the transferability of the models under domain shift is evaluated on 12 human-annotated images from a *Mobotix Q71* ASI with moderate domain shift and on 12 human-annotated HDR images from a *AXIS* ASI with strong domain shift. All images for the benchmark were taken from the benchmark dataset created and presented in section 3.4, where more specific information about the utilized ASIs can be found.

5.2.1 Hyperparameter selection

Fully-supervised baseline model

The fully-supervised baseline model is trained exclusively on the human annotated dataset presented in section 3.2. Training is executed for 40 epochs with a batch size of 4 and a learning rate of 1e-4. The model weights are updated using AdamW [68] with the weight decay set to the default value of 1e-2. The learning rate is scheduled using OneCycleLR learning rate scheduler [69]. The training data is augmented using the augmentations specified in the table 5.1, such as flipping and rotating. Training takes around 20 minutes on the given hardware and is repeated a total of 3 times, to compensate for random fluctuations. The fully-supervised model for the benchmark is selected by choosing the checkpoint with the best mean IoU on the validation set. The best model is selected after 27 epochs with a $mIoU$ of 0.798 as shown in figure 5.1.

Semi-supervised cloud segmentation model

The semi-supervised model is trained on the same human annotated dataset as the fully-supervised model and the weakly-annotated dataset presented in section 3.3. The semi-supervised model is trained for a total of 100 epochs with a batch size of 8 and a learning rate of 5e-4. The same learning rate scheduling and optimiser as utilized to train the fully-supervised baseline model is used to train the semi-supervised model. The augmentations listed in table 5.1 are used as weak augmentations and the augmentations specified in table 5.2 are used as strong augmentations, such as Gaussian blurring and the proposed *CloudMix* augmentation. Each batch consists out of 7 weakly-annotated samples and 1 human-annotated sample to always have some guidance through the human-annotated data for the optimisation process. More samples are dedicated to the weakly-labeled dataset per batch to calculate the consistency loss over more samples and obtain better gradient estimates during back-propagation allowing more stable training. To consider even more weakly-labeled samples per optimization step, gradient accumulation is utilized [70]. Gradient accumulation is set to 8, permitting an effective batch size of 64 with 56 weakly-annotated samples and 8 human-annotated samples per optimization step. The consistency weight λ is statically set to 2. The previously fully-supervised baseline model selected for the benchmark is used as teacher model to generate the pseudo-labels for the semi-supervised model. The training takes about 7 hours on the given hardware. Again, the model for benchmarking is selected by choosing the model checkpoint with the highest mean IoU on the validation set, which is reached after 100 epochs with a value of 0.852 as shown in figure 5.1. All hyperparameters for the described training setups are listed in table 5.3 for comparison.

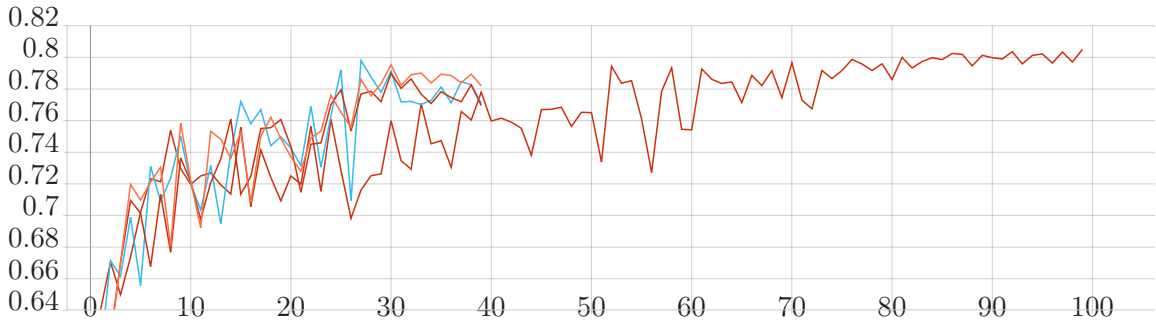


Figure 5.1: Mean IoU on the validation dataset during the training for benchmarking. Fully-supervised runs: blue, orange, red. Semi-supervised run in red.

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

Table 5.1: Utilized weak 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

Table 5.2: Utilized strong data augmentation techniques.

State of the art cloud segmentation model

The current SOTA segmentation model proposed by [13] is used for benchmarking. The model is not trained as part of this thesis but the pre-trained model is used for comparison with the models developed in this thesis. The model was trained on the same human annotated dataset as the baseline and semi-supervised model but employs a U-Net architecture with a ResNet34 backbone, in comparison to the DeepLabv3+ model with ResNet50 backbone employed for the models developed in this thesis. To give a detailed overview of the hyperparameters, the hyperparameter selection for each training setup is listed in table 5.3. A detailed overview of the training process can be found in [13].

Hyperparameter	**Fully-supervised**	**Semi-supervised**	**SOTA** [13]
Input size	512x512	512x512	512x512
Arch., backbone	DeepLabv3+, ResNet50	DeepLabv3+, ResNet50	U-Net, ResNet34
Initialization	ImageNet	ImageNet	self-supervised [71]
Epochs	40	100	2x20
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 [69]	OneCycleLR [69]	OneCycleLR [69]

Table 5.3: Hyperparameter selection for the training of deep cloud segmentation models.

5.2.2 Deterministc semantic segmentation metrics

The deterministic segmentation performance of the three models is evaluated quantitatively with the classification metrics *Accuracy*, *Precision*, *Recall* and the semantic segmentation metric IoU.

Accuracy, Precision and Recall

The *Accuracy* in a semantic segmentation context measures the relative amount of correctly classified pixels related to the total amount of pixels and is defined as

$$Acc = \frac{TP}{TP + TN + FP + FN} \tag{5.1}$$

where TP are the true positives, TN are the true negatives, FP are the false positives and FN are the false negatives.

The *Precision* is calculated separately for each class and measures how the relative amount the true positives, to the sum of true positives and false postives. It is defined as

$$Precision = \frac{TP}{TP + FP} \tag{5.2}$$

The *Recall* is also calculated separately for each class and measures how many of the pixels of the specific class in the ground truth were classified as the respective class. Mathematically it is defined as

$$Recall = \frac{TP}{TP + FN} \tag{5.3}$$

Intersection over Union

The Intersection over Union also known as *Jaccard Index* is a commonly used metric to quantify performance in semantic segmentation tasks. The IoU is defined for a specific class as the fraction of the intersection of the groundtruth labels with the predicted segmentation mask i.e. the true positives, and the union, which is defined as the sum of true positives, false negatives and false positives.

$$IoU_C = \frac{TP}{TP + FP + FN} \tag{5.4}$$

The average IoU over all classes can be calculated as a weighted average of the classwise IoU values:

$$IoU_{AVG} = \frac{1}{N} \sum_{j=1}^{C} IoU_j \cdot w_j \tag{5.5}$$

where N denotes the number of pixels, C the classes and $w_j = TP_j + FN_j$ the support for the respective class.

5.2.3 Deterministic semantic segmentation results

Quantitative evaluation on the benchmark dataset

The semi-supervised model performs best in terms of accuracy and average IoU on the 36 in-domain images of the benchmark dataset. The accuracy is at 68.46%, an improvement of 2.46% over the state-of-the-art. The average IoU is at 55.11%, an improvement of 3.52% over the SOTA. The metrics are lowest for the fully-supervised baseline as shown in table 5.4.

Metric	SOTA [13]	Fully-supervised	Semi-supervised
Accuracy mean	66.00	64.32	**68.46**
IoU mean	51.59	50.25	**55.11**

Table 5.4: Accuracy and IoU mean on the benchmark dataset.

The cloud segmentation is a multi-class problem, so it makes sense to also examine the metrics for each class separately. The following pattern can be observed on the results of the 36 in-domain benchmark images. The semi-supervised model seems to perform best for *mid-layer* and *high-layer* clouds, but not for the *clear-sky* and *low-layer* classes in terms of recall and *IoU*. The *high-layer* recall is improved by 2.46% to 54.83% by a small margin compared to the SOTA. The *high-layer* recall is improved by 25.29% to 67.33% by a much larger margin compared to the SOTA. Similarly, the *IoU* for *mid-layer* is improved by a small margin of 3.34% to a value of 33.75% and improved by 12.20% by a larger margin to a value of 48.77% for *high-layer* clouds compared to the SOTA as shown in table 5.5.

Metric	Class	SOTA [13]	Fully-supervised	Semi-supervised
IoU	Clear-sky	**77.96**	77.15	76.47
IoU	Low-layer	48.23	**49.43**	49.36
IoU	Mid-layer	30.41	26.23	**33.75**
IoU	High-layer	36.57	34.98	**48.77**
Recall	Clear-sky	87.83	**88.32**	81.61
Recall	Low-layer	**74.10**	70.43	63.70
Recall	Mid-layer	52.37	47.45	**54.83**
Recall	High-layer	42.04	41.69	**67.33**
Precision	Clear-sky	87.40	85.92	**92.38**
Precision	Low-layer	58.01	62.37	**68.68**
Precision	Mid-layer	42.04	36.96	**46.75**
Precision	High-layer	**73.74**	68.49	63.89

Table 5.5: Classwise IoU, precision and recall on the benchmark.

The recall for the class *clear-sky* is highest for the fully-supervised model with 88.32%, outperforming the semi-supervised model by 6.71%. For the class *low-layer*, the state-of-the-art model performs best on the in-domain benchmark with an recall of 74.10% outperforming the semi-supervised model by 10.4%.
The same pattern can be observed for *IoU*, but to a much lesser extent. The SOTA model shows with 77.96% the highest *IoU* for *clear-sky* pixels and outperforms the semi-supervised model with 1.49%. For the *low-layer* class no significant difference can be observed as shown in table 5.5.

The large performance gap in terms of recall, which is not as strong by *IoU* and not at all

for precision for the classes *clear-sky* and *low-layer*, could be an indicator of a significant amount of false positives from the-state-of-the-art model and the fully-supervised model for these classes. False positives are not penalised by the recall metric, but are penalised by the IoU and precision metrics.

The classes *clear-sky* and *mid-layer* are the majority classes in the human-annotated dataset presented in section 3.2, as shown in the class data distributions of figure 3.4. The bias towards the majority classes is a common difficulty in deep learning based image classification and semantic segmentation, which has been addressed by [72].

A powerful tool for visualising the strengths and weaknesses of semantic segmentation models are confusion matrices. They are structured as a square matrix where the rows represent the ground truth and the columns represent the classes predicted by the model. The true positives are the diagonal elements, the false positives for each class are all elements in the column except for the diagonal element, and the false negatives for each class are the elements in the specific row except the diagonal elements.

Figure 5.2 shows the confusion matrices for the three models. All values are normalised by the support of the respective class, i.e. by the sum of each row. Thus the diagonals resemble the recall for the respective class of the given row.

The first row for the semi-supervised model on the right shows that the recall for *clear-sky* pixels decreased as more pixels were mistaken for *high-layer* clouds than for the fully-supervised and SOTA models. Conversely, the recall for *high-layer* clouds increased as fewer pixels were mistaken for *clear-sky*, as can be seen in the last row. For *low-layer* clouds more pixels were mistaken for *mid-layer* pixels, and more pixels were mistaken for *high-layer* clouds. For *mid-layer* the confusion with *low-layer* decreased, while the confusion towards *high-layer* increased. Nevertheless, the recall for *mid-layer* clouds improved for the semi-supervised approach.

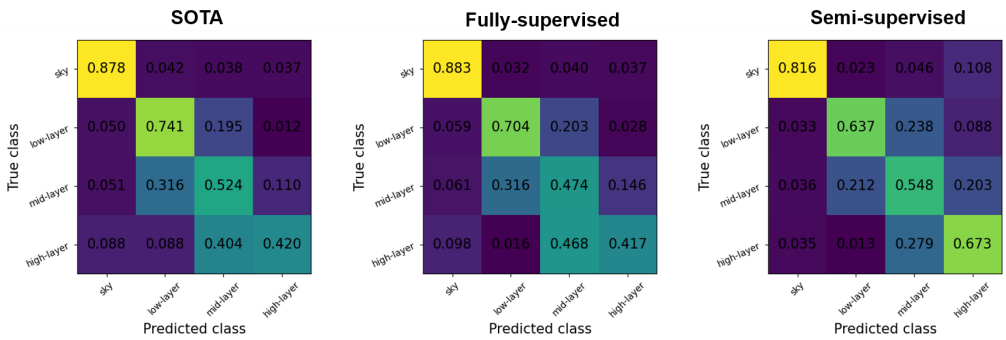


Figure 5.2: Confusion matrices for benchmark dataset.

Quantitative evaluation on the validation dataset

Also on the 154 samples of the validation dataset, the semi-supervised model performs best in terms of accuracy and average IoU. Surprisingly, the fully-supervised model performs almost as well as the semi-supervised model on the validation dataset. The semi-supervised model still performs best in terms of the metrics considered, but only by a small margin. The accuracy improved by 0.47% to 88.67% and the average IoU improved by 0.72% to 80.52% compared to the fully-supervised baseline, as shown in the table 5.6. Interestingly, the fully-supervised baseline outperforms the SOTA on the validation set but is itself outperformed by the SOTA on the benchmark dataset. This could be caused by overfitting by the fully-supervised model to the specific location and camera, as these variables did not change between training dataset and validation dataset.

Metric	SOTA [13]	Fully-supervised	Semi-supervised
Accuracy mean	84.28	88.20	**88.67**
IoU mean	74.71	79.80	**80.52**

Table 5.6: Accuracy and IoU mean on the hold-out validation set.

Qualitative evaluation on the benchmark

So far, the metrics have quantitatively shown that fewer *high-layer* clouds are confused as *mid-layer* clouds. This can also be seen in the first row of figure 5.3, which shows some selected examples of the possible strengths of the semi-supervised approach. The *high-layer cirrus* clouds are incorrectly classified as *mid-layer* and even *low-layer* clouds by the SOTA and fully-supervised models, but are correctly classified as *high-layer* clouds by the semi-supervised model.

The second example shows a scene with again *high-layer cirrus* clouds in the background and *mid-layer altocumulus* clouds in the foreground, which are only correctly classified by the semi-supervised model and misclassified as *low-layer* by the SOTA and baseline. This example also shows nicely, that the semi-supervised models segmented only a small part as *clear-sky* compared to the ground truth. This is an example where even the ground truth could be questioned as it is not very clear where the *high-layer* clouds end and where *clear-sky* begins due to the thin texture of the *cirrus* clouds.

The third and fourth rows show two examples of *mid-layer altocumulus* clouds that were probably misclassified as *low-layer* clouds by the SOTA and the baseline but correctly classified by the semi-supervised model.

The reverse of the previous situation is illustrated in the first two examples of 5.4. In

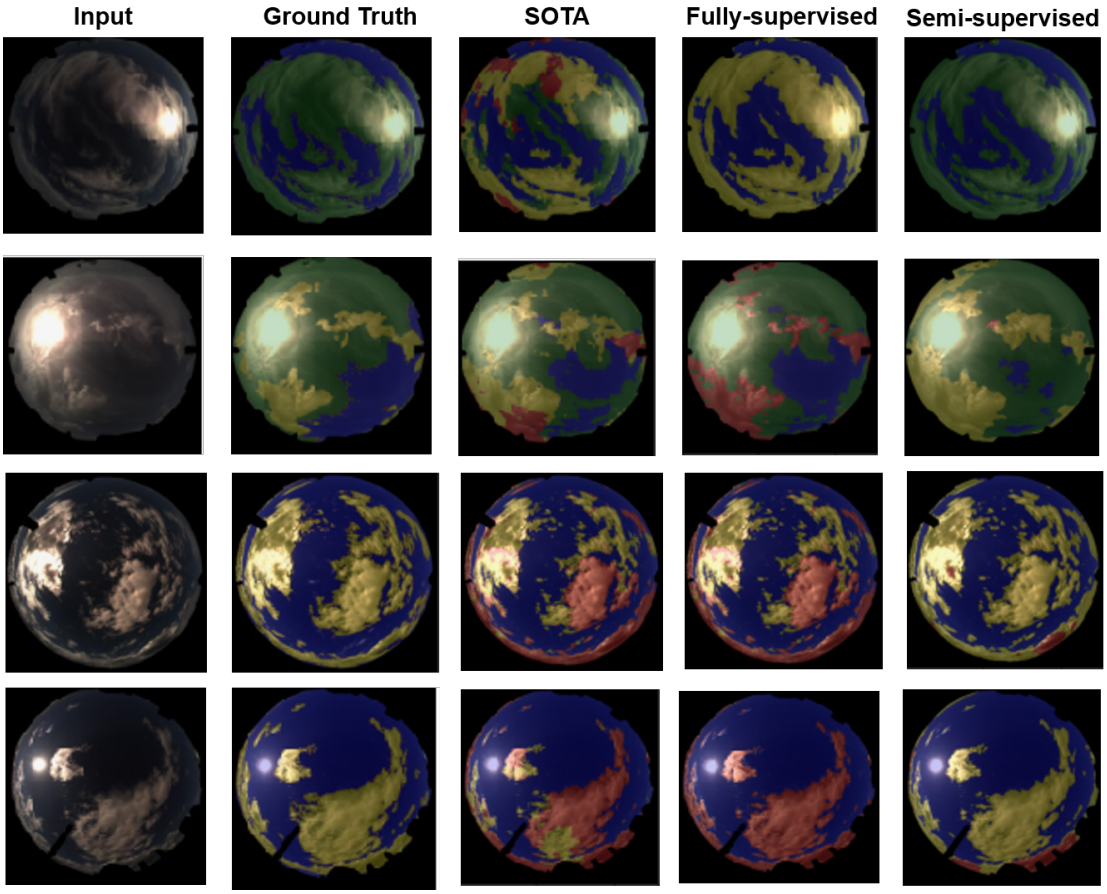


Figure 5.3: Examples from the benchmark resembling the strengths of the semi-supervised model.

both situations, *low-layer cumulus* clouds are mistaken for *mid-layer* clouds, explaining the decrease in *low-layer* accuracy. The third and final example is a good illustration of how complex the scenery can become. The *low-layer* clouds on the bottom are segmented correctly for the most part as well as the *mid-layer* clouds on the right. All three models segmented the as *mid-layer* annotated clouds in the center as *high-layer* clouds. Indeed, the clouds in this part of the image seem to be thinner than in the rest of the image, and it should be questioned whether such misclassifications are harmful or even beneficial for solar nowcasting applications.

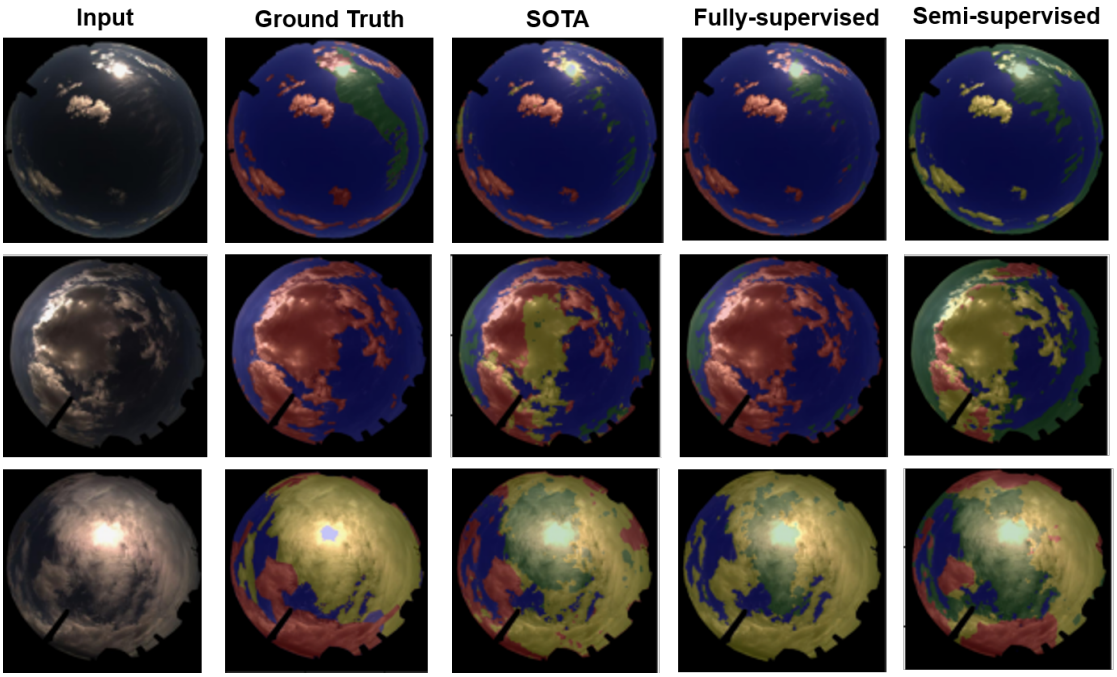


Figure 5.4: Examples from the benchmark resembling challenging conditions for the semi-supervised model.

Overall, the semi-supervised learning showed promising improvements for accuracy and mean IoU on the benchmark dataset comprising 36 all-sky images and the validation dataset comprising 154 all-sky images. From a class-wise perspective the IoU and recall improved for *mid-layer* and *high-layer* clouds with the cost of a lower recall for *low-layer* clouds and *clear-sky*, but still a higher precision for these classes. The shown examples also indicate that the recognition of coherent clouds as a single instance could have improved with the semi-supervised learning, which could be a sign for better contextual understanding of clouds by the semi-supervised model. Also it has been shown, that the ground truth is not clear in all cases, which could motivate to question, if this kind of classification is the most suitable for solar nowcasting applications.

5.2.4 Deterministic performance under domain-shift

The generalization capabilities under domain shift of the three models are evaluated on labeled images of two camera models that are different from the camera models used to acquire the training data. The first camera is a newer *Mobotix Q71* ASI with a different CMOS technology than the *Mobotix Q25/Q26* ASIs from the in-domain benchmark. It captures images with the same exposure time of $160\mu s$ as the cameras used for the training datasets. The second camera is an *AXIS* model that captures HDR images. For each camera model, 12 images were human-annotated and are used for evaluation.

For the samples of the Mobotix Q71 camera, the semi-supervised model outperforms the SOTA and baseline for all classes except for *low-layer* clouds in terms of recall and *IoU*, as can be seen in table 5.7. The average *IoU* is at 60.66% for the predictions of the semi-supervised model, outperforming the baseline and SOTA by a large margin with 17.91%, 23.57% respectively. In particular, *high-layer* clouds appear to be hardly detectable for the SOTA and baseline model for images of this camera model. This may be caused as the images from the Q71 ASI seem a bit darker than the images from the Q25 and Q26 ASIs. The recall for this class is at 13.62% and 28.18% for the two models respectively, while the semi-supervised model was able to identify 64.14% of all pixels annotated as *high-layer* clouds for these samples. For the *low-layer* class the recall is highest for SOTA model with 74.17%, outperforming the semi-supervised model by 32.74% by a wide margin. This isn't resembled in the *IoU* metrics for this class with a value of 20.56% for the SOTA and a 17.55% higher value of 38.11% for the semi-supervised model. Please note that the quantitative results for this benchmark need to be taken with a grain of salt, as the dataset size for this is very small with only 12 images.

Metric	Class	SOTA [13]	Fully-supervised	Semi-supervised
Accuracy	Average	50.36	58.33	**74.21**
IoU	Average	37.09	42.75	**60.66**
IoU	Clear-sky	73.00	74.61	**78.33**
IoU	Low-layer	20.56	36.20	**38.11**
IoU	Mid-layer	29.45	31.10	**50.10**
IoU	High-layer	13.43	22.67	**59.17**
Recall	Clear-sky	80.21	84.44	**87.00**
Recall	Low-layer	**74.17**	69.04	41.42
Recall	Mid-layer	49.42	62.78	**89.23**
Recall	High-layer	13.62	23.18	**64.14**
Precision	Clear-sky	89.04	83.98	**91.54**
Precision	Low-layer	22.15	43.21	**82.68**
Precision	Mid-layer	42.16	38.14	**53.32**
Precision	High-layer	90.68	**91.05**	88.42

Table 5.7: Accuracy, IoU, recall and precision for the Mobotix Q71 out-of domain benchmark.

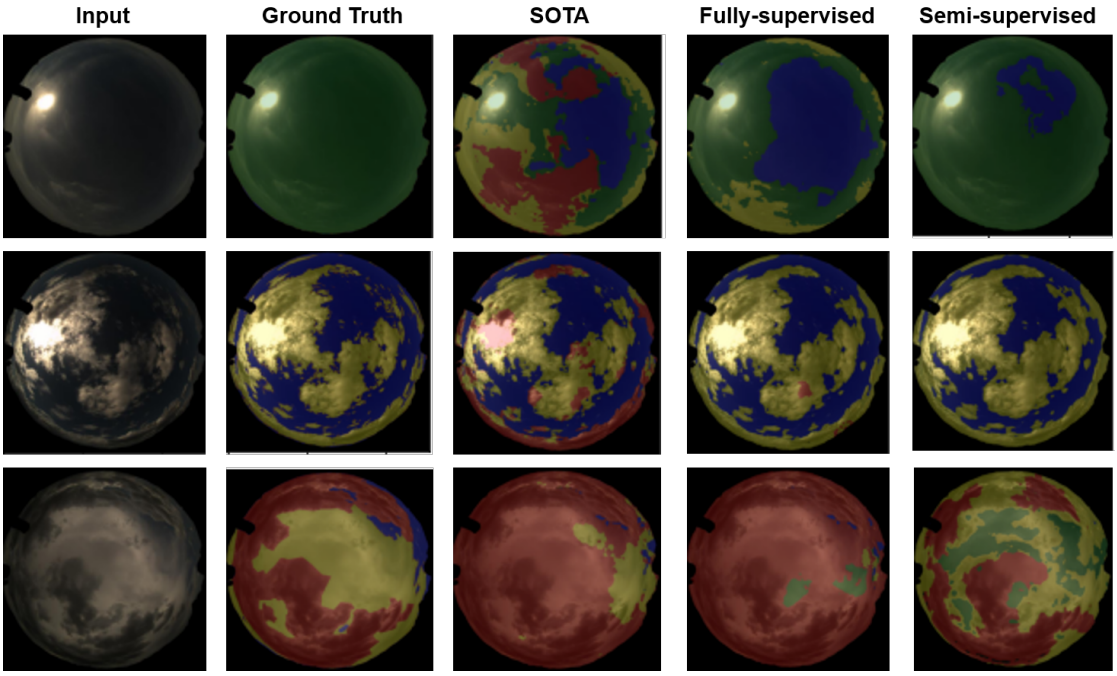


Figure 5.5: Examples from the Mobotix Q71 the out-of domain benchmark.

The improved classification of thin *cirrus* clouds as *high-layer* can also be seen in the first example of the segmentation results for Q71 images in figure 5.5. While the SOTA classifies parts of the scenery as *mid-layer* and even *low-layer*, which is clearly incorrect for this specific case. The semi-supervised model correctly recognises most of the image as *high-layer*. A small part, which is difficult even for a human to annotate due to low local illumination, deviates from the ground truth and is classified as *clear-sky* instead of *high-layer*. The second example shows a *mid-layer* condition correctly classified by the semi-supervised model, while partially misclassified by the SOTA and baseline model as *low-layer* clouds. It is important to note that the state-of-the-art model misclassified *mid-layer* clouds as *low-layer* just in front of the sun disk. Such misclassifications could degrade the performance for solar nowcasting, as the circumsolar region is the most important region in terms of solar irradiance prediction. The third example shows that dark multi-layer overcast conditions appear to be challenging for all three models.

When benchmarking the HDR images, the performance for all three models degrades drastically. In particular, *mid-layer* clouds cannot be reliably detected by the three models. The input image contains clouds of all three classes, but almost every cloud pixel is classified as *high-layer* by the three models, possibly due to the high brightness in the HDR images.

These numbers and the shown examples indicate that cloud segmentation models trained

solely on images with fixed exposure times cannot be naively transferred to the HDR domain. This is to be expected since the domain shift between images with fixed exposure times and HDR images is quite large, as important features for semantic cloud segmentation such as illumination and contrast change drastically between these two domains.

Metric	Class	SOTA [13]	Fully-supervised	Semi-supervised
Accuracy	Average	50.11	**59.23**	57.27
IoU	Average	29.14	**39.58**	35.19
IoU	Clear-sky	41.67	53.21	**63.97**
IoU	Low-layer	6.39	**27.70**	5.92
IoU	Mid-layer	**2.09**	0.0	0.0
IoU	High-layer	49.45	**61.11**	48.23

Table 5.8: Accuracy and IoU for HDR images of AXIS out-of domain benchmark.

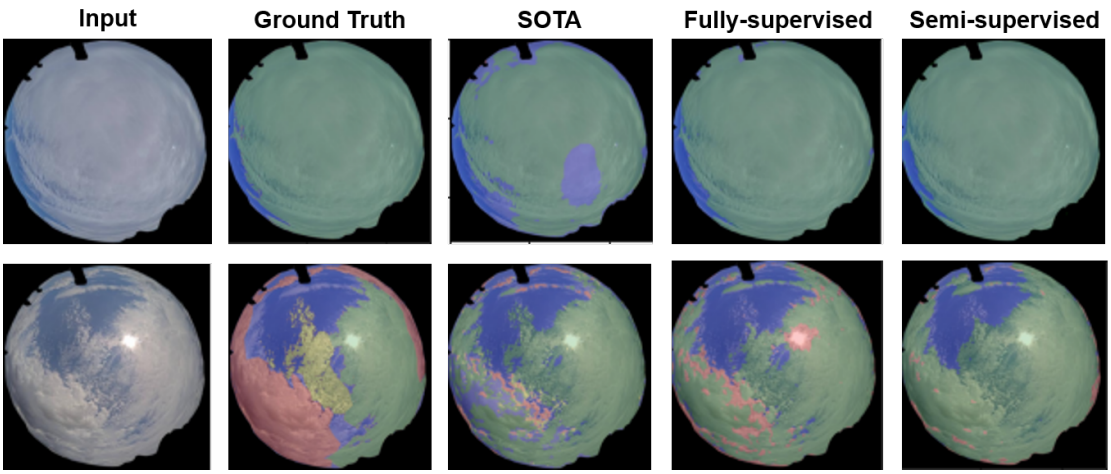


Figure 5.6: Examples from the HDR Mobotix AXIS out-of domain benchmark.

To summarise this section on the generalisation capabilities of the three models under domain shifts, it can be concluded that the evaluation of the samples studied showed the following two trends under consideration of the small sample sizes of the two out-of domain benchmarks:

1. The semi-supervised approach seems to improve the generalization capabilities for small domain shifts, such as changing camera hardware without changing exposure times. This may due to the extensive consistency regularization during semi-supervised learning, which teaches the model to distinguish meaningful features in terms of semantic cloud segmentation.

2. For strong domain shifts, such as when applying a cloud segmentation model to HDR images, the segmentation performance degrades significantly for all approaches. As expected, semi-supervised learning alone cannot bridge such large domain-gaps. The question how to develop a model that also generalizes well in the HDR domain remains to be answered in future research.

5.3 Probabilistic evaluation of the calibration approach

In this section the reliability of the uncertainty estimates of the three cloud segmentation models is investigated. In addition, the calibration technique based on *local temperature scaling* [55] presented in section 4.2 is applied to the semi-supervised model, which performed best in the deterministic benchmark of the previous section. Followed by the evaluation of the applicability of the calibration technique for semantic cloud segmentation with the probabilistic metrics ECE and MCE, and reliability diagrams.

5.3.1 Hyperparameter selection

The validation set of 154 human annotated samples is used for the calibration and is divided with a fixed split into 50% calibration training samples and 50% calibration validation samples. This results in 77 calibration training samples and 77 calibration validation samples available for the calibration. The dataset is artificially augmented using the same weak data augmentation as used for the semi-supervised learning listed in table 5.1. The calibration network is trained for 50 epochs with a batch size of 16 samples and a learning rate of 5e-4. The learning rate is scheduled using the OneCycleLR learning rate scheduler [69]. The calibration network parameters are optimized using the Adam [73] optimizer. The calibration model used for the probabilistic evaluation is the one with the lowest validation loss during the calibration procedure. The calibration takes about 16 minutes on the given hardware. All hyperparameters for the calibration are listed in the table 5.9.

5.3.2 Probabilistic metrics

The probabilistic interpretability will be evaluated quantitatively with the calibration metrics ECE and MCE. In addition, the effectiveness of the applied calibration method is visualized with *reliability diagrams*.

The ECE quantifies how well a given model is calibrated, i.e. measures how well the predicted output probabilities of the model match the actual probabilities of the ground

Hyperparameter	Value
Input size	512x512
Architecture	CNN, adapted from [55]
Initialization	zeros, adapted from [55]
Epochs	50
Batch size	16
Training samples	77 training, 77 validation
Optimizer	Adam [73]
Learning rate	5e-4
Scheduler	OneCycleLR [69]
Data augmentations	weak data augmentations, table 5.1

Table 5.9: Hyperparameter selection for the probabilistic calibration.

truth distribution. It is defined as

$$ECE = \sum_{i=1}^{N} b_i ||p_i - c_i|| \tag{5.6}$$

where N is the number of bins, p_i is the top-1 prediction accuracy in bin i, c_i is the average confidence of predictions in bin i, and b_i is the fraction of data points in bin i [67]. The MCE measures the worst case calibration error by calculating the infinity norm

$$MCE = \max_i (p_i - c_i) \tag{5.7}$$

where i denotes the bin i [67]. To obtain the numerical results presented in the following the number of bins N is set to the default value 15.

5.3.3 Probabilistic calibration results

First the uncalibrated models are compared in terms of ECE and MCE. Interestingly, the semi-supervised model, which performed best on most of the deterministic metrics on the same data, has the worst ECE of all three models, as shown in table 5.10. The ECE is at 0.176 for the semi-supervised model, 0.109 for the SOTA model and lowest with 0.105 for the fully-supervised model, which performed worst on the deterministic benchmark. Meanwhile, the SOTA model shows the highest possible MCE with 1.0, followed by the semi-supervised model with 0.276. The fully-supervised model also shows the lowest MCE with 0.188, similar to the lowest ECE.

In contrast, for the calibrated semi-supervised model, the calibration errors improved significantly. The ECE decreased from 0.176 to 0.037 by a large margin of over 80%. The MCE decreased by over 70% from 0.276 to 0.111. These quantitative results are

promising and may indicate the effectiveness of calibrating semantic cloud segmentation models with *local temperature scaling* to improve the probabilistic interpretability.

An interesting question might be why the semi-supervised model had the highest calibration errors without the calibration post-processing step. A possible explanation could be, that the semi-supervised model was trained for 100 epochs, whereas the SOTA and baseline models were trained for 40 epochs. In other words, the semi-supervised model was optimized 100 times on the same human-annotated samples, while the other two models were optimized only 40 times on the same human-annotated samples. More optimization iterations on the same data often leads to overconfidence, as discussed in [45].

Metric	SOTA [13]	Fully-supervised	Semi-supervised	Calibrated Semi-supervised
ECE	0.109	0.105	0.176	**0.037**
MCE	1.0	0.188	0.276	**0.111**

Table 5.10: ECE and MCE for the state of the art model, the fully-supervised model, semi-supervised model and the calibrated semi-supervised model on the benchmark.

Reliability diagrams, also known as calibration curves, are a visual tool for assessing calibration in different confidence regions [55][45]. They are commonly used for classification tasks and can be transferred to the semantic segmentation domain. The calibration curve for a perfectly calibrated model theoretically lies on the diagonal $y = x$, as indicated by the grey line in the reliability diagrams in figure 5.7. Points above this line indicate underconfidence, while points below indicate overconfidence for the respective confidence region.

The calibration curves for the semi-supervised model and the calibrated semi-supervised model are shown in figure 5.7. The uncalibrated model is far from being perfectly calibrated and shows overconfidence for all three cloud classes for confidence levels above 0.4. For the *mid-layer* class the overconfidence is strongest and the observed relative frequency is below 0.5 for all confidence ranges. In contrast, for the *clear-sky* class, underconfidence is be observed for the higher confidence regions.

The reliability diagrams for the calibrated model show better confidence estimates for the *low-layer* and *high-layer* classes across all confidence regions, as the calibration curves are situated closer to the diagonal after the calibration. For the *clear-sky*, class the underconfidence was shifted from the low confidence region to the high confidence region. For the *clear-sky*, *low-layer* and *high-layer* classes, the calibration curves indicate underconfidence for predictions with high confidence, i.e. predictions with high confidence for these three classes could be trusted with high certainty.

The confidence estimates for the *mid-layer* class improved for the predictions with confi-

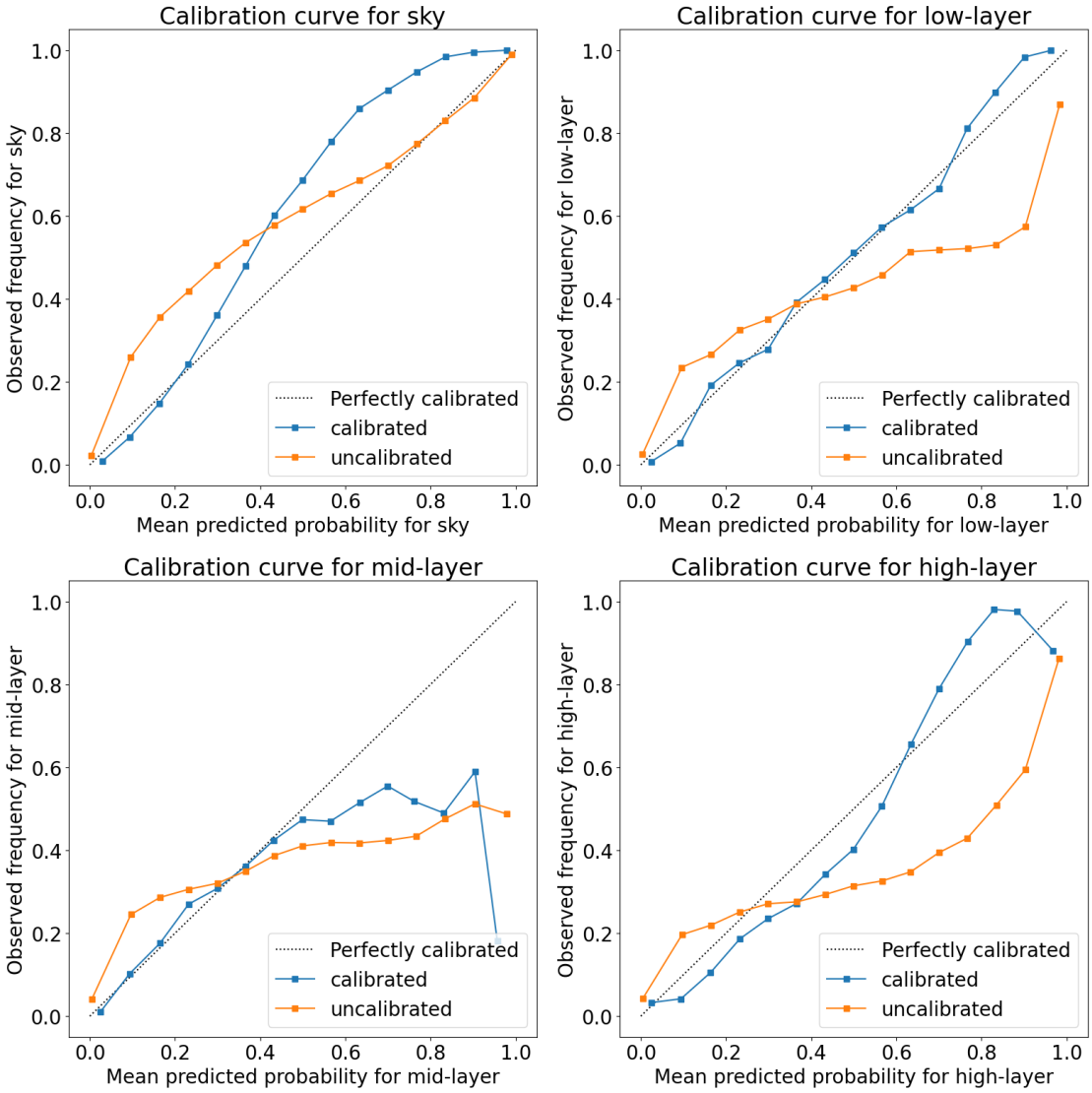


Figure 5.7: Reliability diagrams for the uncertainty estimates of the semi-supervised model and the calibrated semi-supervised model for the 36 in-domain benchmark images.

dence levels below 0.5, but at higher confidence levels the model remains overconfident. This could be due to the uncertainty in the model for this class but also by the uncertainty in the data itself, which cannot be bridged by any calibration method. The *mid-layer* class in the real world is an intermediate class, with strong visual similarities to the *low-layer* and *high-layer* classes. For instance *altocumulus* clouds in the *mid-layer* class are difficult to classify with high confidence and are easily confused with *cumulus* clouds in the *low-layer* class and *cirrocumulus* clouds in the *high-layer* class. This is also true for human experts, when solely one single all-sky image is provided. Often additional information such as ceilometer measurements and inspection of videos of all-sky images is required to identify *mid-layer* clouds with high confidence. Thus, high confidence ground truth predictions may simply not exist for this class based solely on single all-sky image. This may explain why the calibration method seems reach its limits in calibrating *mid-layer* predictions with high confidence.

The potential of the *local temperature scaling* calibration method to improve the confidence estimates of semantic cloud segmentation can be seen in the three examples shown in figure 5.8. In the first example the estimated temperatures are highest for the difficult to classify region with *cirrocumulus* clouds of class *high-layer*, which reduces the estimated confidence in the region significantly. In the second and third examples the temperatures are highest in the circumsolar region, which is usually difficult to classify due to effects such as oversaturation. For less ambiguous regions such as parts with *clear-sky* and the *cumulus* clouds of class *low-layer*, the estimated temperatures are lower and the confidence remains high.

Overall, the results of the *local temperature scaling* calibration method applied to the semantic cloud segmentation task are very promising. Without any additional human-annotation effort and with almost no computational overhead during inference, the calibration metrics on the benchmark of the calibrated semi-supervised model was significantly improved compared to the uncalibrated model. In future works, the confidence estimates could possibly be further improved by using a larger calibration dataset and even a larger architecture for the calibration network.

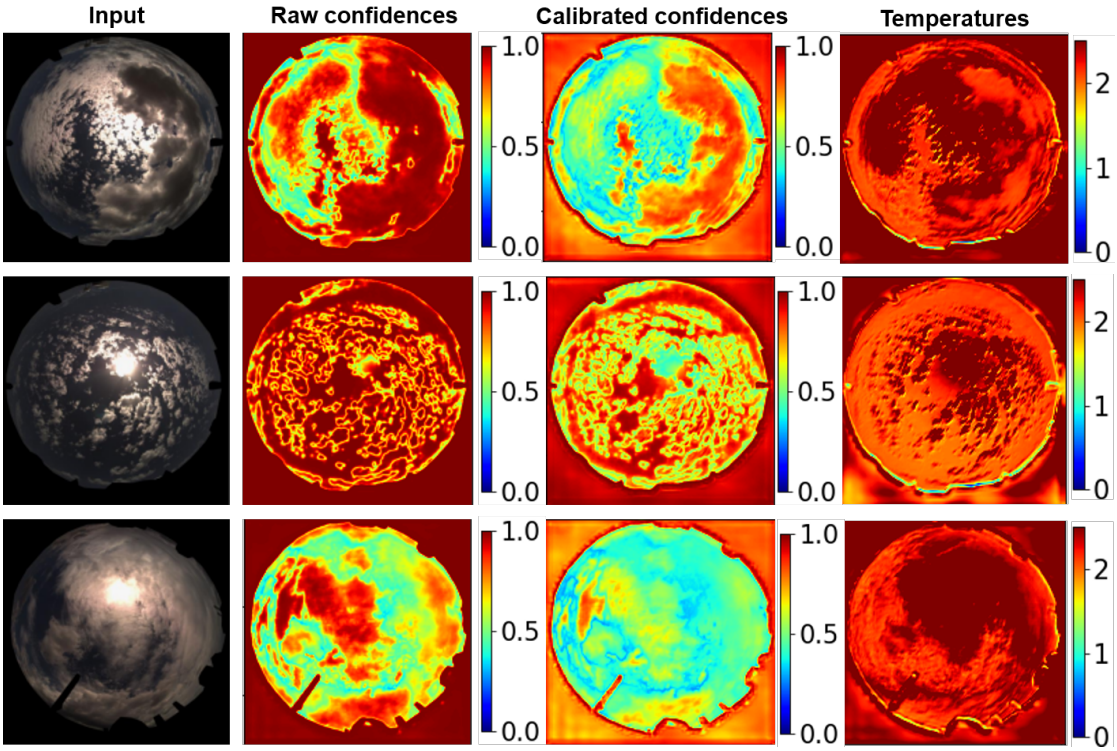


Figure 5.8: Examples for the calibrated confidences on the benchmark for the calibrated semi-supervised model. From left to right: Input image, uncalibrated confidences, calibrated confidences, local temperatures estimated by the calibration network.

6 Conclusion and outlook

This thesis investigated the effectiveness of semi-supervised learning techniques for cloud detection in ground-based imagery. Also, the general probabilistic interpretability of semantic cloud segmentation models and specifically the effectiveness of a *local temperature scaling* [55] probability calibration method for semantic cloud segmentation was investigated.

6.1 Conclusion

The main contributions of this thesis are:

- A new dataset of over 47000 unique all-sky images with image-level weak labels based on heuristics applied to ceilometer measurements.
- A new dataset of mostly complex multi-layer cloud conditions for benchmarking semantic cloud segmentation models with 60 on a pixel-level annotated all-sky images from 5 different all-sky imagers.
- A new approach to semi-supervised learning for semantic cloud segmentation using advanced SOTA training techniques such as image-level weak labels, pseudo-labeling and consistency regularization. As part of the consistency regularization, a novel data augmentation technique called *CloudMix*, tailor-made for semantic cloud segmentation, has been proposed.
- Studies on the effectiveness of the developed semi-supervised learning approach for semantic cloud segmentation.
- Studies on the effectiveness of *local temperature scaling* [55] for the probability calibration of semantic cloud segmentation models.
- Investigation of the transferability of semantic cloud segmentation models to new domains, such as changing camera hardware and the usage of HDR images.

- Development of a probability calibrated semi-supervised learning based semantic cloud segmentation model outperforming the SOTA on deterministic metrics such as accuracy and mean IoU, and on probabilistic metrics such as ECE and MCE on a benchmark with 36 human annotated images from three cameras.

The semantic cloud segmentation distinguishes between the four classes: *clear-sky*, *low-layer*, *mid-layer*, *high-layer*. The model trained with the proposed semi-supervised learning approach showed promising results on the benchmark. The accuracy and the mean IoU could be improved by 2.4% points and 3.5% points respectively compared to the SOTA [13]. Especially the detection of the classes *mid-layer* and *high-layer* seems to benefit from the semi-supervised learning approach with improvements of 3.34% and 12.2% points in *IoU* respectively. For the classification for the classes *clear-sky* and *low-layer* no significant improvements could be observed on the benchmark. The *mid-layer* class remains the most difficult to classify and is still often confused with the neighboring cloud layers due to optical similarities. This is true for both deep learning based cloud segmentation models and human experts.

It is important to note that the number of available images for the benchmarking is limited due to the difficulty of human annotation. Consequently, minor differences in the predictions can result in significant fluctuations in the metrics. Nevertheless, the observed quantitative improvements in the metrics were accompanied by observable qualitative improvements in the evaluation. In particular, the recognition of coherent clouds as a single instance seems to have improved with the semi-supervised learning, which could indicate a better contextual understanding of clouds by the model.

In addition, the semi-supervised learning showed promising results in bridging moderate domain gaps, such as changing camera hardware without changing exposure times. The state-of-the-art model was outperformed by 23.8% points in pixel accuracy and by 23.5% points in mean IoU on an out-of-domain benchmark evaluated on 12 all-sky images from a different ASI than the models were trained on. This could be an indicator of improved generalization capabilities due to the extensive consistency regularization via the semi-supervised learning approach.
As expected, the domain gap from fixed exposure images to HDR images proved to be too large, as none of the existing cloud segmentation models could perform decently on the 12 HDR images evaluated. The question how to develop a model that also generalizes well in the HDR domain remains to be answered in future research.

Also, the *local temperature scaling* probability calibration is effective to improve the probabilistic interpretability of semantic cloud segmentation models. The calibration method was applied to calibrate the semi-supervised model, and the expected calibration error and maximum calibration error metrics decreased significantly from 0.176 to 0.037 and from 0.276 to 0.111 respectively compared to the uncalibrated semi-supervised model on the probabilistic benchmark. Reliability diagrams revealed overconfidence on the part

of the uncalibrated model for all three cloud classes on the evaluated images, which could be mitigated by the calibration method for most parts, except for *mid-layer* predictions with high confidence.

In conclusion, this thesis proposes a novel semi-supervised learning approach for semantic cloud segmentation, which enables learning not only from images but also from ceilometer measurements. Furthermore, it was demonstrated that *local temperature scaling* calibration can significantly improve the probabilistic interpretability of semantic cloud segmentation.

6.2 Outlook

The next step is to integrate the developed probabilistic cloud segmentation model into the existing physical nowcasting system to benchmark it with the SOTA model to validate its usefulness in the target domain for solar irradiance prediction. If this validation is successful, semi-supervised learning could be re-iterated by employing the new probabilistic model as a teacher, to obtain an even better model, followed by the probabilistic calibration. This process may be repeated several times.

Subsequently, the ground truth database could be extended using *active learning* techniques [74][75], which are commonly used in the medical imaging domain [76] to leverage human annotation efforts. In particular, methods that select samples based on the uncertainty, could benefit from the improved uncertainty estimates provided by probabilistic calibration.

In addition, integrating more meaningful information into the cloud segmentation model during learning and inference could help to learn better feature representations. Optical flow [77] or video data [78][79] could be integrated to learn from motion, which is very useful for human experts to distinguish different cloud layers in complex multi-layer scenarios. In particular, *mid-layer* clouds proved to be nearly impossible to classify with high confidence based on the information available in single all-sky images alone, as was shown also in a previous work [13].

I strongly encourage further research on this topic, as I still see a lot of potential to improve deep learning-based cloud segmentation and believe in the possibility of outperforming even human experts at this task in the long run.

Bibliography

[1] C. J. Hahn, W. B. Rossow, and S. G. Warren, "Isccp cloud properties associated with standard cloud types identified in individual surface observations," *Journal of Climate*, vol. 14, no. 1, pp. 11 –28, 2001. DOI: 10.1175/1520-0442(2001)014<0011:ICPAWS>2.0.CO;2. [Online]. Available: https://journals.ametsoc.org/view/journals/clim/14/1/1520-0442_2001_014_0011_icpaws_2.0.co_2.xml.

[2] D. Giggenbach, M. T. Knopp, and C. Fuchs, "Link budget calculation in optical leo satellite downlinks with on/off-keying and large signal divergence: A simplified methodology," *International Journal of Satellite Communications and Networking*, vol. 41, no. 5, pp. 460–476, 2023. DOI: https://doi.org/10.1002/sat.1478. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sat.1478. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sat.1478.

[3] N. B. Blum, S. Wilbert, B. Nouri, *et al.*, "Analyzing spatial variations of cloud attenuation by a network of all-sky imagers," *Remote Sensing*, vol. 14, no. 22, 2022, ISSN: 2072-4292. DOI: 10.3390/rs14225685. [Online]. Available: https://www.mdpi.com/2072-4292/14/22/5685.

[4] H. Wen, Y. Du, X. Chen, *et al.*, "Deep learning based multistep solar forecasting for pv ramp-rate control using sky images," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 2, pp. 1397–1406, 2021. DOI: 10.1109/TII.2020.2987916.

[5] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. de Pison, and F. Antonanzas-Torres, "Review of photovoltaic power forecasting," *Solar Energy*, vol. 136, pp. 78–111, 2016.

[6] J. Marcos, L. Marroyo, E. Lorenzo, D. Alvira, and E. Izco, "Storage requirements for pv power ramp-rate control," *Solar Energy*, vol. 86, no. 10, pp. 2677–2684, 2011.

[7] K. Shields, J. Tovar-Pescador, and F. J. Batlles, "Performance optimization of concentrating solar power plants through use of direct normal irradiance forecasting," *Renewable Energy*, vol. 75, pp. 518–524, 2015.

[8] Z. Peng, D. Yu, D. Huang, J. Heiser, S. Yoo, and P. Kalb, "3d cloud detection and tracking system for solar forecast using multiple sky imagers," *Solar Energy*, vol. 118, pp. 496–519, 2015, ISSN: 0038-092X. DOI: https://doi.org/10.1016/j.solener.2015.05.037. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038092X15002972.

[9] B. Nouri, S. Wilbert, N. Blum, *et al.*, "Probabilistic solar nowcasting based on all-sky imagers," *Solar Energy*, vol. 253, pp. 285–307, 2023, ISSN: 0038-092X. DOI: https://doi.org/10.1016/j.solener.2023.01.060. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038092X23000683.

[10] W. M. Organization, *International Cloud Atlas*. World Meteorological Organization, 2017. [Online]. Available: https://cloudatlas.wmo.int/.

[11] B. Nouri, S. Wilbert, L. Segura, *et al.*, "Determination of cloud transmittance for all sky imager based solar nowcasting," *Solar Energy*, vol. 181, pp. 251–263, 2019, ISSN: 0038-092X. DOI: https://doi.org/10.1016/j.solener.2019.02.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038092X19301306.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[13] Y. Fabel, B. Nouri, S. Wilbert, *et al.*, "Applying self-supervised learning for semantic cloud segmentation of all-sky images," *Atmospheric Measurement Techniques*, vol. 15, no. 3, pp. 797–809, 2022. DOI: 10.5194/amt-15-797-2022. [Online]. Available: https://amt.copernicus.org/articles/15/797/2022/.

[14] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4981–4990.

[15] K. Sohn, D. Berthelot, C.-L. Li, *et al.*, *Fixmatch: Simplifying semi-supervised learning with consistency and confidence*, 2020. arXiv: 2001.07685 [cs.LG].

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

[17] C. Long, J. Sabburg, J. Calbó, and D. Pages, "Retrieving cloud characteristics from ground-based daytime color all-sky images," *Journal of Atmospheric and Oceanic Technology - J ATMOS OCEAN TECHNOL*, vol. 23, May 2006. DOI: 10.1175/JTECH1875.1.

[18] Q. Li, W. Lyu, and J. Yang, "A hybrid thresholding algorithm for cloud detection on ground-based color images," *Journal of Atmospheric and Oceanic Technology*, vol. 28, pp. 1286–1296, Oct. 2011. DOI: 10.1175/JTECH-D-11-00009.1.

[19] A. Heinle, A. Macke, and A. Srivastav, "Automatic cloud classification of whole sky images," *Atmospheric Measurement Techniques*, vol. 3, no. 3, pp. 557–567, 2010. DOI: 10.5194/amt-3-557-2010. [Online]. Available: https://amt.copernicus.org/articles/3/557/2010/.

[20] A. Kazantzidis, P. Tzoumanikas, A. Bais, S. Fotopoulos, and G. Economou, "Cloud detection and classification with the use of whole-sky ground-based images," *Atmospheric Research*, vol. 113, pp. 80–88, 2012, ISSN: 0169-8095. DOI: https://doi.org/10.1016/j.atmosres.2012.05.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169809512001342.

[21] V. Jayadevan, J. Rodriguez, and A. Cronin, "A new contrast-enhancing feature for cloud detection in ground-based sky images," *Journal of Atmospheric and Oceanic Technology*, vol. 32, pp. 209–219, Feb. 2015. DOI: 10.1175/JTECH-D-14-00053.1.

[22] A. Taravat, F. Del Frate, C. Cornaro, and S. Vergari, "Neural networks and support vector machine algorithms for automatic cloud classification of whole-sky ground-based images," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 3, pp. 666–670, 2015. DOI: 10.1109/LGRS.2014.2356616.

[23] W. Xie, D. Liu, M. Yang, *et al.*, "Segcloud: A novel cloud image segmentation model using a deep convolutional neural network for ground-based all-sky-view camera observation," *Atmospheric Measurement Techniques*, vol. 13, no. 4, pp. 1953–1961, 2020. DOI: 10.5194/amt-13-1953-2020. [Online]. Available: https://amt.copernicus.org/articles/13/1953/2020/.

[24] L. Ye, Z. Cao, Y. Xiao, and Z. Yang, "Supervised fine-grained cloud detection and recognition in whole-sky images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 7972–7985, 2019. DOI: 10.1109/TGRS.2019.2917612.

[25] M. Reinhardt, S. Y. Schoger, F. Kurzrock, and R. Potthast, "Convective-scale assimilation of cloud cover from photographs using a machine learning forward operator," *Artificial Intelligence for the Earth Systems*, vol. 2, no. 2, e220025, 2023. DOI: 10.1175/AIES-D-22-0025.1. [Online]. Available: https://journals.ametsoc.org/view/journals/aies/2/2/AIES-D-22-0025.1.xml.

[26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[27] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015. arXiv: 1512.03385 [cs.CV].

[28] O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, 2015. arXiv: 1505.04597 [cs.CV].

[29] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, *Encoder-decoder with atrous separable convolution for semantic image segmentation*, 2018. arXiv: 1802.02611 [cs.CV].

[30] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[31] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[32] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd. Springer, 2009.

[33] R. Balestriero, M. Ibrahim, V. Sobal, *et al.*, *A cookbook of self-supervised learning*, 2023. arXiv: 2304.12210 [cs.LG].

[34] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, pp. 373–440, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:208044535.

[35] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*. MIT Press, 2006.

[36] D.-H. Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:18507866.

[37] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, *Self-training with noisy student improves imagenet classification*, 2020. arXiv: 1911.04252 [cs.LG].

[38] Y. Zou, Z. Yu, B. V. K. V. Kumar, and J. Wang, *Domain adaptation for semantic segmentation via class-balanced self-training*, 2018. arXiv: 1810.07911 [cs.CV].

[39] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 5049–5059.

[40] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow, *Realistic evaluation of deep semi-supervised learning algorithms*, 2019. arXiv: 1804.09170 [cs.LG].

[41] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Toronto, ON, Canada, Technical Report TR-2009, 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[42] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, *Pseudo-labeling and confirmation bias in deep semi-supervised learning*, 2020. arXiv: 1908.02983 [cs.CV].

[43] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, Aug. 2017, ISSN: 2095-5138. DOI: 10.1093/nsr/nwx106. eprint: https://academic.oup.com/nsr/article-pdf/5/1/44/31567770/nwx106.pdf. [Online]. Available: https://doi.org/10.1093/nsr/nwx106.

[44] A. Radford, J. W. Kim, C. Hallacy, *et al.*, *Learning transferable visual models from natural language supervision*, 2021. arXiv: 2103.00020 [cs.CV].

[45] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, *On calibration of modern neural networks*, 2017. arXiv: 1706.04599 [cs.LG].

[46] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," New York, NY, USA: Association for Computing Machinery, 2015, ISBN: 9781450336642. [Online]. Available: https://doi.org/10.1145/2783258.2788613.

[47] M. Bojarski, D. D. Testa, D. Dworakowski, *et al.*, *End to end learning for self-driving cars*, 2016. arXiv: 1604.07316 [cs.CV].

[48] R. M. Neal, *Bayesian Learning for Neural Networks*. Berlin, Heidelberg: Springer-Verlag, 1996, ISBN: 0387947248.

[49] Y. Gal and Z. Ghahramani, *Dropout as a bayesian approximation: Representing model uncertainty in deep learning*, 2016. arXiv: 1506.02142 [stat.ML].

[50] A. Kendall and Y. Gal, *What uncertainties do we need in bayesian deep learning for computer vision?* 2017. arXiv: 1703.04977 [cs.CV].

[51] S. Fort, H. Hu, and B. Lakshminarayanan, *Deep ensembles: A loss landscape perspective*, 2020. arXiv: 1912.02757 [stat.ML].

[52] J. Ho, A. Jain, and P. Abbeel, *Denoising diffusion probabilistic models*, 2020. arXiv: 2006.11239 [cs.LG].

[53] J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin, *Diffusion models for implicit image segmentation ensembles*, 2021. arXiv: 2112.03145 [cs.CV].

[54] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," 1999. [Online]. Available: https://api.semanticscholar.org/CorpusID:56563878.

[55] Z. Ding, X. Han, P. Liu, and M. Niethammer, *Local temperature scaling for probability calibration*, 2021. arXiv: 2008.05105 [cs.CV].

[56] "Plataforma solar de almería." (2024), [Online]. Available: http://www.psa.es/es/gen/objetivos.php (visited on 05/05/2024).

[57] K. Widener and C. Long, *All sky imager*, 2004. [Online]. Available: https://www.freepatentsonline.com/y2004/0169770.html.

[58] M. Hasenbalg, P. Kuhn, S. Wilbert, B. Nouri, and A. Kazantzidis, "Benchmarking of six cloud segmentation algorithms for ground-based all-sky imagers," *Solar Energy*, vol. 201, pp. 596–614, 2020, ISSN: 0038-092X. DOI: https://doi.org/10.1016/j.solener.2020.02.042. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038092X2030147X.

[59] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, *Microsoft coco: Common objects in context*, 2015. arXiv: 1405.0312 [cs.CV].

[60] B. Nouri, S. Wilbert, P. Kuhn, *et al.*, "Real-time uncertainty specification of all sky imager derived irradiance nowcasts," *Remote Sensing*, vol. 11, no. 9, 2019, ISSN: 2072-4292. DOI: 10.3390/rs11091059. [Online]. Available: https://www.mdpi.com/2072-4292/11/9/1059.

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

[62] L. Lu, M. Yin, L. Fu, and F. Yang, "Uncertainty-aware pseudo-label and consistency for semi-supervised medical image segmentation," *Biomed. Signal Process. Control.*, vol. 79, p. 104 203, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:252474564.

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

[64] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, *Cutmix: Regularization strategy to train strong classifiers with localizable features*, 2019. arXiv: 1905.04899 [cs.CV].

[65] A. Paszke, S. Gross, F. Massa, *et al.*, *PyTorch: An imperative style, high-performance deep learning library*, 2019. arXiv: 1912.01703 [cs.LG]. [Online]. Available: http://arxiv.org/abs/1912.01703.

[66] S. Marcel and D. Rodriguez, *Torchvision the machine-vision package of Torch*, In Proceedings of the 18th ACM international conference on Multimedia, MM '10, 2010. DOI: 10.1145/1873951.1874254.

[67] T. P. L. team, *Torchmetrics: Machine learning metrics for distributed, scalable pytorch applications*, 2022. [Online]. Available: https://github.com/Lightning-AI/metrics.

[68] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[69] L. N. Smith, "Super-convergence: Very fast training of neural networks using large learning rates," *arXiv preprint arXiv:1708.07120*, 2018.

[70] P. Goyal, P. Dollár, R. Girshick, *et al.*, "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[71] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, *Deep clustering for unsupervised learning of visual features*, 2019. arXiv: 1807.05520 [cs.CV].

[72] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[73] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[74] J. Wu, J. Chen, and D. Huang, *Entropy-based active learning for object detection with progressive diversity constraint*, 2022. arXiv: 2204.07965 [cs.CV].

[75] S. Mittal, J. Niemeijer, J. P. Schäfer, and T. Brox, *Best practices in active learning for semantic segmentation*, 2023. arXiv: 2302.04075 [cs.CV].

[76] H. Wang, Q. S. Jin, S. Li, S. Liu, M. Wang, and Z. Song, "A comprehensive survey on deep active learning and its applications in medical image analysis," *ArXiv*, vol. abs/2310.14230, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264426546.

[77] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan, *Learning features by watching objects move*, 2017. arXiv: 1612.06370 [cs.CV].

[78] T. Zhou, F. Porikli, D. J. Crandall, L. Van Gool, and W. Wang, "A survey on deep learning technique for video segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7099–7122, 2023. DOI: 10.1109/TPAMI.2022.3225573.

[79] M. Gao, F. Zheng, J. J. Yu, C. Shan, G. Ding, and J. Han, "Deep learning for video object segmentation: A review," *Artificial Intelligence Review*, vol. 56, Apr. 2022. DOI: 10.1007/s10462-022-10176-7.

List of Figures

List of Tables