OPTIMIZING SEMANTIC CLOUD SEGMENTATION VIA VIDEO OBJECT SEGMENTATION FOR SOLAR FORECASTING

Summary

Solar irradiance is influenced by atmospheric factors such as aerosols and clouds. Among these, clouds are the primary driver of intra-hour variability in solar irradiance. Accurate nowcasting of solar irradiance allows to optimize solar field production and its integration with the electricity grid. The use of machine learning (ML) techniques for the detection and classification of clouds in all-sky images is promising, but it is not without challenges. A key limitation is the strong dependency on large volumes of high-quality, pixel-level annotated data for supervised training. This study addresses the challenge of limited annotated data for cloud detection using ML. By applying video object segmentation (VOS), manually annotated images are extended with the addition of automatic annotations across image sequences. The quality of these synthetic masks was assessed by training multiple Convolutional Neural Networks (CNNs), each using a different propagation interval for the initial manually annotated image. This enables systematic comparison of interval effects on accuracy and robustness. Results demonstrate that integrating VOS-generated data enhances model accuracy, offering a solution for training under data scarcity.

Keywords: cloud detection, solar irradiance, nowcasting, machine learning, video object segmentation, convolutional neural network, annotated data, all-sky images.

Field of application for the findings: Solar Resources and Energy Meteorology, Solar Forecasting, Remote Sensing.

Addressed audience: Researchers in solar energy, machine learning practitioners, renewable energy engineers, grid operators

1. Aim and approach

DLR created a manually annotated dataset (Fabel et al., 2022) designed to train semantic cloud segmentation models, comprising 770 pixel-level labeled sky images. The dataset includes different cloud layers (low, medium, and high), according to the World Meteorological Organization (WMO) classification. The dataset contains a wide range of scenarios with different cloud formations, sun elevation angles, and atmospheric turbidity levels. However, this dataset is relatively small. As the training of machine learning (ML) models typically requires larger datasets we propose to automatically create annotated images. Video object segmentation (VOS) is a technique that segments and tracks objects frame by frame in a video. In Semi-Supervised VOS (SVOS) a mask of the object to be segmented is given for one frame and the model propagates it throughout the following frames. In this case, for each of the manually annotated images, a sequence is created using existing consecutive images, hereafter referred to as frames. The sampling time between the frames is 30 s. The propagation of the mask with SVOS was done for 10 frames, that corresponds to the 10 previous and 10 subsequent images of each manually annotated image. The idea of using previous and subsequent images is to maximize the use of each manually annotated image. The proposal of using 10 frames was made heuristically (Gut, 2025).

The “Cutie” model (Cheng et al. 2024), which integrates memory-based and query-based approaches to improve segmentation accuracy, was selected as the final SVOS model for this study. We fine-tuned the pre-trained Cutie model on a weakly labeled dataset (Gut, 2025), which enhanced its performance by improving temporal coherence and contextual understanding. This weakly labeled dataset, generated in a previous work (Magiera et al., 2025), combines binary cloud segmentation with ceilometer cloud height measurements and cloud state heuristics. It consists of 948 distinct image sequences, totaling 47,000 images. Fine-tuning the Cutie model using domain-specific, weakly labelled data improves the quality of the SVOS model for the sky camera-specific task, thereby improving segmentation accuracy.

The benefit of the annotated dataset generated with SVOS is evaluated by training and benchmarking a set of semantic cloud segmentation models, i.e. convolutional neural networks (CNNs). Figure 1 shows an overview of the proposed approach. In order to evaluate the mask propagation up to a given frame, hereafter referred to as step, a model is trained using all images generated up to that frame. This process is repeated for each step resulting in one model per step, allowing the comparison of mask propagation. Furthermore, the performance of CNNs was studied when they were pre-trained using the previously mentioned weakly labeled dataset (Magiera et al., 2025).

The dataset for training the semantic cloud segmentation models, consisting of the 770 manually annotated images, was divided in three subsets: training, validation and test in a 60:20:20 ratio. The data was evenly distributed to have all different classes represented in all subsets (Gut, 2025). Only manually annotated images were used for the validation and test subsets to avoid any potential bias from inaccuracies in the SVOS-generated annotations. The training subset, on the other hand, includes all remaining manually annotated images, along with their corresponding automatically annotated images.

The performance of the trained segmentation models is evaluated using the Intersection over Union (IoU) metric (eq. 1), which is widely adopted in computer vision for assessing semantic segmentation tasks.

$$\text{mean IoU} = \frac{1}{|\text{Classes}|}\sum_{c\,\in\,\text{Classes}}\frac{\text{Target}_c \cap \text{Prediction}_c}{\text{Target}_c \cup \text{Prediction}_c} = \frac{1}{|\text{Classes}|}\sum_{c\,\in\,\text{Classes}}\frac{\text{TP}_c}{\text{TP}_c \cup \text{FP}_c \cup \text{FN}_c} \qquad \text{(eq. 1)}$$

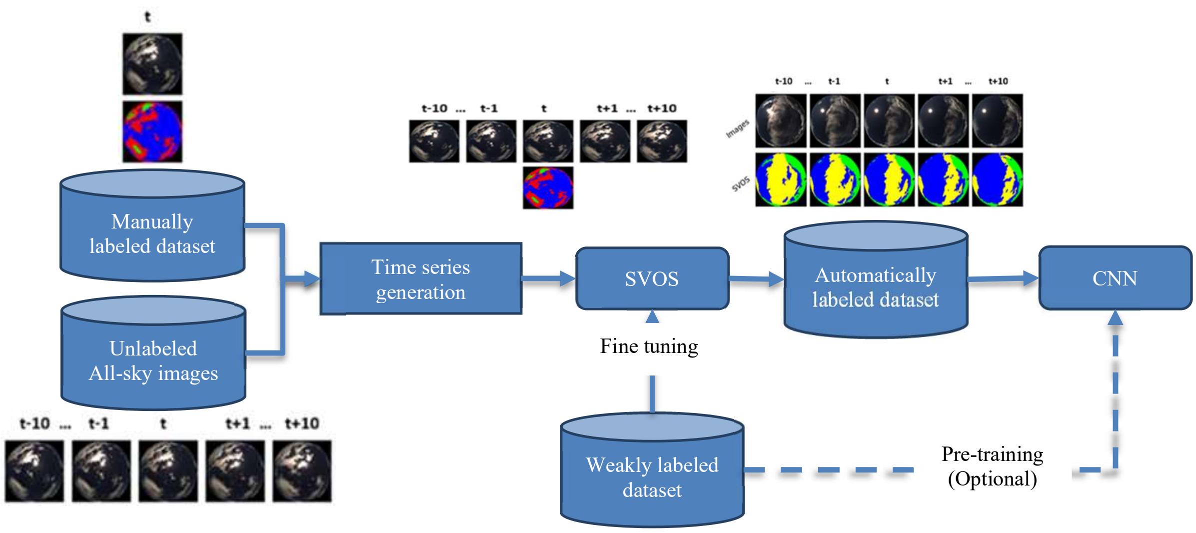

Fig. 1: Overview of the proposed approach.

2. Scientific innovation and relevance

The primary scientific innovation of this work is the use of SVOS, which has shown to be a promising solution for segmentation tasks with limited labeled data, achieving significant improvements. These innovations enable our semantic cloud segmentation model to improve by 6.4% against the baseline.

Finally, this work also lays the foundation for future research like the use of time series information, the incorporation of confidence weighting (e.g. decreasing weights for label propagation) or the relaxation of labels to smooth class boundaries of propagated targets.

3. Results and conclusions

Figure 2 shows the resulting IoU obtained from evaluating the set of trained CNNs on the test subset, with and without pre-training using weakly labeled dataset (Magiera et al., 2025). When no pre-train was done results show an improvement of 3.4% against the baseline for propagation step 5 (IoU of 0.771). The baseline corresponds to the scenario in which the CNN is trained exclusively on manually annotated data, i.e., without any propagation steps (propagation step 0). Furthermore, this number is surpassed by the pre-trained CNNs that show substantial performance gains for all steps, having the optimal result when propagating initial mask up to step 8 (IoU of 0.794), corresponding to an improvement of 6.4% against the baseline. From that point on, performance does not improve and we can see how it begins to degrade.

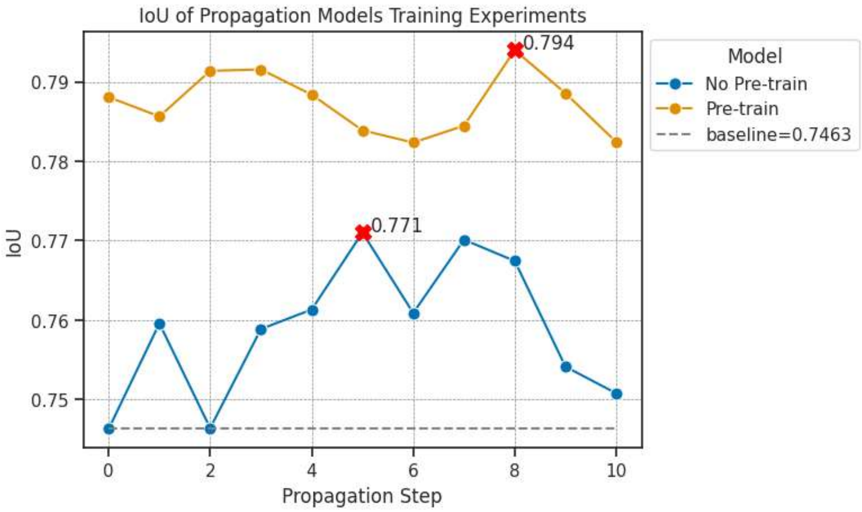

Fig. 2: IoU of propagation models training experiments

The full contribution will provide a detailed description of the entire workflow, including fine-tuning the SVOS model and pre-training the semantic cloud segmentation model using the weakly labeled dataset. It will also present a comprehensive analysis comparing the performance of different SVOS models, as well as exploring potential improvements achieved through advanced ground truth generated from infrared images.

4. References

Cheng, H. K., Oh, S. W., Price, B., Lee, J. Y., & Schwing, A. 2024. Putting the object back into video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3151-3161).

Fabel, Y., Nouri, B., Wilbert, S., Blum, N., Triebel, R., Hasenbalg, M., ... & Pitz-Paal, R. 2022. Applying self-supervised learning for semantic cloud segmentation of all-sky images. *Atmospheric Measurement Techniques*, *15*(3), 797-809.

Gut, R. 2025. *Development of advanced cloud classification and segmentation models for solar energy applications using deep learning and synthetic data augmentation* (Master Thesis, Hochschule Kempten).

Magiera, D., Fabel, Y., Nouri, B., Blum, N., Schnaus, D., & Zarzalejo, L. F. 2025. Advancing semantic cloud segmentation in all-sky images: A semi-supervised learning approach with ceilometer-driven weak labels. *Solar Energy*, *300*, 113822.