Solar Energy 311 (2026) 114515

Contents lists available at ScienceDirect

Solar Energy

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

Multi-frame cloud prediction in all-sky images from RGB images and segmented masks

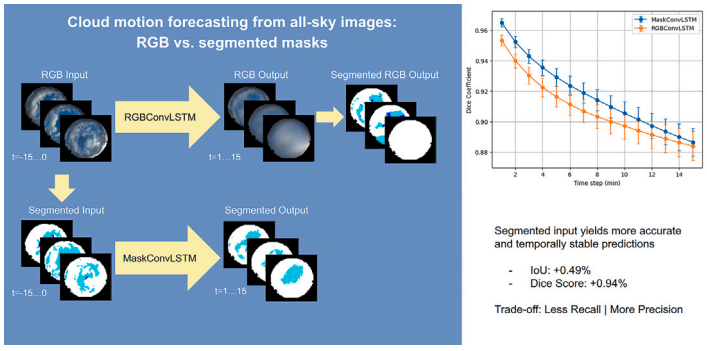
Javier Gatón[a, b,*], Roberto Román[a, b], Cesar Guzman[c], Daniel González-Fernández[a, b], Bruno Longarela[a, b], Carlos Toledano[a, b], Ramiro González[a, b]

[a] Grupo de Óptica Atmosférica (GOA-UVa), Universidad de Valladolid, Paseo de Belén 7, Valladolid, 47011, Spain
[b] Laboratory of Disruptive Interdisciplinary Science (LaDIS), Universidad de Valladolid, Paseo de Belén 7, Valladolid, 47011, Spain
[c] DriMT AI Space, Valga, 68204, Estonia

HIGHLIGHTS

- RGB and semantic mask inputs compared for short-term all-sky cloud prediction.
- Evaluation framework proposed to compare both input representations.
- Semantic mask inputs improve semantic label space agreement.
- Average gains of 0.49% IoU and 0.94% Dice score observed.
- Results show a trade-off between recall reduction and precision improvement.

GRAPHICAL ABSTRACT



ARTICLE INFO

2000 MSC:
68T45
68U10
86A10

PACS:
07.05.Pj
07.05.Mh
92.60.Jq
42.68.-w
42.68.Ge

Keywords:
All-sky images
Clouds
Cloud motion prediction
Artificial intelligence
Semantic segmentation
Next-frame prediction

ABSTRACT

This paper presents a comparative study on the impact of input representation on deterministic artificial intelligence models for short-term multi-frame prediction in all-sky images. This work compares a model operating on 8-bit RGB all-sky images with a model that shares the same backbone, but operates directly on semantically segmented masks that encode cloud-related classes. Using an available sky segmentation model, predictions are evaluated in the segmentation label space using segmenter-derived masks as a proxy reference. Within this evaluation framework, the use of semantic masks as input for short-term prediction leads to improved temporal stability and higher agreement across standard segmentation metrics such as intersection over union, Dice coefficient, and categorical cross-entropy. While these results suggest potential relevance for weather and solar energy nowcasting applications, further validation against physical irradiance measurements is required.

* Corresponding author at: Grupo de Óptica Atmosférica (GOA-UVa), Universidad de Valladolid, Paseo de Belén 7, Valladolid, 47011, Spain.
Email address: gaton@goa.uva.es (J. Gatón).

https://doi.org/10.1016/j.solener.2026.114515
Received 15 September 2025; Received in revised form 21 January 2026; Accepted 8 March 2026
Available online 18 March 2026

Nomenclature

RGB red, green and blue
LSTM long short-term memory
RNN recurrent neural network
GOA-UVa Atmospheric Optics Group - *Grupo de Óptica Atmosférica* - of the University of Valladolid
MSE Mean Squared Error
IoU Intersection over Union
CCE Categorical Cross-Entropy

1. Introduction

Solar energy has become one of the most important renewable sources for electricity production. It possesses the potential to reduce the dependency on fossil fuels due to its reduced cost, mitigating the effects of climate change [1]. Solar energy is based on converting the incoming solar radiation, which is multi-spectral, encompassing a wavelength range between 280 and 3000 nm. The amount of solar radiation reaching a definite Earth's surface point mainly depends on two factors: 1) astronomical factors, defined by the geographic coordinates, date and time; and 2) atmospheric and surface factors such as gases, aerosols, clouds and surface albedo. Solar radiation at Earth's surface is formed by two components: 1) the direct solar radiation, or beam radiation, which is the solar radiation that hasn't been scattered by the atmosphere; and 2) the diffuse radiation, which is the solar radiation that reaches Earth's surface after its direction is changed by atmospheric scattering [2,3]. Under cloud-free conditions the direct component constitutes the dominant fraction of the total solar radiation, while under overcast conditions the diffuse radiation becomes the main contribution [4]. The relative proportion of direct and diffuse radiation plays a critical role in solar energy production [5].

Clouds, formed by water droplets and ice crystals in suspension, are the main atmospheric factor modulating the incoming solar radiation, especially the direct component. This limits the amount of solar radiation that reaches the Earth's surface, and therefore the energy production of the photovoltaic systems. Therefore, predicting the position and evolution of clouds plays a key role in solar nowcasting, which refers to the short-term forecasting of solar irradiance [6]. Accurate cloud position estimation over time horizons of minutes to a few hours allows for better management and optimization of solar energy production, significantly reducing the associated uncertainties [7]. However, cloud formation and movement depend on a wide range of atmospheric factors such as wind direction and speed, vertical convection, humidity or temperature gradients. These variables are difficult to model or observe with high accuracy and sufficient spatial and temporal resolution, resulting in chaotic behavior and making short-term prediction particularly challenging [8].

Almost all state-of-the-art techniques for predicting cloud position rely on all-sky imagery [9], captured by all-sky cameras or similar devices. These specialized instruments generally use fisheye lenses to capture the entire sky dome in a single frame, enabling real-time tracking of cloud dynamics. Current neural network-based methods can be categorized according to how they model spatiotemporal dynamics into three categories: 1) deterministic video prediction networks like TrajGRU [10], U-Net-based networks [11] or ConvLSTM [12] and derived networks like PredRNN [13]; 2) probabilistic and generative video prediction networks like VideoGPT [14] or GAN-based networks [15,16]; and 3) networks that combine neural networks with physical knowledge like PhyDNet [17] or SkyGPT [18].

While these recent works leverage advanced architectures or physics-informed constraints, they all rely on 8-bit red, green and blue (RGB) images as input. None, to our knowledge, investigate the effect of using semantically segmented masks as a structured input representation, where cloud location and type are encoded explicitly, reducing the number of unknowns to calculate. The idea of using masks with information on the location of the clouds and their kind as the input and the output of the model, instead of images, is new. Predicting segmented futures has shown substantially better results than segmenting predicted RGB frames in other domains [19]. However, its impact on cloud position nowcasting has not been explored and remains unclear. This motivates a controlled comparison between both kinds of input representations.

In this framework, the main objective of this work is to investigate the impact of input representation on short-term cloud motion prediction. To assess this, we compare the performance of models operating on RGB all-sky images with models operating directly on semantically segmented masks. This is done using a sky segmentation model. To enable a controlled comparison, model performance is evaluated in the segmentation label space, using segmenter-derived masks as a proxy reference.

The main contributions of this work are:

- A controlled comparison between RGB-based and mask-based representations for short-term all-sky sequence prediction using a common ConvLSTM backbone.
- An evaluation framework that enables comparisons in a shared semantic label space.
- An analysis of temporal robustness and per-class behavior, highlighting both advantages and limitations of operating in a discrete semantic space.

Due to its simplicity ConvLSTM has been chosen as a representative deterministic video prediction architecture to serve as a common backbone on which to perform this comparison. The conclusions of this work are to be applied within the class of deterministic video prediction networks. Exploring this with other network families and approaches is a direction for future work.

The scope of this study is subject to several limitations. Model performance is evaluated using segmenter-derived masks as a proxy reference and is therefore conditioned on the behavior of the adopted sky segmentation model. Furthermore, the relatively low spatial resolution of the input data (64×64) may hinder the representation of fine-grained cloud structures, such as thin clouds or sharp boundaries. The reported results should be interpreted as reflecting representation effects within this evaluation framework rather than as direct validation against physical irradiance measurements.

2. Materials and methods

2.1. Data collection and preparation

The SKIPP'D [20] dataset, which contains 57,803 videos in *.hdf5* format, has been used as the all-sky image source for this study. This dataset includes all-sky videos collected under various weather conditions and from multiple geographic locations, rendering it highly generalizable and well-suited for robust model training and evaluation. Each video comprises a sequence of 31 8-bit RGB frames with a resolution of 64×64 pixels. Each frame represents a sky image with a temporal difference of one minute from the previous and the next one in the sequence.

This dataset comes with a predefined split into two subsets: a training subset of 53,336 videos, and a test subset of 4467 videos. For model development, the training subset has been further split into training and validation sets using a random 90% to 10% partition, respectively. Each video is also divided into two subgroups: the "log" subgroup which

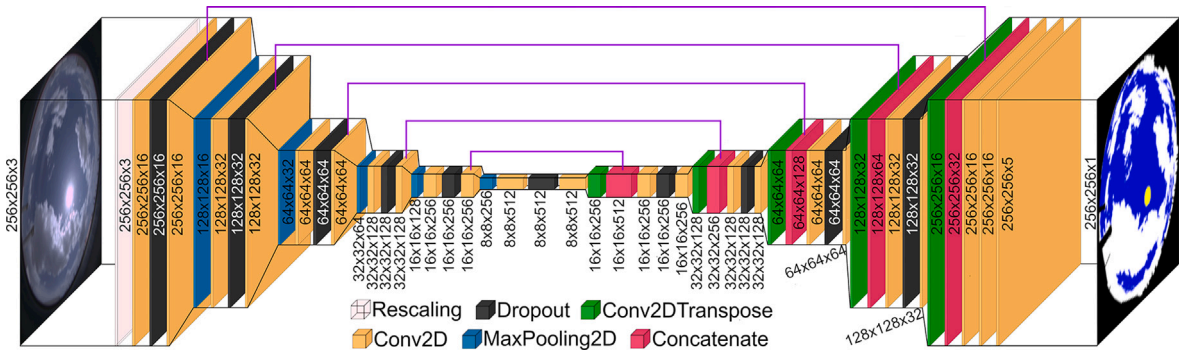


Fig. 1. Architecture of the GOA-UVa all-sky segmentation U-Net model.

Fig. 2. (a) An RGB all-sky image from the SKIPP'D dataset. (b) Its semantically segmented mask obtained with the GOA-UVa all-sky segmentation U-Net model. The classes of the segmented image are: cloud-free, cloud, thin cloud, sun and other, and they are represented with cyan, white, dark blue, yellow and black respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

consists of the first 16 frames, spanning from minute $t = -15$ to minute $t = 0$, and the "pred" subgroup which encompasses the latter 15 frames, from minute $t = +1$ to $t = +15$. The "log" subgroup is intended to be the model input, while the "pred" subgroup serves as the prediction target.

The images have been preprocessed before using them. All the images have a noisy black frame surrounding the sky image. That frame has been homogenized, by applying a manually created black mask over all the images.

2.2. Ancillary segmentation model

A semantic segmentation model, developed by the Atmospheric Optics Group - *Grupo de Óptica Atmosférica* - of the University of Valladolid (GOA-UVa), has been used to obtain a segmented sky mask for each one of the images in the SKIPP'D dataset and from the images generated by the all-sky images model down the line. This segmentation model, the GOA-UVa all-sky segmentation U-Net model, is available on Zenodo [21], and it automatically classifies each pixel into five predefined sky condition classes: cloud-free, cloud, thin cloud, sun, and other.

2.2.1. Architecture

The GOA-UVa all-sky segmentation U-Net model follows a U-Net architecture, a well established convolutional neural network designed for semantic segmentation [22]. The model operates on all-sky RGB images with a resolution of 256 × 256 pixels. Fig. 1 shows a detailed scheme of the architecture of the GOA-UVa all-sky segmentation U-Net model.

2.2.2. Semantic classes and label interpretation

The segmentation model classifies each pixel into five sky categories: cloud, cloud-free, sun, thin cloud and other. Cloud corresponds to opaque cloud formations, while cloud-free represents clear sky pixels. Then, sun designates the solar disk when it is not covered by any kind of cloud structure. The thin cloud category is more complex as it does not exclusively correspond to thin cloud formations. It encompasses clouds like thin cirrus but also visually ambiguous regions that lie between the cloud-free and cloud categories, including class boundaries and low opacity structures. It can be interpreted as a low-confidence cloud. This class represents intrinsic semantic ambiguity and uncertainty, mainly defined by radiometric attenuation rather than well-defined spatial structures. Finally, other signifies any kind of element unrelated to sky conditions, like buildings, landscape elements or camera borders. An instance of this segmentation can be seen in Fig. 2.

2.2.3. Segmentation model performance

The GOA-UVa all-sky segmentation U-Net model has been used to generate the segmented sky masks for the ground-truth all-sky images. These masks are used as a proxy for physical cloud observations in this work, and it is essential to determine how far this pseudo ground-truth is from the actual one.

For this purpose an independent small test set of 48 annotated all-sky images has been put together, obtained through GOA-UVa's all-sky cameras. This set has been fed to the model, and the model's output has been compared against the ground truth sky masks. The results can be seen in Fig. 3, showing both confusion matrices for recall and precision. Recall, also called sensitivity, is the number of correct results, or true positives, relative to the number of expected positive results; and precision, also known as positive predictive value, is the number of correct results relative to the number of all results reported as positive by the predictor [23]. These matrices show that the model predicts the other, cloud-free and cloud classes with recall and precision values always higher than 87% and 75%, respectively. The predictions are less aligned for the sun class, with a recall of 63%, mistakenly predicting it as cloud-free 16% of the time, or as cloud and thin cloud 11% of the time for both classes.

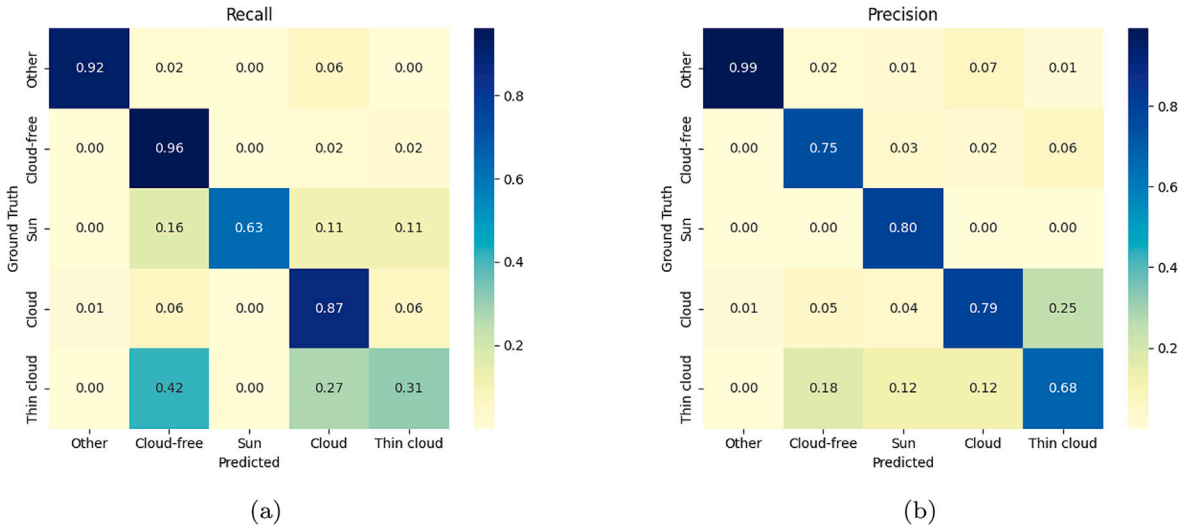


Fig. 3. (a) Row-normalized confusion matrix of the segmentation model showing the distribution of predicted classes given the ground truth labels, corresponding to per-class recall. (b) Column-normalized confusion matrix of ground-truth labels given the predicted classes, corresponding to per-class precision.

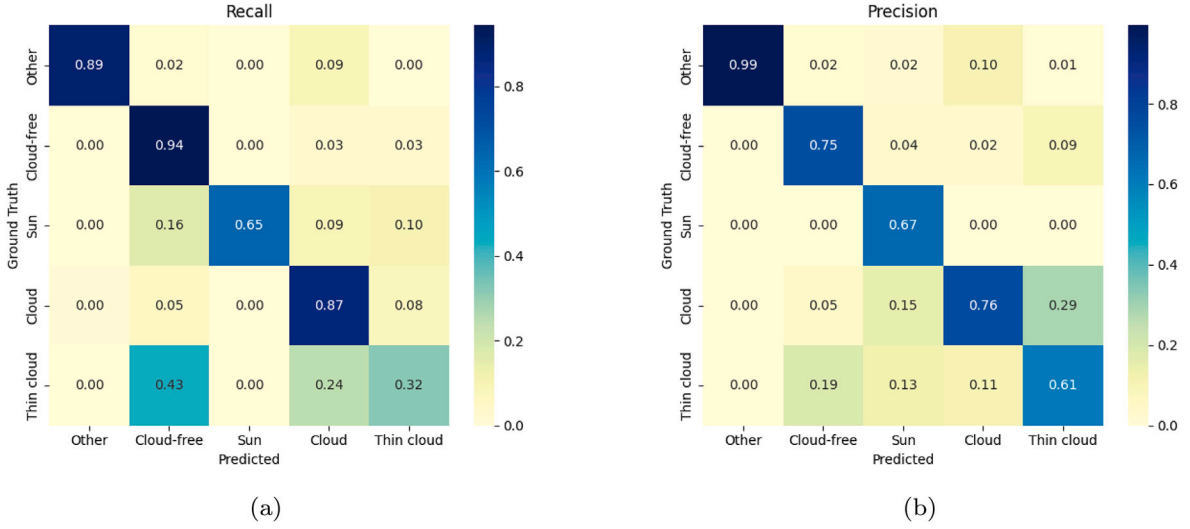


Fig. 4. (a) Row-normalized confusion matrix of the segmentation model evaluated on images that were downsampled to 64×64 resolution and then upsampled to 256×256, showing the distribution of predicted classes given the ground truth labels, corresponding to per-class recall. (b) Column-normalized confusion matrix of ground-truth labels given the predicted classes for the same images, corresponding to per-class precision.

Nevertheless, the results show a precision of 80% for the sun class, indicating a low rate of false positives, which mainly are actual thin cloud pixels (13%). The sun disk is a small high-contrast element that can be sensitive to small spatial errors in its segmented contour, especially when it is surrounded or partly covered by clouds, or when it is near the horizon. Finally, the predictions greatly underperform for the thin cloud class, with a recall of 31%. As mentioned, the thin cloud class represents a visually ambiguous category, not defined by clear spatial patterns. Its labeling involves on multiple occasions gradual radiometric differences without discrete boundaries. This results in a relatively high degree of subjectivity, limiting the achievable agreement. It is semantically located at the boundary of what the human eye would classify as cloud and cloud-free. This is shown in the results, where the false negatives are split between those two semantic categories (27% and 42%).

2.2.4. *Impact of spatial downsampling on model performance*

The used segmentation model operates on all-sky images with 256 × 256 pixel resolution, higher than the 64 × 64 pixel resolution of the SKIPP'D dataset. Therefore, in order to use SKIPP'D data with the GOA-UVa all-sky segmentation U-Net model, the images must be resized to match the input dimensions of the segmentation model, and the resulting output masks must then be upsampled back to the original SKIPP'D resolution. This hinders the model's ability to perceive fine cloud structures and involves the creation of interpolation artifacts in the images. This can negatively affect the model's effective performance in this work and therefore its impact should be assessed. To do so, we have downsampled the images of the same test set used in Section 2.2.3 to a 64 × 64 resolution and then upsampled them to the original 256 × 256 one to compare them against the original annotated masks. The downsampling has been performed through area-based interpolation, and the upsampling through bilinear interpolation, as that is the same interpolation used to upsample the SKIPP'D dataset in this work.

The results are shown in Fig. 4, indicating a slight decrease in precision. This decrease mainly affects the sun class, going from a precision of 80% to 67%. This translates into the model more frequently misclassifying cloud or thin cloud pixels as sun ones. This is probably due

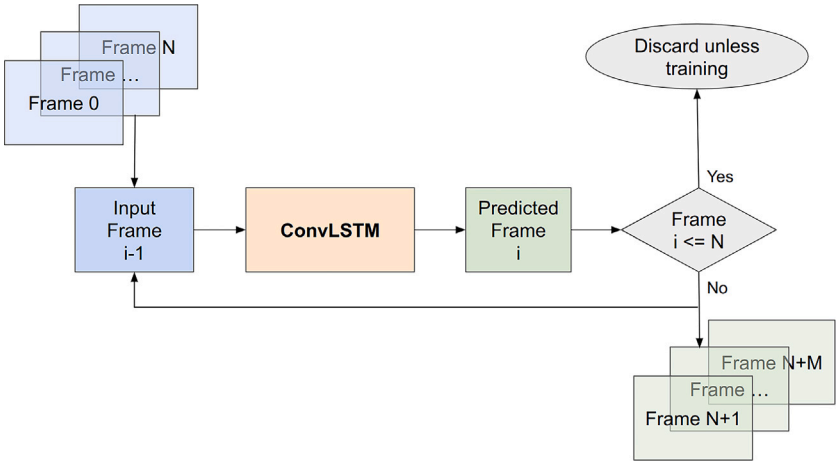


Fig. 5. Workflow of the ConvLSTM networks used in this work. $N+1$ input frames are fed to the network one at a time, extracting their information into its hidden states. Then, the last generated frame is fed as new input until M output frames are generated.

to the loss of small cloud-related features that indicate that the sun is covered by clouds. Other categories are not significantly affected, although overall precision is slightly reduced.

2.3. Base architecture for short-term multi-frame prediction

The prediction models used in this work are based on a common neural network backbone belonging to the Convolutional Long Short-Term Memory (ConvLSTM) family. ConvLSTM networks are well established in the state of the art, and they have been employed in the task of predicting cloud movement multiple times [18,24]. This kind of network has been chosen in this study due to the relative simplicity of its architecture and its training process, which is especially important when adapting it to different input representations. In this work, two models based on the same ConvLSTM backbone are considered: one operating on all-sky RGB images, and another operating on semantically segmented masks. For clarity, these configurations are referred to as RGBConvLSTM (RGB-based) and MaskConvLSTM (mask-based), respectively.

ConvLSTM networks are based on long short-term memory (LSTM) networks [25], which are a type of recurrent neural network (RNN) composed of LSTM cells. These cells maintain an internal hidden state, which is updated at each time step based on the current input and the previous hidden state. The output of the LSTM cell is generated by combining both elements, allowing the network to retain temporal information across sequences. LSTM cells combined with convolutional operations are what shape the ConvLSTM cells. These ConvLSTM cells are typically organized into two-dimensional layers, known as ConvLSTM2D layers, where the convolutional operations are applied across the spatial dimensions of the input (i.e., height and width). This structure makes them particularly suitable for processing sequences of images, such as video frames or time series of sky images, where both spatial and temporal patterns must be learned simultaneously. Both networks used in this work (RGBConvLSTM and MaskConvLSTM networks) share a similar underlying structure based on ConvLSTM2D layers.

ConvLSTM networks can be applied with different workflows. In this work, both RGBConvLSTM and MaskConvLSTM networks use the same approach for cloud motion forecasting. Their workflow is described in Fig. 5. Internally ConvLSTM networks only predict one frame at a time. To generate videos, the available input frames are fed to the network, producing one output image for each of them. During training, these output images will be used to improve the network, while in production these images will be discarded. It's important to feed the network with all the input data even if the output is going to be discarded, because the network needs to extract all the information and store it inside its inner hidden states. Once all the available frames have been used, the network starts to receive the last predicted frame as the input, creating new frames in this way from that point onward, which will form the predicted video.

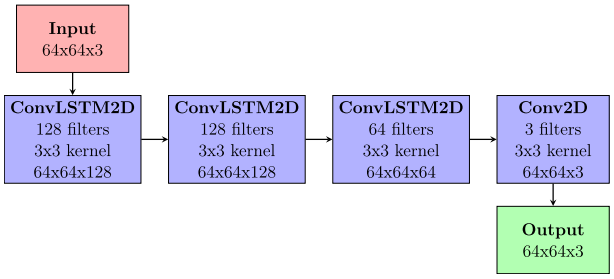


Fig. 6. RGBConvLSTM model's structure.

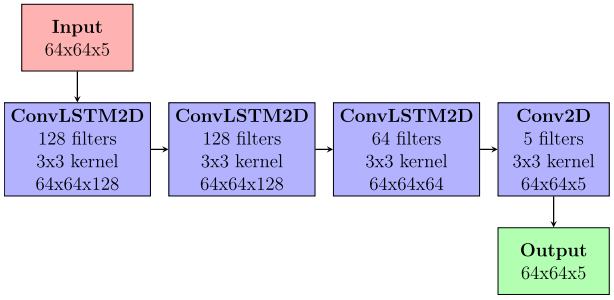


Fig. 7. MaskConvLSTM model's structure.

2.4. Network design and model training

As mentioned, two model instances are implemented in this study, both based on the same ConvLSTM network architecture. The first configuration, RGBConvLSTM model, processes sequences of 64×64 RGB all-sky images. The second configuration, MaskConvLSTM model, processes sequences of 64×64 semantically segmented masks representing multiple cloud-related classes. Although both configurations share the same network design, they are trained independently under different input and output representations and therefore result in different sets of learned parameters. The only distinction between both model instances lies in the input and output representations, allowing the effect of the representation choice to be isolated. Both models are implemented using

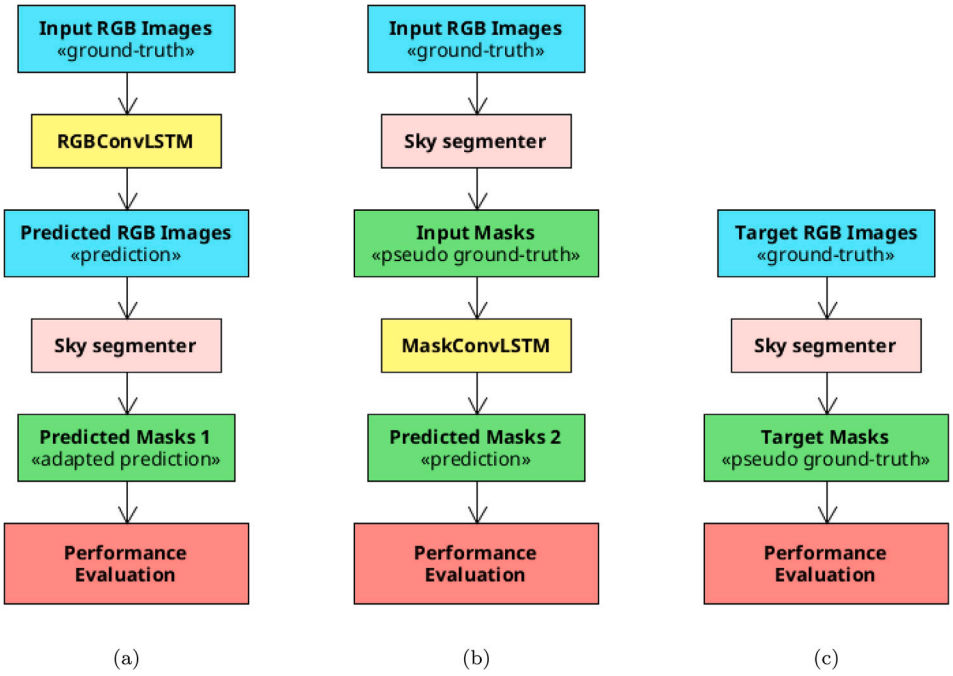


Fig. 8. Methodology used to render both models' outputs comparable in the semantic label space. (a) Ground-truth input RGB images processed through RGBConvLSTM model, whose RGB predictions are segmented obtaining masks. (b) Ground-truth input RGB images segmented and fed to the MaskConvLSTM model obtaining predicted semantic masks. (c) Ground-truth target RGB images segmented to obtain the proxy reference masks for evaluation.

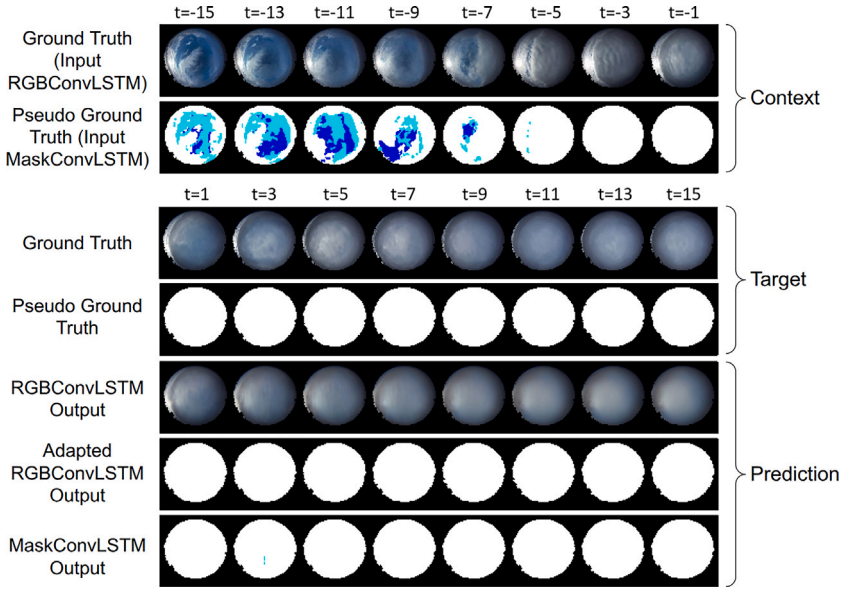


Fig. 9. MaskConvLSTM model and adapted RGBConvLSTM model predictions for Case 1. The first two rows show the last observed frames from $t = -15$ to $t = -1$ minutes, sampled every 2 minutes. Although the frame at $t = 0$ is part of the input context, it is not displayed here. The following two rows present the ground truth future frames from $t = +1$ to $t = +15$ minutes, also sampled every 2 minutes. The last three rows display the models' predictions for the same forecast horizon.

the PyTorch 2.3.1 deep learning framework [26] and trained on a GPU cluster equipped with NVIDIA A100 (40 GB memory) graphics cards.

2.4.1. RGBConvLSTM model

The first model instance uses a classic ConvLSTM architecture. Its structure is shown in detail in Fig. 6 and follows the design commonly adopted in the literature [18]. The design consists of three ConvLSTM2D layers with dimensions 128, 128 and 64, all of them with a kernel size of 3x3 and a stride of 1. After that, there's a convolutional 2D layer (Conv2D) with 64 input channels and 3 output channels, to create the output images.

Following the training setup described in the same work [18], this model has been trained for 50 epochs by minimizing the Mean Squared Error (MSE) loss function. Two optimization techniques have also been adopted accordingly: teacher-forcing [27] and the automatic reduction of learning rate [28]. Teacher-forcing has been done with a teacher-forcing ratio starting at 1 and decreasing by 0.03 per epoch until reaching zero. The automatic reduction of learning rate has been

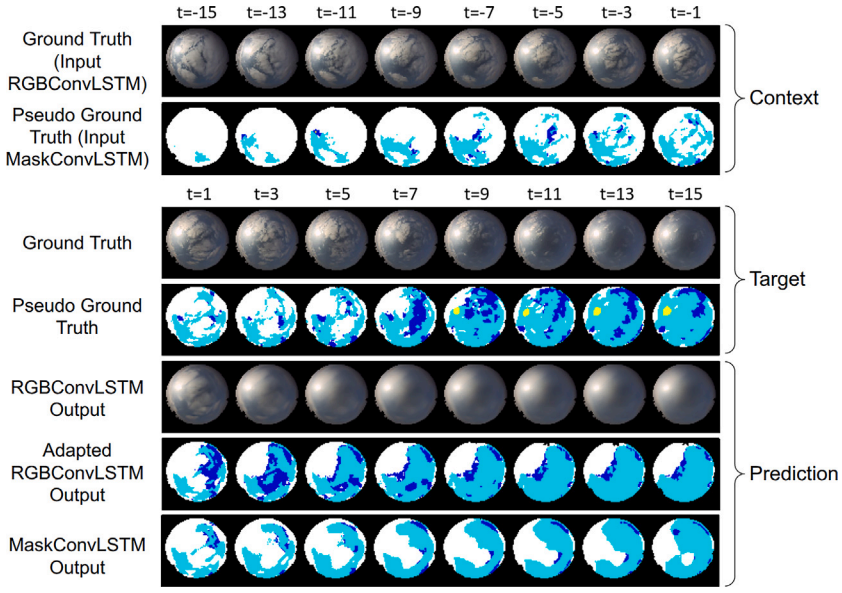


Fig. 10. Similar to Fig. 9 but for Case 2.

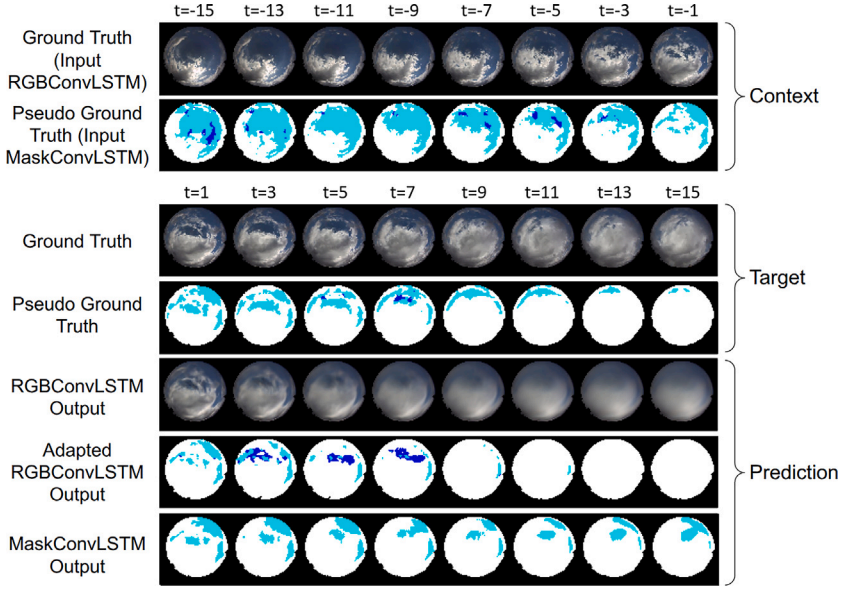


Fig. 11. Similar to Fig. 9 but for Case 3.

performed by waiting for 3 epochs without improvements on the validation subset to reduce the learning rate, with a reduction factor of 0.3, and with 0.001 as the initial learning rate. The final model parameters have been selected from the epoch yielding the best performance on the validation set.

2.4.2. *MaskConvLSTM model*

The second model instance (MaskConvLSTM) proposed in this work corresponds to the same ConvLSTM architecture described in the previous section but adapted to use semantically segmented information. Its structure, shown in Fig. 7, is identical to that of the aforementioned RGBConvLSTM model, with the exception of the input and output representations. Specifically, instead of working with three color channels (RGB images), it operates with five-channel inputs corresponding to the five semantic classes defined by the segmenter model. Consequently, the feedback mechanism explained in Section 2.3, where the latest frame is reused as input, has been implemented using class probability maps obtained via a softmax operation.

Like the RGBConvLSTM model, MaskConvLSTM model has been trained for 50 epochs, using the same optimization techniques and protocol. Categorical Cross-Entropy (CCE) has been used as the optimization objective, as it is naturally suited for multi-class semantic segmentation tasks and aligns with the evaluation metrics employed in this study. This choice is consistent with standard practice in semantic segmentation and allows the model to directly optimize class-wise probability distributions. The use of different loss functions across configurations reflects the different output representations rather than architectural differences, and is aligned with the goal of isolating the impact of input representation. The model corresponding to the epoch with the best performance on the validation set has been selected for evaluation.

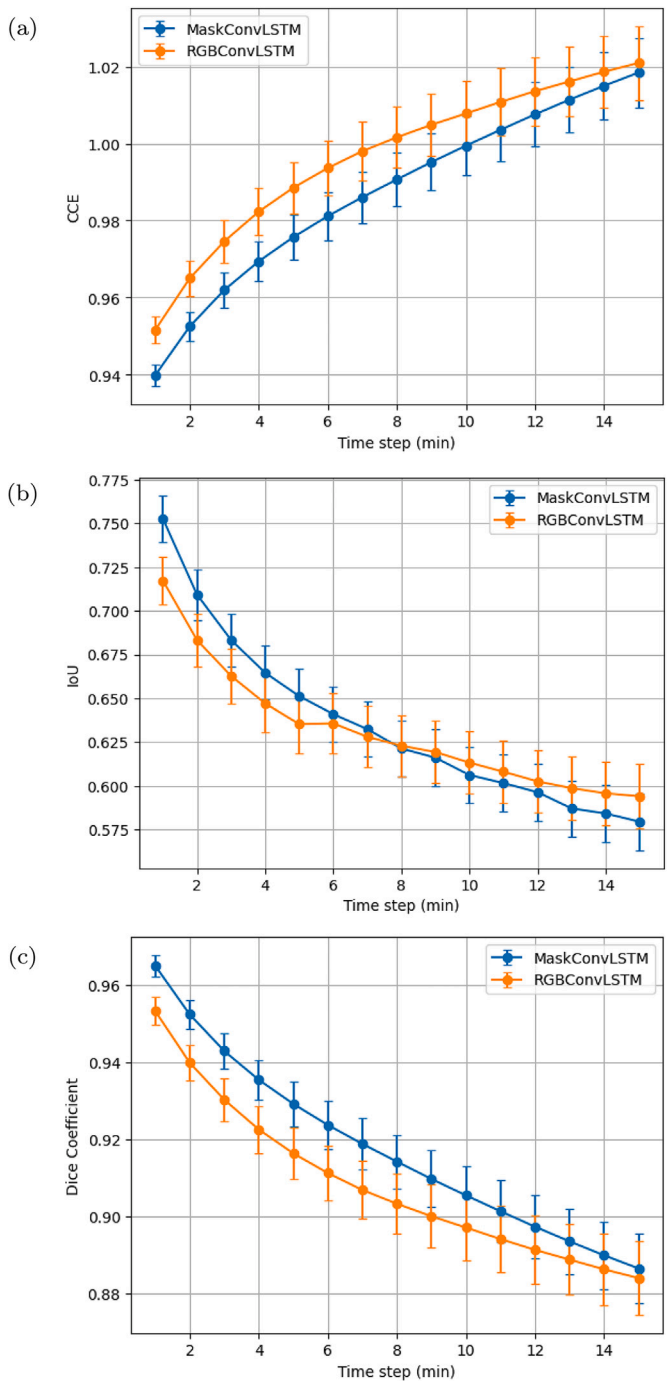


Fig. 12. Segmentation performance of RGBConvLSTM model and MaskConvLSTM model over the 15-minute prediction horizon against the test set using: (a) CCE, (b) IoU and (c) Dice Coefficient. Each curve shows the mean score across all test sequences for a given timestep. Error bars indicate the 95% confidence interval of the mean.

2.5. Comparison framework

RGBConvLSTM model and MaskConvLSTM model operate with different types of data, so it is not possible to carry out a direct performance comparison between them. The performance for predicting all-sky segmented classes has been evaluated in this work for both implemented model instances, following the process shown in Fig. 8. In order to render their outputs comparable, RGBConvLSTM model's predictions have been processed with the GOA-UVa all-sky segmentation U-Net model, mapping them into the same semantic label space as the outputs that MaskConvLSTM model generates. The same segmentation model has also been applied to the ground-truth frames of the test dataset, to generate the reference masks used for evaluation. This is done for the input images because the MaskConvLSTM model needs segmented masks as its input, and for the target images this is needed because in order to assess the performance of MaskConvLSTM model's predictions and RGBConvLSTM model's adapted predictions, one needs to compare them against segmented masks, not full images.

The goal of this comparison framework is not to benchmark an end-to-end cloud nowcasting pipeline against physical ground truth, but to isolate the effect of working in a semantic label space for sequence prediction. Therefore, predictions based on RGB images are mapped into the same label space as the predictions based on semantically segmented masks, using a common segmentation model.

This evaluation and comparison have been carried out using three different metrics: the CCE loss function [29]; the Sørensen-Dice coefficient, commonly referred to as the Dice coefficient [30]; and the Intersection over Union (IoU), also known as the Jaccard Index [31]. These metrics reflect classification accuracy at the pixel level. The performance of the predictions at each time step has been assessed with these metrics.

CCE measures the alignment between the predicted class probability distribution and the reference label. It provides a measure of probabilistic consistency, but it is not directly visually interpretable. In contrast, IoU and the Dice coefficient explicitly quantify the spatial agreement between predicted and reference masks. IoU measures, for each class, the ratio of the area of the intersection of the class pixels in the prediction and reference masks to the area covered by their union. This makes it very sensitive to boundary errors and small spatial displacements. The Dice coefficient works in a similar way, measuring the fraction of relevant pixels that have been correctly identified, being less affected by class imbalances and small boundary errors [32]. Using these three metrics together allows us to assess not only probabilistic correctness (CCE) but also spatial accuracy and robustness to class imbalance (IoU and Dice), which is particularly relevant for all-sky segmentation where cloud and cloud-free pixels dominate over less frequent classes such as sun or thin cloud.

The CCE loss function is usually utilized to evaluate semantic segmentation models, where each pixel is treated as an independent classification problem. CCE has an inverse relation with predictive performance, where the lower the CCE, the higher the performance is. It's widely used as the training's loss function and as the evaluation metric for comparing models' performances (see Sections 6.2 and 6.5 of [33]), but it's not easily interpretable in absolute terms, especially for more than two categories. From its definition, one can infer that random predictions over N classes tend to yield a CCE of $log(N)$. This study has worked with five categories, so $log(5) \approx 1.609$ is a good reference for what both models should at least beat.

The other metrics used, IoU and the Dice coefficient, are better suited for comprehensively analyzing the performance of this sort of models, as they are widely used for evaluating the quality of semantic segmentation tasks [34,35]. The IoU evaluates how well the predicted segmentation coincides with the real segmentation, measuring the intersection over the union of the predicted pixels and the target ones. The Dice coefficient is very similar to the IoU, but it is more tolerant of class boundary mismatches, penalizing small discrepancies less, and gives more weight to true positives, allowing minority classes to have a greater influence on the result. This is useful for unbalanced datasets where some classes dominate over others, which is something that applies to the used dataset, because in most frames there are many more cloud or cloud-free pixels than sun or thin cloud pixels.

This comparison framework is not bound to the dataset and models used in this work. Since the study focuses on input representation rather than on a particular network architecture, its methodology could be applied to any set of sky image sequences and prediction architectures.

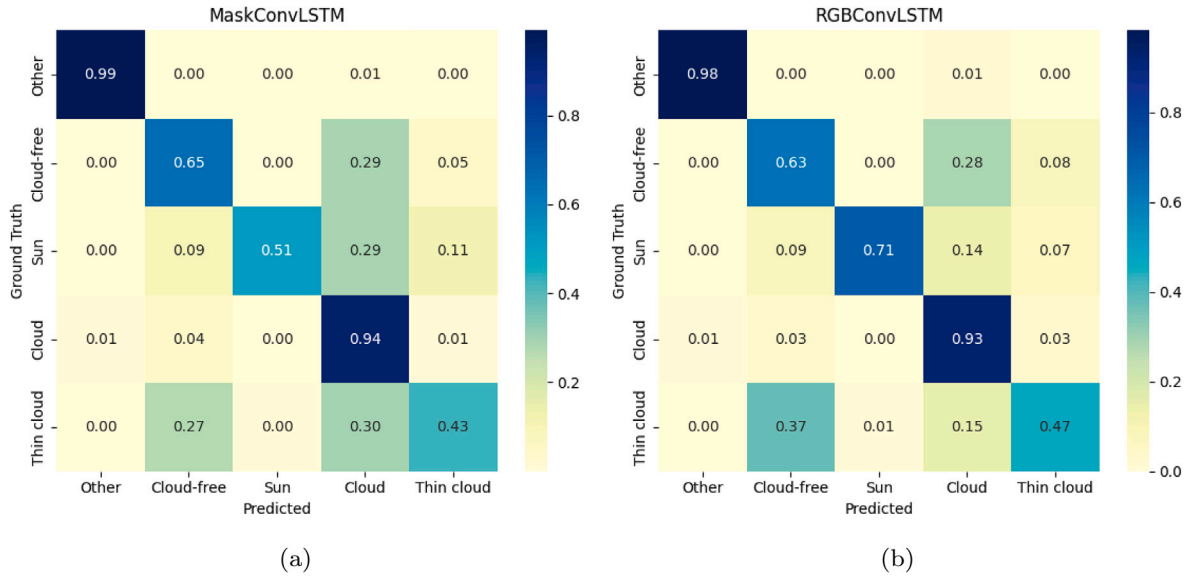


Fig. 13. Row-normalized confusion matrices averaged across all time steps for (a) MaskConvLSTM model and (b) RGBConvLSTM model. Each matrix shows the distribution of predicted classes given the ground truth labels, corresponding to per-class recall.

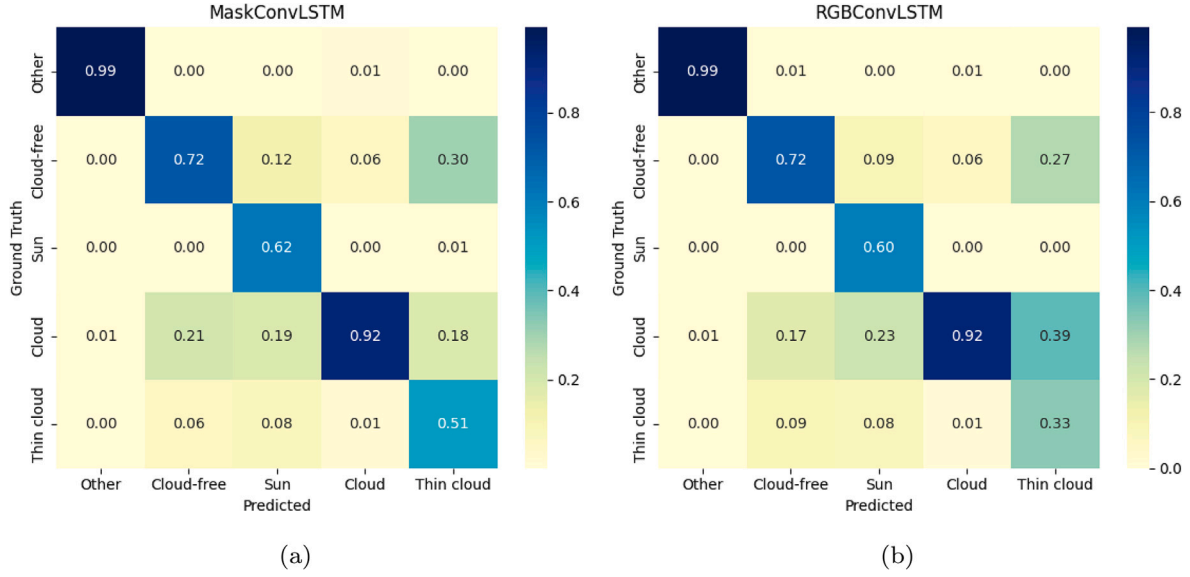


Fig. 14. Column-normalized confusion matrices averaged across all time steps for (a) MaskConvLSTM model and (b) RGBConvLSTM model. Each matrix shows the distribution of ground-truth labels given the predicted classes, corresponding to per-class precision.

The only requirement would be the availability of a suitable semantic segmentation model for the given dataset.

3. Results

3.1. Qualitative comparison

For a qualitative comparison, Figs. 9–11 present the segmented all-sky mask predicted by both models for three different cases of the test dataset: from high cloud cover with thin clouds to overcast (Case 1; Fig. 9), from high cloud cover to cloud-free (Case 2; Fig. 10), and from medium cloud cover to overcast (Case 3; Fig. 11). These three selected cases from the test set are not representative enough, but comprise varying atmospheric conditions, thus serving to highlight qualitative characteristics observed more broadly.

In Case 1, where the sky evolves from partly covered to totally overcast conditions (Fig. 9), both models' predictions look quite accurate and precise, replicating the change to overcast conditions. For Case 2, shown in Fig. 10, the sky goes from partly cloudy to completely cloud-free. In this situation, both models partially reproduce the transition, but neither accurately predicts the final cloud-free conditions. They correctly identify the trend toward clearing; however they underestimate the rate of dissipation, leaving residual cloud pixels that are not present in the ground truth. In Case 3 (Fig. 11) the sky varies from partly cloudy to completely overcast by low clouds. In this case, MaskConvLSTM model is not able to precisely replicate the remainder of overcast sky, as its prediction ends up showing a mostly overcast sky but with a notable cloud-free area. On the other hand, RGBConvLSTM model accurately imitates ground-truth's behavior, forecasting a constant overcast sky.

In general, the RGBConvLSTM model's outputs tend to become increasingly blurred over time due to repeated processing without explicit class separation. This effect accumulates across frames, leading

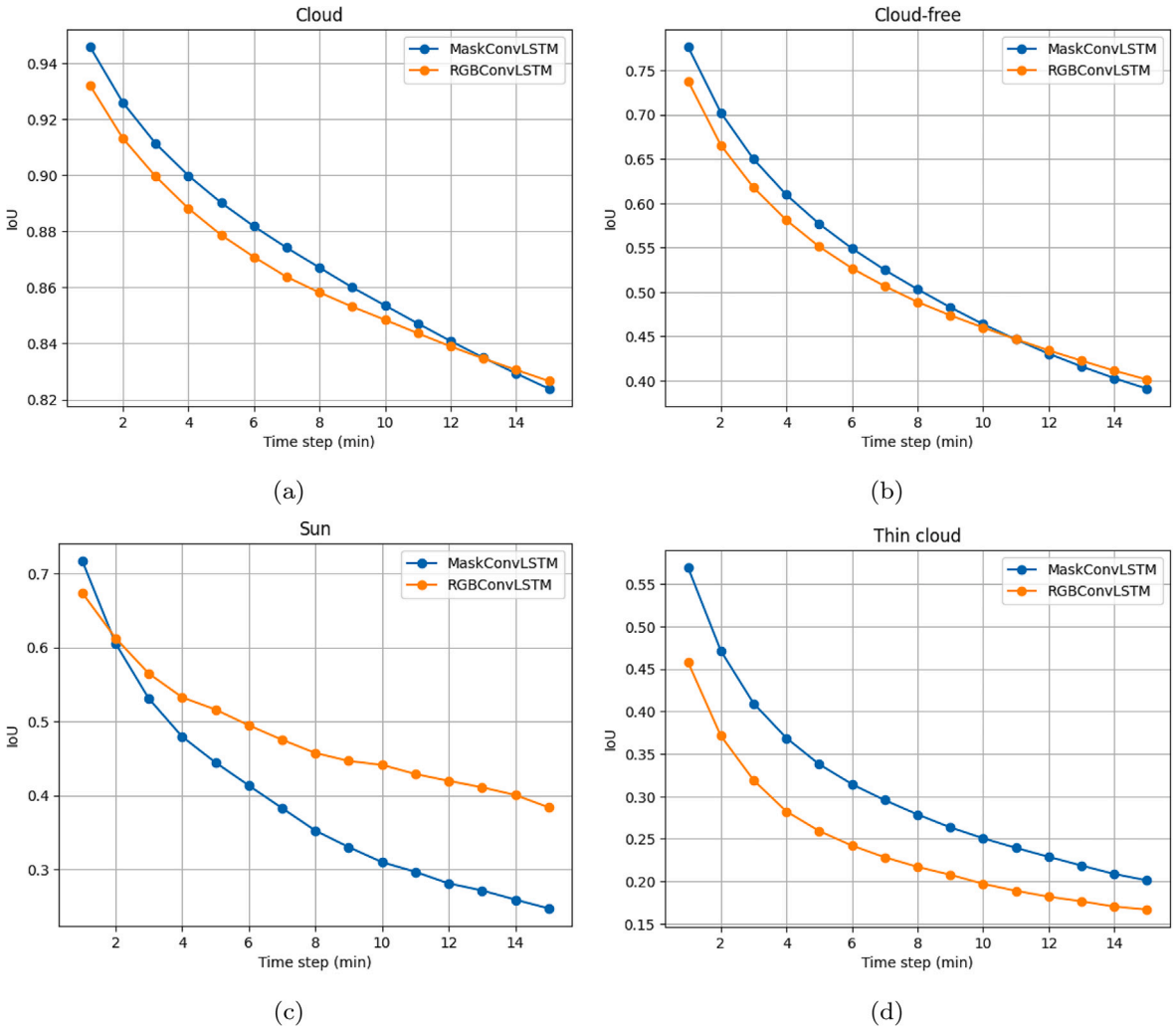


Fig. 15. IoU of RGBConvLSTM model and MaskConvLSTM model for the classes (a) Cloud, (b) Cloud-free, (c) Sun and (d) Thin cloud for each timestep.

to a gradual loss of spatial structure and semantic clarity. As a consequence, these degraded RGB predictions may fall outside the distribution of natural sky images, causing the sky segmentation model to produce unreliable masks. A similar behavior is observed internally in MaskConvLSTM model, where predictions are affected by progressive smoothing. This is not directly observed in the output because the most probable class is selected for each pixel. However, the underlying uncertainty remains and results in occasional failures.

These results are not conclusive in determining which model provides better predictions. In some cases, like Case 2, RGBConvLSTM model achieves better results, while in others, like Case 3, MaskConvLSTM model delivers a better performance. To rigorously determine which model presents the best performance, a quantitative analysis has been carried out in the next section using all the data from the test set.

3.2. Quantitative comparison

The quantitative performance has been evaluated using the metrics mentioned in Section 2.5. Fig. 12(a) reports the CCE metric, which has an inverse relationship with the performance, for both models. RGBConvLSTM model (adapted to segmentation) consistently shows slightly higher CCE values (from 0.952 to 1.021), hence a worse performance, than MaskConvLSTM model (CCE from 0.940 to 1.018). The confidence intervals confirm that the difference in means is more pronounced in early time steps, while it becomes less robust in the later stages of the prediction horizon. As expected, CCE increases as the time steps progress in both models, indicating a gradual decline in performance over time. For the initial frames performance decreases at a similar pace in both models. After that, MaskConvLSTM model's performance worsens faster than RGBConvLSTM model's.

The IoU values for both models are displayed in Fig. 12(b). MaskConvLSTM model achieves superior performance to RGBConvLSTM model only during the first seven time steps. After that, the roles are reversed and RGBConvLSTM model is the one which yields a better performance. On average, MaskConvLSTM model achieves a 0.49% increase in IoU compared to RGBConvLSTM model.

Fig. 12(c) shows that both models achieve high Dice coefficients. However, MaskConvLSTM model again performs better across all time steps. MaskConvLSTM model's performance decreases faster than RGBConvLSTM model's for the latter frames, as it did with CCE. MaskConvLSTM model maintains a Dice coefficient above 0.9 across eleven time steps, while RGBConvLSTM model drops below that threshold by the nine minute time step. On average, MaskConvLSTM model achieves a 0.94% improvement in Dice coefficient.

Both the CCE and the Dice coefficient reveal that the performance in the segmentation label space of RGBConvLSTM model is lower than for MaskConvLSTM model especially in the early prediction steps. When considering only the first five frames, MaskConvLSTM model achieves a Dice coefficient that is 1.26% higher than RGBConvLSTM model.

To assess whether the observed performance differences between models are statistically significant, a paired t-test has been conducted for

Table 1
Per-class recall over time for MaskConvLSTM model and RGBConvLSTM model. It shows the proportion of ground-truth pixels for each class that were correctly predicted at each time step ($t = 1 \dots 15$). Each column block corresponds to one model (MaskConvLSTM model on the left, RGBConvLSTM model on the right), with identical class labels across both.

Time step	MaskConvLSTM					RGBConvLST				
	Other	Cloud free	Sun	Cloud	Thin cloud	Other	Cloud free	Sun	Cloud	Thin cloud
t = 1	0.995	0.865	0.828	0.977	0.680	0.985	0.824	0.857	0.968	0.720
t = 2	0.993	0.808	0.734	0.970	0.587	0.984	0.761	0.816	0.958	0.660
t = 3	0.992	0.766	0.663	0.964	0.529	0.984	0.721	0.787	0.950	0.607
t = 4	0.992	0.731	0.611	0.959	0.491	0.984	0.691	0.761	0.943	0.560
t = 5	0.991	0.702	0.574	0.954	0.461	0.983	0.666	0.746	0.938	0.526
t = 6	0.991	0.677	0.540	0.950	0.438	0.983	0.645	0.729	0.934	0.496
t = 7	0.990	0.655	0.506	0.947	0.418	0.983	0.629	0.715	0.931	0.468
t = 8	0.990	0.635	0.471	0.943	0.399	0.983	0.615	0.698	0.928	0.441
t = 9	0.990	0.616	0.445	0.940	0.382	0.983	0.601	0.683	0.926	0.420
t = 10	0.990	0.598	0.421	0.936	0.368	0.982	0.589	0.675	0.924	0.396
t = 11	0.989	0.582	0.404	0.933	0.356	0.982	0.577	0.660	0.922	0.376
t = 12	0.989	0.567	0.384	0.929	0.344	0.982	0.564	0.644	0.921	0.360
t = 13	0.989	0.554	0.371	0.926	0.332	0.982	0.552	0.631	0.920	0.347
t = 14	0.989	0.541	0.355	0.922	0.320	0.982	0.539	0.613	0.918	0.334
t = 15	0.989	0.529	0.341	0.919	0.310	0.981	0.527	0.592	0.918	0.325

each timestep and evaluation metric. The results (not shown) indicate that the improvements achieved by MaskConvLSTM model are statistically significant across the first 13 forecast steps for CCE and the Dice coefficient, and for 5 steps for IoU. Therefore, it can be established that MaskConvLSTM model obtains better results only for the first forecast steps compared to RGBConvLSTM model within this framework.

While IoU scores are good for the initial frames but decline rapidly after them, the Dice coefficient remains consistently high, indicating that overall segmentation performance within this framework is preserved despite increasing spatial misalignments. As previously explained, this can be due to unbalances in the datasets, where some classes appear way more than others, and slight mismatches in class boundaries. IoU penalizes this imbalances harsher than the Dice coefficient.

The analyzed statistical indices provide a general overview of the performance of the models in predicting the sky mask of the upcoming frames under the considered setup. However, they do not offer direct information about the models' performance in forecasting the individual elements of the all-sky images (i.e., the five segmented classes).

To address this, the predicted classes of the models have been compared with the ground truth classes of the prediction targets of the test dataset using row-normalized confusion matrices, which reflect per-class recall. The results are presented as confusion matrices in Fig. 13 for each model. This figure shows that both models display similar confusion patterns, with the main difference being that MaskConvLSTM model misclassifies more frequently actual sun and thin cloud pixels as cloud or cloud-free. It is important to note that sun and thin cloud pixels represent only 0.2% and 1.9% of the test set, respectively, whereas cloud and cloud-free pixels account for 44.9% and 10.9%. As a result, MaskConvLSTM model's focus on the dominant classes leads to overall better segmentation performance, with the cost of an increased misclassification of less-frequent but meteorologically relevant classes. As expected, the class other is well identified for both models, since it generally remains constant in all the frames. To complement this analysis, Fig. 14 presents column-normalized confusion matrices, which reflect per-class precision. These matrices show that MaskConvLSTM model's predictions of infrequent classes such as sun or thin cloud, are more reliable than RGBConvLSTM model's despite the low recall observed in these classes. This highlights a trade-off between sensitivity and reliability for infrequent classes.

Another way to analyze the per-class performance of each model is by observing the class-wise behavior of any of the metrics from Section 2.5 over the prediction horizon. This has been done only for IoU as it is the most suitable for a comprehensive performance analysis. It is shown in Fig. 15. MaskConvLSTM model reports better results for all classes except for sun, where RGBConvLSTM model performance exhibits a much slower decrease over time, yielding better results just after the first predicted frame. This suggests that RGBConvLSTM model tends to overestimate sun pixels' persistence over time at the expense of the other categories. It is also possible that the segmenter model's initial limitations regarding sun pixels are also responsible for this behavior.

The results of Figs. 13 and 14 are for all time steps together. To observe the dependence of the performance of the models on the time step, Table 1 respectively shows the success rate or recall (rate at which a model predicts the correct class) of the models for each time step value. The other class is well predicted by both models, as expected since this class does not present significant temporal changes, with recall values above 98% for all time steps. The success rate of MaskConvLSTM model for cloud-free class decreases from 86.5% after 1 minute to 52.9% after 15 minutes; it is always higher than that for the RGBConvLSTM model, especially for the first steps. For the cloud class, MaskConvLSTM model is the model with higher success rate values, although never more than 1.5% over RGBConvLSTM model's. The success rate of MaskConvLSTM model strongly decreases from the first to the last time step, down to 48.6% and 37%, for the classes sun and thin cloud, respectively. This decrease is lower for the RGBConvLSTM model, 26.5% for sun and 39.5% for thin cloud, showing better results than MaskConvLSTM model for these two classes.

4. Conclusions

This study presents a controlled comparison of input representations for short-term multi-frame prediction in all-sky images. This is done by evaluating the same ConvLSTM backbone under two different types of data: regular RGB sky images for RGBConvLSTM model, and semantically segmented masks for MaskConvLSTM model. Within this evaluation framework, results indicate that operating in a semantic mask space can improve temporal prediction stability. This is shown across multiple evaluation metrics, including CCE, IoU and the Dice coefficient, particularly in the early prediction frames where performance decay is slower in the segmentation-based approach. While MaskConvLSTM model improves agreement for dominant classes, it underperforms for thin cloud and sun, which are meteorologically relevant despite their low frequency. This limitation is likely amplified by the quality of the segmentation targets, therefore, the results for these classes should be carefully interpreted.

Both training targets and evaluation references are derived from the same segmentation model, thus the reported results should be interpreted as agreement within the segmenter's output space rather than

as direct validation against physical cloud observations. In addition, the use of low-resolution (64 × 64) all-sky imagery constrains the representation of fine-grained cloud structures, and further validation against higher-resolution data and physical irradiance measurements remains necessary.

These findings suggest that incorporating semantic segmentation as a preprocessing step may enhance the model's ability to capture spatiotemporal patterns. Unlike prior methods based solely on 8-bit RGB imagery, our approach provides a structured representation that leads to greater stability and accuracy. While this suggests potential relevance for solar energy nowcasting and other applications, further validation against physical irradiance measurements is required.

Nevertheless, the dependence on high-quality segmentation and real-time processing remains a limitation. Future work should address these challenges while exploring extending this methodology to other spatiotemporal sequence prediction problems, and combining it with more advanced temporal models such as transformers or physics-constrained networks.

CRediT authorship contribution statement

Javier Gatón: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Roberto Román:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Cesar Guzman:** Writing – review & editing, Conceptualization. **Daniel González-Fernández:** Writing – review & editing. **Bruno Longarela:** Writing – review & editing. **Carlos Toledano:** Writing – review & editing, Funding acquisition. **Ramiro González:** Writing – review & editing.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used *ChatGPT-4o* exclusively to assist with grammar checking and spelling correction of the manuscript. No content, analysis, results, or scientific claims were generated by *ChatGPT-4o*. All text, figures, and conclusions were produced by the authors, who take full responsibility for the content of the publication.

Funding

This work was supported by the Ministerio de Ciencia e Innovación (MICINN), with the grant no. PID2021-127588OB-I00. This work is part of the project TED2021-131211B-I00375 funded by MCIN/AEI/10.13039/501100011033 and European Union, "NextGenerationEU"/PRTR and is based on work from COST Action CA21119 HARMONIA. Financial support from the Department of Education, Junta de Castilla y León, and FEDER Funds is gratefully acknowledged (Reference: CLU-2023-1-05). The authors acknowledge the support of the Spanish Ministry for Science and Innovation to ACTRIS ERIC. This work was supported as part of EUBURN-RISK (S2/2.4/F0327), an Interreg Sudoe Programme project co-funded by the European Union.

Declaration of competing interest

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

Code and data availability

Access to the code for the models implemented in this study and their trained coefficients is provided in the repository: https://github.com/javgat/SkyMaskConvLSTM.

References

[1] F. Creutzig, P. Agoston, J.C. Goldschmidt, G. Luderer, G. Nemet, R.C. Pietzcker, The underestimated potential of solar energy to mitigate climate change, Nat. Energy 2 (9) (2017) 17140, https://doi.org/10.1038/nenergy.2017.140
[2] M. Iqbal, An Introduction to Solar Radiation, Academic Press, Orlando, FL, 1983, https://www.osti.gov/biblio/5596615.
[3] K.N. Liou, An Introduction to Atmospheric Radiation, International Geophysics, Academic Press, 2002, https://books.google.es/books?id=mQ1DiDpX34UC.
[4] D.G. Erbs, S.A. Klein, J.A. Duffie, Estimation of the diffuse radiation fraction for hourly, daily and monthly-average global radiation, Sol. Energy 28 (4) (1982) 293–302, https://doi.org/10.1016/0038-092X(82)90302-4, https://www.sciencedirect.com/science/article/pii/0038092X82903024.
[5] J. Chantana, S. Ueno, Y. Ota, K. Nishioka, T. Minemoto, Uniqueness verification of direct solar spectral index for estimating outdoor performance of concentrator photovoltaic systems, Renew. Energy 75 (2015) 762–766, https://doi.org/10.1016/j.renene.2014.10.059, https://www.sciencedirect.com/science/article/pii/S0960148114006867.
[6] K.A. Browning, Conceptual models of precipitation systems, Weather Forecast. 1 (1) (1986) 23–41, https://doi.org/10.1175/1520-0434(1986)001<0023:CMOPS>2.0.CO;2, https://journals.ametsoc.org/view/journals/wefo/1/1/1520-0434_1986_001_0023_cmops_2_0_co_2.xml.
[7] Y. Chu, H. Pedro, M. Li, C. Coimbra, Real-time forecasting of solar irradiance ramps with smart image processing, Sol. Energy 114 (2015) 91–104, https://doi.org/10.1016/j.solener.2015.01.024
[8] D. Yang, J. Kleissl, C.A. Gueymard, H.T.C. Pedro, C.F.M. Coimbra, History and trends in solar irradiance and PV power forecasting: a preliminary assessment and review using text mining, Sol. Energy. 168 (2018) 60–101, Advances in Solar Resource Assessment and Forecasting. doi:https://doi.org/10.1016/j.solener.2017.11.023, https://www.sciencedirect.com/science/article/pii/S0038092X17310022.
[9] J.M. Arrais, A. Cerentini, B.J. Martins, T.Z.L. Chaves, S.L.M. Neto, A. von Wangenheim, Systematic review on ground-based cloud tracking methods for photovoltaics nowcasting, Am. J. Clim. Change 13 (3) (2024) 452–476, https://doi.org/10.4236/ajcc.2024.133021
[10] Z. Yu, Z. Tan, S. Ma, W. Yan, Nowcast for cloud top height from Himawari-8 data based on deep learning algorithms, Meteorol. Appl. 30 (3) (2023) e2130, https://doi.org/10.1002/met.2130, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/met.2130.
[11] L. Berthomier, B. Pradel, L. Perez, Cloud cover nowcasting with deep learning, in: 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA), 2020, pp. 1–6, https://doi.org/10.1109/IPTA50016.2020.9286606
[12] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. Kin Wong, W. Chun Woo, Convolutional LSTM network: a machine learning approach for precipitation nowcasting, arXiv:1506.04214, 2015, https://arxiv.org/abs/1506.04214.
[13] Y. Wang, H. Wu, J. Zhang, Z. Gao, J. Wang, P.S. Yu, M. Long, PredRNN: a recurrent neural network for spatiotemporal predictive learning, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2) (2023) 2208–2225, https://doi.org/10.1109/TPAMI.2022.3165153
[14] W. Yan, Y. Zhang, P. Abbeel, A. Srinivas, VideoGPT: video generation using VQ-vae and transformers, arXiv:2104.10157, 2021, https://arxiv.org/abs/2104.10157.
[15] Y. Son, X. Zhang, Y. Yoon, J. Cho, S. Choi, LSTM–GAN based cloud movement prediction in satellite images for PV forecast, J. Ambient Intell. Humanized Comput. 14 (9) (2023) 12373–12386, https://doi.org/10.1007/s12652-022-04333-7
[16] H. Mansourifar, A. Moskovitz, B. Klingensmith, D. Mintas, S.J. Simske, Gan-based satellite imaging: a survey on techniques and applications, IEEE Access 10 (2022) 118123–118140, https://doi.org/10.1109/ACCESS.2022.3221123
[17] V. Le Guen, N. Thome, Disentangling physical dynamics from unknown factors for unsupervised video prediction, CoRR arXiv:2003.01460, 2020, https://arxiv.org/abs/2003.01460.
[18] Y. Nie, E. Zelikman, A. Scott, Q. Paletta, A. Brandt, SkyGPT: probabilistic ultra-short-term solar forecasting using synthetic sky images from physics-constrained VideoGPT, Adv. Appl. Energy 14 (2024) 100172, https://doi.org/10.1016/j.adapen.2024.100172, https://www.sciencedirect.com/science/article/pii/S2666792424000106.
[19] P. Luc, N. Neverova, C. Couprie, J. Verbeek, Y. LeCun, Predicting deeper into the future of semantic segmentation, arXiv:1703.07684, 2017, https://arxiv.org/abs/1703.07684.
[20] Y. Nie, X. Li, A. Scott, Y. Sun, V. Venugopal, A. Brandt, Skipp'd: a sky images and photovoltaic power generation dataset for short-term solar forecasting, Sol. Energy 255 (2023) 171–179, https://doi.org/10.1016/j.solener.2023.03.043, https://www.sciencedirect.com/science/article/pii/S0038092X23002037.
[21] R. Román, J. Gatón, D. González-Fernández, S. Herrero-Anta, C. Herrero del Barrio, B. Longarela, J.L. Martín Marcos, R. González, GOA-UVa All-Sky Segmentation U-Net Model, Zenodo, 2026, https://doi.org/10.5281/zenodo.18894938
[22] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, arXiv:1505.04597, 2015, https://arxiv.org/abs/1505.04597.
[23] C.J. Van Rijsbergen, Information Retrieval, Butterworths, 1975, https://cir.nii.ac.jp/crid/1970023484860037059.
[24] P. Desai, S. Ch, S. Chakraborty, S. Ansuman, S. Bhandari, S. Kardiguddi, Next frame prediction using convlstm, J. Phys. Conf. Ser. 2161 (2022) 012024, https://doi.org/10.1088/1742-6596/2161/1/012024
[25] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735
[26] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison,

A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: an imperative style, high-performance deep learning library, arXiv:1912.01703, 2019, https://arxiv.org/abs/1912.01703.

[27] J.F. Kolen, S.C. Kremer, A field guide to dynamical recurrent networks, in: A Field Guide to Dynamical Recurrent Networks, John Wiley & Sons, 2001, p. 202, https://books.google.es/books?id=NWOcMVA64aAC&pg=PA202&redir_esc=y#v=onepage&q&f=false.

[28] F. Chollet, et al., ReduceLROnPlateau - Keras Documentation, 2015, https://keras.io/api/callbacks/reduce_lr_on_plateau/ (accessed: 8 October 2024).

[29] K.P. Murphy, Probabilistic Machine Learning: an Introduction, MIT Press, 2022, http://probml.github.io/book1.

[30] A. Carass, S. Roy, A. Gherman, J.C. Reinhold, A. Jesson, T. Arbel, O. Maier, H. Handels, M. Ghafoorian, B. Platel, A. Birenbaum, H. Greenspan, D.L. Pham, C.M. Crainiceanu, P.A. Calabresi, J.L. Prince, W.R.G. Roncal, R.T. Shinohara, I. Oguz, Evaluating white matter lesion segmentations with refined sørensen-dice analysis, Sci. Rep. 10 (1) (2020) 8242, https://doi.org/10.1038/s41598-020-64803-w

[31] P. Jaccard, Étude comparative de la distribution florale dans une portion des Alpes ET du Jura, Soc. Vaud. Sci. Nat. 37 (142) (1901), https://doi.org/10.5169/seals-266450

[32] S. Jadon, A survey of loss functions for semantic segmentation, in: 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE, 2020, pp. 1–7, https://doi.org/10.1109/cibcb48159.2020.9277638, http://dx.doi.org/10.1109/CIBCB48159.2020.9277638.

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

[34] F. Milletari, N. Navab, S.-A. Ahmadi, V-Net: fully convolutional neural networks for volumetric medical image segmentation, arXiv:1606.04797, 2016, https://arxiv.org/abs/1606.04797.

[35] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick, P. Dollár, Microsoft COCO: common objects in context, arXiv:1405.0312, 2015, https://arxiv.org/abs/1405.0312.