Improving CNN Generalization for Photovoltaic Nowcasting Under Data Scarcity Through Sky Image Hybrid Augmentation Approaches

Markos A. Kousounadis-Knousen , Velissarios Theocharis , Athina P. Georgilaki , Pavlos S. Georgilakis *

Posted Date: 27 April 2026

doi: 10.20944/preprints202604.1794.v1

Keywords: photovoltaic power forecasting; nowcasting; deep learning; computer vision; sky images; data augmentation; resampling; convolutional neural networks; data scarcity

 doi:10.20944/preprints202604.1794.v1

Article

Improving CNN Generalization for Photovoltaic Nowcasting Under Data Scarcity Through Sky Image Hybrid Augmentation Approaches

Markos A. Kousounadis-Knousen, Velissarios Theocharis, Athina P. Georgilaki and Pavlos S. Georgilakis *

School of Electrical and Computer Engineering, National Technical University of Athens, 15780, Athens, Greece
* Correspondence: pgeorg@power.ece.ntua.gr

Abstract

Reliable photovoltaic (PV) power forecasting based on deep learning typically requires large historical datasets to capture the high temporal and spatial variability of solar irradiance. However, in many real-world applications, data availability is limited to short observation periods, hindering the effective training of deep learning models. This paper investigates how sky image data augmentation techniques can improve the generalization capability of Convolutional Neural Networks (CNNs) trained under data scarcity. Three augmentation-based oversampling methods—SMOTE, Mixup-kNN, and Mixup-RP—are evaluated, along with two novel hybrid strategies that combine them in-parallel and in-series configurations. The proposed framework is validated on two distinct PV power nowcasting case studies, in which the original sky image training datasets span less than one month. Experimental results show average performance improvements of up to 50% on external validation data when training the CNN on the augmented datasets compared to the original base datasets, demonstrating that accurate PV power nowcasting is feasible even under data-scarce conditions typical of newly installed PV systems, and highlighting the potential of data-efficient learning approaches for renewable energy applications.

Keywords: photovoltaic power forecasting; nowcasting; deep learning; computer vision; sky images; data augmentation; resampling; convolutional neural networks; data scarcity

1. Introduction

Increased energy consumption has been associated with improved economic performance and higher living standards across modern societies; however, it has also led to significant environmental challenges that contribute to the ongoing climate crisis [1]. In response, the European Union has adopted an ambitious strategy to gradually replace fossil fuels with Renewable Energy Sources (RES) for electric power generation. Over the past decade, this transition has been driven largely by the growing deployment of Photovoltaic (PV) systems, which have been connected to the electrical grid predominantly at the distribution level [2]. The increasing penetration of such non-dispatchable PV systems necessitates the modernization of conventional electric power systems through the adoption of smart power electronics, dynamic modeling techniques, and Artificial Intelligence (AI), to mitigate the negative impact of the induced uncertainty and ensure reliable and efficient operation [3].

Solar power forecasting is one of the most effective methods for the reliable integration of PV systems into the electrical grid, as it enables optimal planning and proactive control while reducing operational costs [4]. Forecasting horizons vary depending on the downstream task, ranging from a few seconds ahead (nowcasting), to minutes ahead (ultra-short-term forecasting), hours ahead (short-term forecasting), and up to days or even weeks ahead (mid- and long-term forecasting) [4]. For ultra-short- and short-term horizons, remote sensing data from satellites and ground-based sky cameras have emerged as particularly promising input sources. These types of data provide detailed information on cloud formations in the form of images, thereby enabling the application of computer

 doi:10.20944/preprints202604.1794.v1

vision techniques for tasks such as cloud detection and cloud motion modeling within solar power forecasting frameworks [4].

Recent advances in AI have enabled the partial or complete replacement of traditional physics-informed computer vision techniques with data-driven deep learning methods for solar power forecasting [4]. For ultra-short-term horizons, where ground-based sky cameras provide images with the required spatiotemporal resolution for minutes-ahead analysis, two primary approaches have emerged [5]: i) directly forecasting the target variable (e.g., solar irradiance or PV power) from sequences of obtained sky images through deep learning (end-to-end modeling), e.g., with the usage of Convolutional Neural Networks (CNNs) or Vision Transformers (VT), and ii) splitting the forecasting task into two stages by first predicting future sky images (image forecasting), and then deriving the target variable from these images (nowcasting), using chained models based on deep learning or hybrid methods. End-to-end modeling for sky-image-based solar power forecasting was first explored in [6], which employed the Stanford University Neural network for Solar Electricity Trend (SUNSET) model [7] for 15-minute-ahead forecasting using sky image sequences. Since then, various end-to-end solar power forecasting architectures have been explored, including Convolutional Long Short-Term Memory (ConvLSTM) models, ResNet-based models, and encoder-decoder frameworks [8,9]. Nevertheless, recent studies suggest that end-to-end models may struggle to predict the highly non-linear dynamics of solar power generation directly from sky images, indicating that splitting the forecasting task into subtasks can be advantageous [5].

A major drawback of AI-based models for solar power forecasting is their strong dependence on large volumes of high-quality historical data for effective training [10]. Despite the growing availability of PV-related datasets, data scarcity remains a significant challenge, particularly with the widespread deployment of distributed PV systems and the adoption of more complex model architectures. In general, data requirements increase with the number of trainable parameters, making it unrealistic to assume that sufficient historical data will always be available, especially for sky image data and small-scale local PV installations. In addition, sky image datasets are often imbalanced, as certain sky conditions are over-represented due to the prevailing climatic conditions of the target location [11]. Since machine learning models typically optimize objective functions based on average errors, under-represented conditions tend to be inadequately learned during training. This issue is particularly pronounced in regions dominated by clear sky conditions, where sky images depicting clouds are under-represented. As a result, models become biased towards clear sky patterns, which are inherently easier to predict and generally do not require complex deep learning architectures.

Based on the above, it is evident that acquiring sufficiently large and balanced historical sky image datasets for training AI-based solar power forecasting models remains a challenge. In recent years, data scarcity has been addressed through transfer learning, where models are pre-trained on large global datasets and subsequently finetuned using local data [12]. Although this approach can substantially reduce the need for extensive local datasets, it still relies on access to large-scale data repositories that adequately resemble the conditions of the target location. On the other hand, dataset imbalance has primarily been addressed through classification-based approaches. In classification-based approaches, sky images are partitioned into distinct classes either to train separate sky-condition-specific models for each class [13,14], or to construct a more balanced training subset for a single model through targeted sampling [15]. However, while these approaches can improve the representation of diverse sky conditions, they do not resolve the data scarcity problem, increasing the risk of overfitting or training collapse [16].

Dataset augmentation has emerged as a promising approach to address the combined challenges of data scarcity and imbalance in the context of AI-based electric power system applications [10]. In contrast to data resampling, which only replicates existing samples, data augmentation enriches datasets by generating new synthetic samples [4], thereby reducing the risk of overfitting and improving the generalization capability of deep learning models. In the context of image data, augmentation can be conducted using simple transformations, such as Gaussian noise injection [17], color casting [18], and brightness adjustment [19]. More advanced approaches include data-driven techniques, such as the Synthetic Minority Oversampling Technique (SMOTE) [20] and Mix-up k-

 doi:10.20944/preprints202604.1794.v1

Nearest Neighbors (kNN) [21], as well as deep generative AI models such as Generative Adversarial Networks (GANs) [22].

Despite the growing interest in image data augmentation for RES-related applications, such as PV panel soiling localization [23], limited research has been conducted on sky image data augmentation for solar power forecasting. A thorough exploration of various augmentation methods in the context of sky-image-based PV power forecasting, including noise injection, color transformations, and image mixing, was first presented in [11]. The different augmentation methods were systematically evaluated based on the performance of the SUNSET model on two tasks: a nowcasting task and a 15-minute-ahead forecasting task. In [24], the sky image dataset was augmented using translational, vertical, and temporal transformations, improving irradiance forecasting performance of three deep learning models across multiple horizons up to 10 minutes ahead. In [25], augmentation techniques including color adjustments, cropping, and rotation, were combined with transfer learning to improve sky-image-based PV power nowcasting performance of two deep CNN-based models. Similarly, [26] incorporated several augmentation techniques, such as translations, scaling, and flipping, within a comprehensive sky image pre-processing pipeline to improve solar irradiance forecasting accuracy.

Nevertheless, some important research gaps remain. Most notably, existing studies apply sky image data augmentation in data-abundant settings, as the datasets used in [11], [24]-[26] span two or three years at minute-scale temporal resolutions, resulting in several hundred thousand sky images. Such data availability is not representative of newly deployed PV systems, particularly in small-scale and resource-constrained environments. Furthermore, [24]-[26] do not explicitly address dataset imbalance, as augmentation is applied across the entire dataset, aiming only to enhance diversity. Only [11] considers dataset imbalance, by splitting the dataset into two subsets based on the error distribution of a baseline forecasting model and applying augmentation selectively to the higher-error subset. However, this approach relies on a relatively simple classification scheme, which may not fully capture the underlying structural imbalances present in the dataset.

This paper presents a novel sky image augmentation framework aimed at improving the generalization capability of CNNs for PV power nowcasting under data scarcity. Unlike existing studies, the proposed framework applies augmentation to a base training dataset comprising only a few days of data, enabling a systematic evaluation of its effectiveness when data availability is limited. Accordingly, the objective of augmentation is not only to mitigate dataset imbalance, but also to increase dataset size to a level sufficient for training deep CNN models. Furthermore, dataset imbalances are identified using a recently proposed sky image clustering approach [14], which results in detailed clusters representing diverse sky conditions. To further address the limited size of the base training dataset and enhance data diversity, new hybrid augmentation approaches based on image mixing are also explored. The proposed framework is evaluated using two distinct sky image datasets, spanning several weeks of unseen data. The main contributions of this paper are summarized as follows:

1. The introduction of a novel, holistic framework to evaluate the impact of different sky image augmentation methods on PV power nowcasting performance under data scarcity. The proposed framework integrates dataset clustering, resampling, and hybrid augmentation, and is evaluated using the SUNSET model on two small-scale sky image datasets.
2. The development of novel hybrid data-driven augmentation strategies based on image mixing. SMOTE, Mixup-kNN, and Mixup-Random Pair (RP) are combined both in series and in parallel, to generate more diverse synthetic sky images and further improve nowcasting performance.
3. The employment of a state-of-the-art automatic sky image clustering approach to identify detailed clusters and reveal underlying dataset imbalances. Clustering is performed with respect to the downstream task, i.e., PV power nowcasting, and clusters are characterized as critical and non-critical for augmentation based on their size and their associated nowcasting errors.

The remainder of this paper is organized as follows: Section 2 presents the sky image datasets, the PV power nowcasting settings, and the proposed sky image augmentation framework. The

 doi:10.20944/preprints202604.1794.v1

experimental results are presented in Section 3. Section 4 discusses the experimental findings and summarizes the main insights. Concluding remarks are provided in Section 5.

2. Materials and Methods

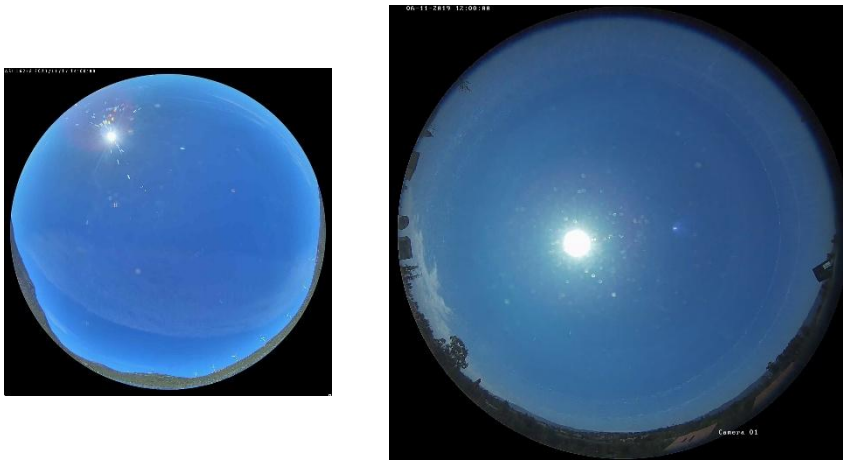
2.1. Sky Image Datasets

2.1.1. *Archon* Dataset

The *Archon* dataset [27] contains RGB sky images captured at a PV system in Greece between 16/11/2023 and 06/01/2024 using a professional All-Sky Imager (ASI)-16 ground-based sky camera with a fisheye lens and a 180° Field Of View (FOV). The images have a resolution of 1536 × 1536 pixels, 96 DPI, and 8-bit color depth, and are captured at 1-minute intervals during daylight hours. An example image captured under clear sky conditions at midday is shown in Figure 1(a). The ASI-16 camera is installed near a 1.2 kW PV system, from which power measurements are also recorded at a 1-minute resolution.

2.1.2. SKIPP'D Dataset

The publicly available *SKIPP'D* dataset [28] contains RGB sky images captured at a rooftop PV system at the Stanford Campus in the US between 2017 and 2019 using a commercial HIKVISION surveillance camera with a fisheye lens and a 180° FOV. The camera has a resolution of 2048 × 2048 pixels and 8-bit color depth, and records videos at 20 frames per second. In this paper, images are extracted at 1-minute intervals during daylight hours only for the period between 01/02/2019 and 31/07/2019. An example image captured under clear sky conditions at midday is shown in Figure 1(b). The surveillance camera is installed approximately 125 m away from the 30 kW PV system, from which power measurements are also recorded at a 1-minute resolution.



(**a**) (**b**)

Figure 1. Example sky images under clear sky conditions at midday: (**a**) *Archon* dataset; (**b**) *SKIPP'D* dataset.

2.2. PV Power Nowcasting

2.2.1. Mathematical Formulation

As mentioned in Section 1, end-to-end ultra-short-term PV power forecasting using sky images is inherently complex, and deep learning models often struggle to capture the non-linear relationships between sequences of past sky images and future PV power values. In recent years, deeper AI models based on VTs have demonstrated promising performance in end-to-end ultra-short-term PV power forecasting; however, training such models requires even larger datasets compared to more conventional deep learning models such as CNNs, to accommodate their billions of parameters. Therefore, this paper focuses on PV power nowcasting under the assumption that

 doi:10.20944/preprints202604.1794.v1

future sky images have already been predicted using a chained model, such the auto-encoder-like CNN proposed in [16].

In this paper, the PV power nowcasting problem is mathematically formulated as follows:

$$\hat{y}(t) = \mathrm{f}\big(\boldsymbol{x}(t)\big) \tag{1}$$

where $\boldsymbol{x}(t)$ is the matrix corresponding to the sky image recorded at time t, $\hat{y}(t)$ is the predicted PV power at time t, and f(·) is the underlying function that maps the output (PV power) to the input (sky image). In other words, the target objective is to estimate PV production at a given time t using only the recorded sky image at time t.

2.2.2. PV Power Nowcasting Model

To extract the mapping function f(·) in (1), the CNN-based SUNSET model [7] is employed. The architecture of SUNSET, illustrated in Figure 2, comprises two convolutional blocks followed by a fully connected block. The model receives a single sky image as input, downscaled to 64 × 64 pixels to reduce computations, with pixel values normalized to [0, 1] per color channel. Each convolution block contains three layers: a convolutional layer, a batch normalization layer, and a pooling layer. The convolutional layers use 3 × 3 kernels with unit stride and padding to preserve input dimensions for the feature maps. Pooling layers then down-sample these feature maps by a factor of two. The output of the convolutional blocks is flattened and passed to the fully connected block, which contains two layers with 1024 neurons each. All hidden layers use ReLU activation. The output layer uses linear activation and contains a single neuron, which represents the predicted PV power.

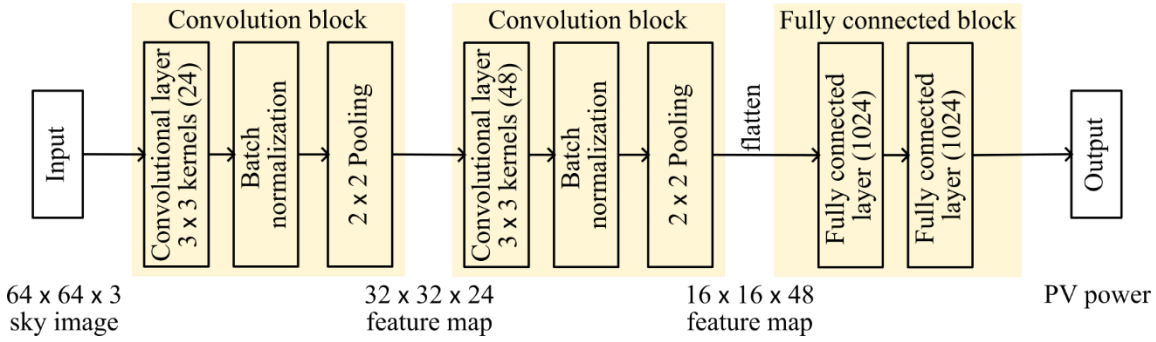


Figure 2. Architecture of the SUNSET PV power nowcasting model [7].

SUNSET is a deep learning computer vision model capable of directly extracting point values from input images. It has demonstrated state-of-the-art performance and is widely regarded as a reliable PV power nowcasting model. Nevertheless, the primary motivation for selecting SUNSET in this study is its relatively low number of trainable parameters compared to other models in the same category. Specifically, SUNSET comprises only a few million trainable parameters [12], making it feasible to train effectively with datasets consisting of only a few thousand sky images. In this context, augmenting a limited base dataset becomes a practical approach for enabling robust model training. In contrast, deeper CNN-based architectures such as ResNet-50, as well as models based on VTs, typically involve tens of millions to billions of parameters. Consequently, they require substantially larger training datasets, which are often unavailable under data-scarce conditions.

2.3. Sky Image Augmentation Methods

The base datasets presented in Section 2.1 are augmented using SMOTE, Mixup-kNN, Mixup-RP, as well as novel in-series or in-parallel combinations of these image mixing techniques. Generative AI models such as GANs are not considered in this study, as they require substantial amounts of training data to produce reliable synthetic images, making them unsuitable for data-scarce environments. Simpler augmentation techniques, such as Gaussian noise injection, brightness adjustment, and color casting, were also evaluated; however, preliminary experiments indicated that their impact is limited compared to more advanced mixing methods. This contrasts with the findings

 doi:10.20944/preprints202604.1794.v1

in [11], where nowcasting performance differences among augmentation techniques were relatively small, indicating that when the base dataset is sufficiently large, model performance becomes less sensitive to the choice of augmentation method due to the sufficient diversity of the base dataset. In contrast, under data-scarce conditions, the limited diversity of the base dataset necessitates more sophisticated augmentation methods, which can generate greater variability and thus have a bigger impact on nowcasting performance.

2.3.1. Synthetic Minority Oversampling Technique

With SMOTE [20], a synthetic image is generated through linear interpolation between the original image and one of its k nearest neighbors, as follows:

$$\tilde{x}_i = \lambda x_{p,i} + (1-\lambda) x_{q,i} \; \forall \, j \in [1, N] \tag{2}$$

where $\tilde{x}_i$ is the i^{th} pixel of the synthetic image $\tilde{\boldsymbol{x}}$, $x_{p,i}$ is the i^{th} pixel of the original image $\boldsymbol{x}_p$, $x_{q,i}$ is the i^{th} pixel of the selected neighbor image $\boldsymbol{x}_q$, and λ is a random coefficient that follows a uniform distribution in [0, 1] and controls the interpolation ratio. Figure 3(a) depicts an augmentation example for the *Archon* dataset using SMOTE. The corresponding PV power value of the synthetic sky image is calculated as the weighted average of the PV power values of the two original images, using the same coefficient λ.

2.3.2. Mixup-kNN

Mixup-kNN [21] is conceptually similar to SMOTE, as synthetic images are generated through linear interpolation between the original image and one of its k nearest neighbors. However, in this case, the interpolation coefficient λ is sampled from a Beta(a, a) distribution, whose probability density function is given by:

$$\text{pdf}(\lambda; a, a) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} \cdot \lambda^{\alpha-1} \cdot (1-\lambda)^{\alpha-1}, \lambda \in [0, 1] \tag{3}$$

where $a \in (0, \infty)$ is the shape parameter of the Beta distribution and $\Gamma(\cdot)$ is the gamma function. Figure 3(b) depicts an augmentation example for the *Archon* dataset using Mixup-kNN. The corresponding PV power value of the synthetic sky image is calculated as the weighted average of the PV power values of the two original images, using the coefficient λ.

2.3.3. Mixup-RP

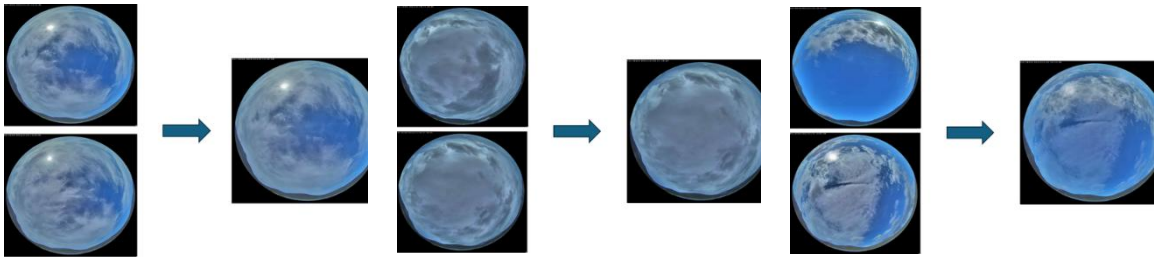
In contrast to SMOTE and Mixup-kNN, Mixup-RP [21] is not restricted to the k nearest neighbors but generates synthetic images through linear interpolation between the original image and any randomly selected image from the base dataset. The interpolation coefficient λ follows the same Beta(a, a) distribution as in Mixup-kNN. Figure 3(c) depicts an augmentation example for the *Archon* dataset using Mixup-RP. Compared to SMOTE and Mixup-kNN, Mixup-RP typically produces synthetic sky images with greater diversity, as randomly selected images from the dataset can differ significantly from the original image. The corresponding PV power value of the synthetic sky image is calculated as the weighted average of the PV power values of the two original images, using the coefficient λ.

2.4. Data Scarcity Environment Simulation

At design time, the sky image datasets presented in Section 2.1 are first split into training and testing subsets to prevent data leakage. To emulate data-scarce conditions, only 15% and 10% of the *Archon* and *SKIPP'D* datasets, respectively, are allocated to the training subsets, which is substantially lower than the typical allocation of at least 50% used in AI model development. Consequently, the training subset of *Archon* consists of 5782 sky images (approximately 10 days of data), whereas the training set of *SKIPP'D* consists of 9663 sky images (approximately 16 days of data). This limited proportion of training data, combined with the already small size of the sky image datasets, reflects realistic data-scarce scenarios, such as newly deployed PV systems with only a few days or weeks of

 doi:10.20944/preprints202604.1794.v1

recorded measurements. These training subsets serve as the base datasets which are subsequently clustered, analyzed for imbalances, and augmented accordingly.



(**a**) (**b**) (**c**)

Figure 3. Sky image augmentation examples for the *Archon* dataset: (**a**) SMOTE ($\lambda = 0.4$); (**b**) Mixup-kNN ($\alpha = 0.6$); (**c**) Mixup-RP ($\alpha = 0.8$).

2.5. Sky Image Dataset Clustering

Identifying imbalances in sky image datasets requires an appropriate data classification strategy rather than relying on simple classification schemes (e.g., sunny – cloudy – overcast). To this end, the automatic sky image clustering framework recently introduced in [14] is employed, primarily due to its ability to generate multiple clusters without relying on pre-defined ground truth labels. This multi-class partitioning enables a more detailed representation of the base dataset structure and facilitates the detection of subtle imbalances. At operation time, the framework proceeds as follows: a total of 49 handcrafted features are extracted from the current sky image, including spectral, textural, cloud coverage-related, and non-instantaneous features. These features are subsequently projected into a latent space using a hybrid dimensionality reduction technique that combines Principal Component Analysis with shallow fully-connected auto-encoders [29]. The sky image is then assigned to one of several pre-defined clusters, which have been created at design time using k-Means clustering [30]. More information on the employed sky image clustering approach can be found in [14].

2.6. Proposed Sky Image Dataset Resampling and Augmentation

In contrast to [11], which characterizes sky images individually as either critical or non-critical for augmentation, this paper conducts a cluster-level analysis to resample the base dataset and address imbalances. Specifically, to determine whether a cluster should be considered critical for augmentation, all sky images within the cluster are jointly evaluated through aggregation (e.g., averaging), which improves robustness to noise that may arise from the criticality criterion (such as the performance of a PV nowcasting model). Furthermore, operating at the cluster level enables cluster-specific augmentation strategies and more flexible handling of individual clusters, thereby expanding the possible ways to process the base dataset and address imbalances.

In the proposed method, each sky image cluster is characterized as critical or non-critical for augmentation based on two criteria: i) the cluster size relative to the base dataset, and ii) similar to [11], the associated downstream task performance, i.e., the average PV power nowcasting error. The proposed cluster-level criticality criterion is formulated as follows:

$$c_i \subseteq C \Leftrightarrow (|c_i| \leq r_1|D|) \wedge (\varepsilon_i \geq r_2 P) \quad \forall\, i \in \{1, 2, \ldots, n\} \tag{4}$$

where c_i denotes cluster i, C is the set of critical clusters (critical dataset), D is the base dataset, $|c_i|$ is the size of cluster i, n is the total number of clusters, ε_i is the PV power nowcasting error associated with cluster i, P is the installed capacity of the PV system, and r_1, r_2 are predefined coefficients determining the criticality thresholds. In other words, a cluster is considered critical for augmentation only if it is both relatively underrepresented in the base dataset and associated with relatively high PV power nowcasting errors. If a cluster is associated with low nowcasting errors (e.g., a cluster representing clear sky conditions), augmentation is unnecessary, even if the cluster is small, since such sky images do not significantly contribute to the overall nowcasting error. On the other

 doi:10.20944/preprints202604.1794.v1

hand, if a cluster is associated with high nowcasting errors but is also sufficiently large, it is excluded from augmentation, as its sky images are already adequately represented in the base dataset.

A schematic overview of the proposed sky image dataset augmentation framework for PV power nowcasting under data-scarce settings is provided in Figure 4. The base dataset D is first partitioned into n clusters using the sky image clustering method described in Section 2.5. The SUNSET model of Figure 2 is then trained separately on each cluster of the base dataset to solve the nowcasting problem defined in (1). Training performance is evaluated using the Root Mean Square Error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_{\mathrm{tr}}}\sum_{i=1}^{N_{\mathrm{tr}}}(\hat{y}(i) - y(i))^2} \quad (5$$

where $\hat{y}(i)$ is the predicted PV power for training sample i, $y(i)$ is the actual PV power for training sample i, and N_{tr} is the total number of training samples. Based on (4), each cluster is characterized as critical or non-critical for augmentation. The critical dataset C is resampled N times to increase dataset volume and achieve better balance with the non-critical dataset. The resampled sky images are subsequently augmented using one of the augmentation techniques described in Section 2.3 or one of the proposed hybrid augmentation methods described in Section 2.7, to create the augmented dataset $(NC)^*$. Finally, the base dataset is combined with the augmented dataset to form the final balanced dataset B, which is used to train the SUNSET model for PV power nowcasting:

$$B = D \cup (NC)^* \quad (6$$

2.7. Proposed Hybrid Augmentation Methods

Two novel hybrid augmentation strategies are introduced in this paper, combining SMOTE, Mixup-kNN, and Mixup-RP: i) in-parallel augmentation, and ii) in-series augmentation. In the in-parallel approach, instead of using a single augmentation method for all N copies of the critical dataset, N_1 copies are created using one augmentation technique, while the remaining $N_2 = N - N_1$ copies are created using an alternative technique:

$$\begin{cases} \tilde{x}_i^j = \lambda_1 x_{p,i} + (1-\lambda_1)x_{q,i}, & j \in [1, N_1] \\ \tilde{x}_i^j = \lambda_2 x_{p,i} + (1-\lambda_2)x_{q,i}, & j \in (N_1, N] \end{cases} \quad (7$$

where λ_1, λ_2 are the interpolation coefficients associated with the two selected augmentation methods. The in-parallel approach is expected to further increase dataset diversity, as it leverages two distinct augmentation techniques to generate synthetic sky images, thereby exploiting the complementary strengths of each method.

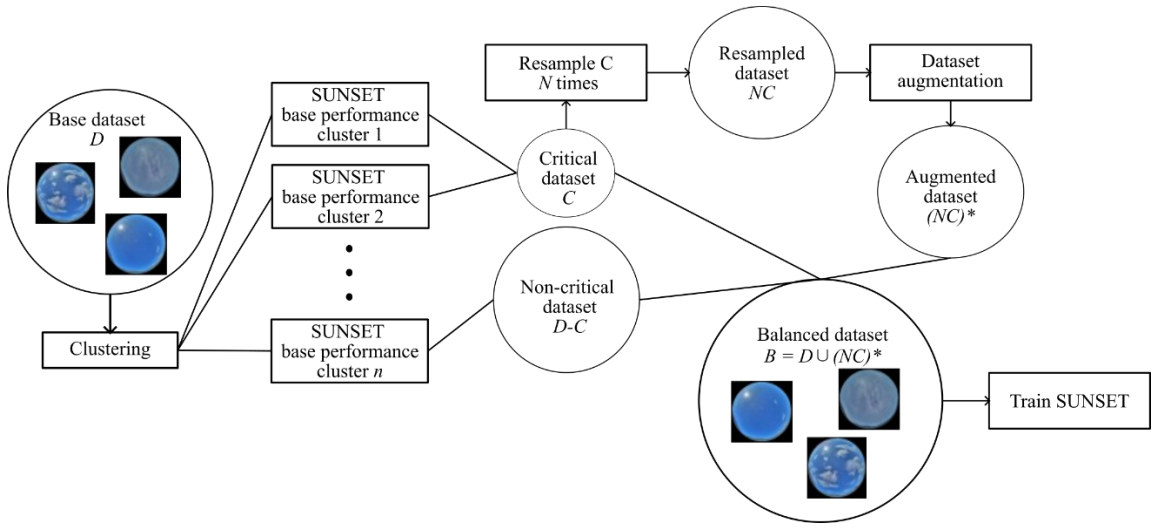


Figure 4. Schematic illustration of the proposed holistic sky image dataset augmentation framework for PV power nowcasting under data-scarce settings.

 doi:10.20944/preprints202604.1794.v1

In the in-series approach, two augmentation methods are applied sequentially to the images of the critical dataset to create N copies:

$$\tilde{x}_i^j = \lambda_2\left(\lambda_1 x_{p,i} + (1-\lambda_1)x_{q,i}\right) + (1-\lambda_2)\tilde{x}_{q,i}, \qquad \forall\, j \in [1,N] \tag{8}$$

where $\tilde{x}_{q,i}$ is the i^{th} pixel of the selected neighbor image $\tilde{\boldsymbol{x}}_q$ corresponding to the augmented original image $\tilde{\boldsymbol{x}}_p$, and λ_1, λ_2 are the interpolation coefficients associated with the two selected augmentation methods applied in series. The in-series approach is also expected to further increase dataset diversity, as sky images undergo two successive augmentation steps, leading to synthetic samples that differ more substantially from the original images. Figure 5 depicts an augmentation example for the *Archon* dataset of the proposed in-series hybrid approach, combining Mixup-kNN with Mixup-RP.

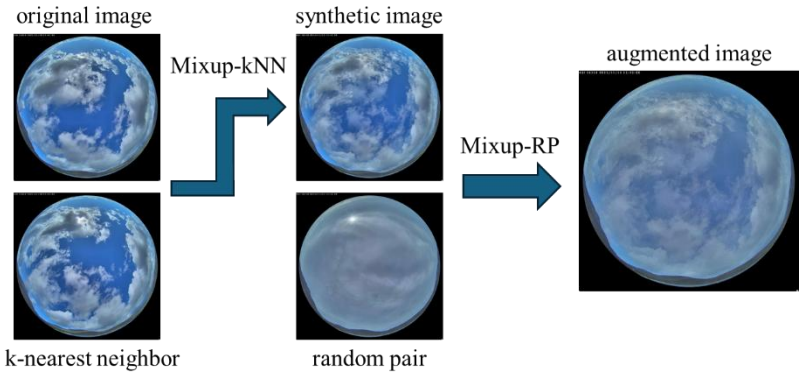


Figure 5. Sky image augmentation example of the proposed in-series hybrid approach, combining Mixup-kNN ($\alpha = 0.6$) and Mixup-RP ($\alpha = 0.8$) for the *Archon* dataset.

3. Results

3.1. Experimental Setup

All methods and models were implemented in Python, and all experiments were conducted on an Intel(R) Core (TM) i7-8700 CPU (3.20GHz, 6 cores) desktop computer with 8 GB of RAM. SUNSET is trained by minimizing the Mean Square Error (MSE) using the Adam optimization algorithm [31], with a learning rate of 3×10^{-6}, a batch size of 256, and a maximum of 100 epochs. The augmented datasets are split into training and validation subsets using a 70:30 ratio, with the validation subset serving for internal evaluation to identify critical sky image clusters, as well as to prevent overfitting during training. The interpolation coefficient λ is randomly sampled for each augmentation of every sky image of the critical dataset. For Mixup-kNN and Mixup-RP, the shape parameter α of the Beta distribution was set to 1 and 0.6, respectively, after finetuning using grid search over the interval [0, 1].

3.2. Clustering Results

Following [16], eight clusters are created during design time for both datasets. Figure 6 presents the clustering results for the *Archon* and *SKIPP'D* datasets, along with the relative proportion of each cluster within the base dataset and a representative sky image of each cluster. Qualitatively, clustering is driven by factors such as cloud coverage, cloud distribution, solar elevation, solar disk occlusion, turbulence levels, and raindrops appearance. As shown in Figure 6, sky images are not uniformly distributed across clusters, revealing inherent dataset imbalances. Clear sky or nearly clear sky conditions dominate both datasets, accounting for approximately 50% and 45% of the sky images in the *Archon* (clusters 1 and 2) and *SKIPP'D* (clusters 1 and 4) datasets, respectively. Overcast conditions (clusters 7 and 8) are well represented in the *Archon* dataset, comprising 26.52% of the sky images. The remaining clusters of the *Archon* dataset are relatively underrepresented, corresponding to scattered/broken clouds with varying cloud coverage and solar disk visibility, with shares ranging from 2% to 10%. In the *SKIPP'D* dataset, only cluster 5 corresponds to scattered/broken clouds,

 doi:10.20944/preprints202604.1794.v1

accounting for 14.48% of the sky images. Overcast conditions in the *SKIPP'D* dataset are represented by clusters 7 and 8, which mainly differ in the dominant color, with cluster 7 relatively well represented (12.98%) compared to the cluster 8 (1.4%). The *SKIPP'D* dataset also includes clusters primarily associated with low solar elevation (cluster 2 – 10.44%), high turbulence levels (cluster 3 – 12.34%), and clear sky conditions following rainfall (cluster 6 – 3.57%).

3.3. SUNSET Per-cluster Base Nowcasting Performance

Table 1 presents the per-cluster nowcasting RMSE achieved by SUNSET using the base training datasets. Specifically, SUNSET is trained separately on each cluster using the corresponding training subset, and the RMSE is calculated using the respective validation subset. For both datasets, RMSE values remain relatively high across all clusters – no lower than 8.75% and 7.1% of the nominal PV capacity for the *Archon* and *SKIPP'D* datasets, respectively. This is primarily due to the limited number of training samples within each cluster, which is insufficient for effectively training a deep learning model such as SUNSET. In the *SKIPP'D* dataset, where most clusters contain slightly more sky images than in the *Archon* dataset, PV power nowcasting errors are generally lower. Nevertheless, despite the negative impact of data scarcity on training reliability, the results still provide useful insights into the sky conditions associated with the largest PV power nowcasting errors.

As expected, clusters corresponding to clear sky conditions (clusters 1 and 2 in the *Archon* dataset and clusters 1 and 4 in the *SKIPP'D* dataset) exhibit the lowest PV power nowcasting errors due to their low variability and relatively large volumes of training data, so these clusters are naturally non-critical for augmentation. In contrast, RMSE reaches values as high as 58% of the nominal PV capacity for certain clusters in both datasets; a result primarily driven by the combination of very limited training samples (fewer than 5% of the total sky images) and highly variable sky conditions. Overall, underrepresented clusters associated with high variability (clusters 4 and 5 in the *Archon* dataset and cluster 6 in the *SKIPP'D* dataset) lead to the highest errors. For the remaining clusters, RMSE varies depending on both cluster size and dominant sky conditions, ranging from 20% to 43.5% for the *Archon* dataset and from 9.4% to 40% for the *SKIPP'D* dataset.

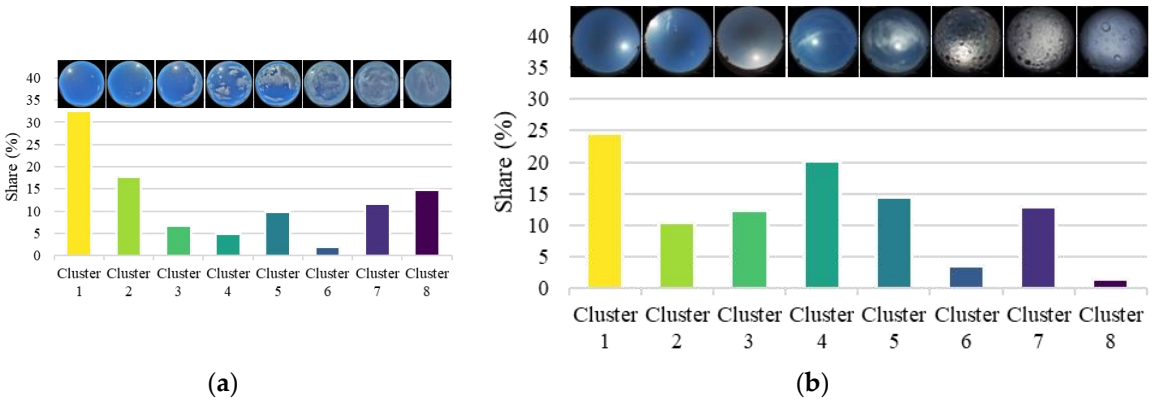


Figure 6. Clustering results along with a representative image of each cluster for both sky image base datasets: (**a**) *Archon* dataset; (**b**) *SKIPP'D dataset*.

Table 1. SUNSET per-cluster nowcasting performance for both sky image base datasets. Clusters characterized as critical for augmentation are in bold.

Archon			*SKIPP'D*		
Cluster	**Samples**	**RMSE (kW)**	**Cluster**	**Samples**	**RMSE (kW)**
1	1877	0.17	1	2382	2.21
2	1023	0.105	**2**	**1009**	**12.04**
3	385	0.464	3	1192	5.95
4	**281**	**0.685**	4	1947	2.13
5	**565**	**0.62**	5	1399	8.04
6	118	0.248	**6**	**345**	**17.37**

 doi:10.20944/preprints202604.1794.v1

7	**674**	**0.522**	7	1254	6.04
8	859	0.254	8	135	2.82

Based on the results of Table 1, the criticality coefficients r_1 and r_2 in (4) are set to 0.15 and 0.4, respectively. Accordingly, a cluster is characterized as critical for augmentation if it contains fewer than 15% of the total sky images and is associated with PV power nowcasting RMSE values exceeding 40% of the installed PV capacity. Nevertheless, r_1 and r_2 can be easily adjusted to reflect the specific characteristics of other datasets, the selected error metric ε_i, and the desired sensitivity to nowcasting errors. Based on these thresholds, clusters 4, 5, and 7 are characterized as critical in the *Archon* dataset (26.29% of the total dataset), and clusters 2 and 6 for the *SKIPP'D* dataset (14.01% of the total dataset). In the *Archon* dataset, clusters 4 and 5 correspond to scattered cloudy conditions, whereas cluster 7 represents overcast conditions; thus, augmenting the sky images of these clusters will improve the representation of both partly cloudy and overcast conditions in the final balanced dataset. Although clusters 3 and 6 are also underrepresented, their associated PV power nowcasting errors are comparatively lower, making them less critical for augmentation. In the *SKIPP'D* dataset, both clusters characterized as critical correspond to challenging conditions for PV power nowcasting. Cluster 2 represents sky images with low solar elevation, characterized by high color and PV power variability, whereas cluster 6 represents clear or nearly clear sky conditions with residual raindrops on the camera lens, indicating recent rainfall. Cluster 8, although underrepresented and associated with overcast conditions, is not considered critical for augmentation due to its relatively low nowcasting error; moreover, similar conditions are already represented by cluster 7.

3.4. PV Power Nowcasting Performance with Sky Image Augmentation

Preliminary experiments revealed that resampling the critical clusters $N = 7$ and $N = 14$ times results in optimal performance for the *Archon* and *SKIPP'D* datasets, respectively. Following the procedure outlined in Section 2.6, the resampled dataset NC is then augmented once with each of the basic augmentation methods described in Section 2.3, as well as with the two proposed hybrid augmentation approaches introduced in Section 2.7. This process results in multiple distinct balanced datasets $D \cup (NC)^*$, on which SUNSET is trained for PV power nowcasting. Model performance is subsequently evaluated using the external testing subset (Section 2.4), which comprises the majority of the original data and represents previously unseen sky images.

Table 2. PV power nowcasting performance of SUNSET on the testing subset after training with different augmentations of both sky image base datasets. Results for the base and resampling datasets are included as baselines.

Training dataset	**Augmentation method**	*Archon* **RMSE (kW)**	*SKIPP'D* **RMSE (kW)**
Base (D)	–	0.1675	3.10
Resampling ($D \cup NC$)	–	0.1100	2.53
$D \cup (NC)^*$	SMOTE	0.0902	2.39
$D \cup (NC)^*$	Mixup-kNN	0.0905	2.16
$D \cup (NC)^*$	Mixup-RP	0.0990	2.45
$D \cup (NC)^*$	Hybrid (in-parallel)	0.0892	1.99
$D \cup (NC)^*$	Hybrid (in-series)	**0.0808**	**1.54**

Table 2 presents the PV power nowcasting performance of SUNSET on the testing subset after training on balanced datasets generated using five augmentation methods: SMOTE, Mixup-kNN, Mixup-RP, SMOTE in-parallel with Mixup-kNN, and Mixup-kNN followed by in-series augmentation with SMOTE. For reference, baseline results are also reported for training on the original dataset D and on the resampled dataset $D \cup NC$. When trained solely on the base dataset, SUNSET achieves RMSE values of 0.1675 kW and 3.10 kW for the *Archon* and *SKIPP'D* datasets, respectively, corresponding to 13.96% and 10.33% of their nominal PV capacities. Although this marks a significant improvement compared to the per-cluster results presented in Table 1 – primarily due to the larger size of the base dataset compared to the cluster sizes – it remains suboptimal for

 doi:10.20944/preprints202604.1794.v1

accurate PV power nowcasting using sky images. Notably, even simple resampling of the critical clusters leads to significant performance gains of 34.33% for *Archon* and 18.39% for *SKIPP'D*, reducing RMSE to 9.17% and 8.43% of the nominal PV capacities, respectively.

Sky image augmentation using SMOTE, Mixup-kNN, and Mixup-RP, further improves performance, leading to similar SUNSET results, with RMSE values ranging from 7.2% to 8.25% of the nominal PV capacity across both datasets. However, the best PV power nowcasting performance is achieved when SUNSET is trained on balanced datasets generated using the two proposed hybrid augmentation approaches. Note that Table 2 reports only the RMSE corresponding to the optimal in-series and in-parallel configurations, i.e., Mixup-kNN followed by SMOTE (in-series), Mixup-kNN applied in-parallel with Mixup-RP and $N_1 = 5$ (*Archon*), and SMOTE applied in-parallel with Mixup-kNN and $N_1 = 6$ (*SKIPP'D*). For both datasets, the proposed in-series augmentation approach delivers the best performance, reducing RMSE to 0.0808 kW and 1.54 kW for the *Archon* and *SKIPP'D* datasets, respectively. These values correspond to 6.73% and 5.13% of the nominal PV capacities, representing an approximate 50% nowcasting error reduction compared to training on the base dataset. Similarly, the proposed in-parallel augmentation approach improves PV power nowcasting performance by 46.75% and 35.81% compared to training on the *Archon* and *SKIPP'D* base datasets, respectively.

Table 3 presents the PV power nowcasting performance of SUNSET using different configurations of the hybrid in-series augmentation. In general, all combinations lead to improved performance compared to training on the base and resampled datasets. For both datasets, the best performance is achieved when Mixup-kNN is applied to sky images, followed by SMOTE. For the *Archon* dataset, comparable performance is also obtained when Mixup-RP is applied after Mixup-kNN, as well as when Mixup-RP is applied after SMOTE. For the *SKIPP'D* dataset, the remaining combinations result in similar performance, with RMSE values ranging from 5.83% to 6.07% of the nominal PV capacity.

Table 3. PV power nowcasting performance of SUNSET on the testing subset after training with different augmentations of both sky image datasets using the proposed hybrid in-series approach.

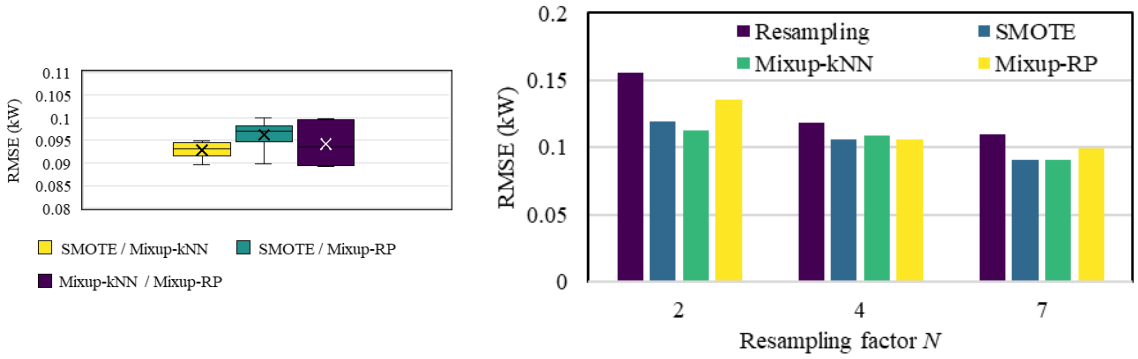
Hybrid in-series augmentation method	*Archon* **RMSE (kW)**	*SKIPP'D* **RMSE (kW)**
SMOTE → Mixup-kNN	0.0902	1.76
SMOTE → Mixup-RP	0.0876	1.82
Mixup-kNN → SMOTE	**0.0808**	**1.54**
Mixup-kNN → Mixup-RP	0.0817	1.80
Mixup-RP → SMOTE	0.0958	1.77
Mixup-RP → Mixup-kNN	0.1001	1.75

Figure 7(a) presents a boxplot of the PV power nowcasting performance of SUNSET using different configurations of the hybrid in-parallel augmentation on the *Archon* dataset. In this setup, all combinations of SMOTE, Mixup-kNN, and Mixup-RP are evaluated across different dataset splits. Specifically, from the seven resampled copies of the critical dataset, N_1 are augmented using one method, while the remaining are augmented using the other, with $N_1 \in \{1, 2, 3, 4, 5, 6\}$. It is observed that the SMOTE / Mixup-RP in-parallel combination results in the lowest RMSE, but also exhibits the highest variance, with values ranging from 0.0892 kW and 0.0998 kW. In contrast, consistent with the results of the *SKIPP'D* dataset, the SMOTE / Mixup-kNN in-parallel combination results in the most stable performance and achieves a minimum RMSE of 0.0896 kW, which is very close to the overall optimum. Consequently, SMOTE / Mixup-kNN can be considered as the most robust in-parallel augmentation approach across both datasets.

Figure 7(b) presents the PV power nowcasting performance of SUNSET on the *Archon* dataset using SMOTE, Mixup-kNN, and Mixup-RP, under different resampling factors $N \in \{2, 4, 7\}$. In general, the RMSE decreases as the resampling factor increases, highlighting the beneficial effect of mitigating data scarcity and data imbalances on overall nowcasting performance. Beyond $N = 7$, performance improvements were marginal and thus did not justify further increases in the resampling factor. Among the evaluated augmentation methods, Mixup-RP exhibits the most

 doi:10.20944/preprints202604.1794.v1

pronounced performance gains with increasing the resampling factor. This may be attributed to its inherent formulation, as it generates synthetic samples by randomly combining images, thereby promoting higher diversity but also introducing additional noise. As the number of generated samples increases, this noise is progressively mitigated through averaging effects, leading to more stable and accurate model performance.



(**a**) (**b**)

Figure 7. PV power nowcasting performance of SUNSET on the *Archon* testing subset using: (**a**) different in-parallel hybrid augmentations with varying splits, and (**b**) SMOTE, Mixup-kNN, and Mixup-RP, with varying resampling factors.

4. Discussion

The experimental results highlight the strong influence of both the size and distribution of historical datasets on PV power nowcasting using sky images, as well as the feasibility of achieving reliable performance under data-scarce conditions given appropriate pre-processing. Under such conditions, even simple resampling of specific data clusters leads to significant performance improvements. On one hand, targeted resampling mitigates inherent dataset imbalances by enhancing the representation of sky images associated with higher PV power nowcasting errors. This enables a deep CNN-based model like SUNSET to learn more uniformly across diverse sky conditions during training and to generalize more effectively when applied to unseen data. On the other hand, models like SUNSET comprise millions of trainable parameters and therefore require sufficiently large datasets for effective training. In this context, increasing the training set size, even by duplicating existing historical samples, can be beneficial when data are limited. These effects are demonstrated throughout this study, from the comparison between the per-cluster performance and the aggregated baseline performance to the improvements achieved by increasing the resampling ratio.

Another key insight from the experimental results concerns the value of data augmentation. Regardless of the augmentation method, SUNSET consistently performs better when trained on augmented datasets than on merely resampled data. By generating synthetic sky images, augmentation increases dataset diversity and, thus, improves CNN generalization under data-scarce conditions. Overall, Mixup-kNN and SMOTE leads to slightly better results compared to Mixup-RP, as they generate synthetic samples by combining neighboring sky images; thus, the created samples are more realistic and introduce less additional noise. Notably, it is the proposed hybrid augmentation approaches that deliver the best PV power nowcasting performance. Their advantage lies not only in enhancing dataset diversity but also in leveraging the complementary strengths of individual augmentation methods, thereby mitigating noise propagation. For in-parallel augmentation, the Mixup-kNN / SMOTE combination provides the most robust performance for both datasets, as it integrates two neighbor-based methods with different sampling distributions, producing diverse yet realistic synthetic sky images. Similarly, in the in-series setting, applying SMOTE after Mixup-kNN exhibits superior performance. Sampling from a Beta distribution, Mixup-kNN first generates a diverse layer of intermediate images, on top of which SMOTE subsequently

 doi:10.20944/preprints202604.1794.v1

performs refining augmentation by subtly mixing similar images using a uniform distribution, leading to more stable and effective augmentation.

It is important to emphasize that this paper explores the potential of accurate PV power nowcasting using sky images under data-scarce conditions, within the broader context of ultra-short-term PV power forecasting using AI and deep learning. As distributed PV systems and data-driven models continue to expand, data abundance cannot always be assumed; thus, developing methods that remain robust under limited data is essential for the reliable integration of PV systems into modern power systems, particularly in terms of reducing operating costs, emissions, and power quality issues. In this context, the results demonstrate that accurate PV power nowcasting is achievable even in extreme data scarcity scenarios, e.g., for newly installed PV systems with only a few weeks of data, provided that appropriate pre-processing is applied. Nevertheless, it should be noted that, under data-abundant conditions, the effectiveness of the proposed framework and the associated insights may be less pronounced, and alternative approaches from the related literature should be also investigated.

5. Conclusions

This paper introduces a holistic sky image augmentation framework for PV power nowcasting using deep CNNs under data-scarce conditions. In the proposed framework, the original limited base dataset is initially clustered using a state-of-the-art sky image clustering method, and each cluster is characterized as critical or non-critical for augmentation based on its relative size and its associated average PV power nowcasting error. Clusters characterized as critical are then resampled and augmented using various data-driven image augmentation methods, along with novel hybrid in-series and in-parallel approaches. The proposed framework is evaluated using two distinct sky image datasets comprising several weeks of data. Experimental results demonstrate that accurate PV power nowcasting is feasible even under limited historical data availability, with the use of augmented datasets, created by the proposed augmentation methods, leading to reductions in average nowcasting error of up to 50%, driven by enhanced CNN generalization due to increased dataset balance, volume, and diversity.

Author Contributions: Conceptualization, M.A.K.-K., V.T. and P.S.G.; methodology, M.A.K.-K., V.T. and P.S.G.; software, M.A.K.-K., V.T. and A.P.G.; validation, M.A.K.-K., V.T., A.P.G. and P.S.G.; formal analysis, M.A.K.-K., V.T. and P.S.G.; investigation, M.A.K.-K., V.T. and A.P.G.; resources, M.A.K.-K. and P.S.G.; data curation, M.A.K.-K., V.T. and P.S.G.; writing—original draft preparation, M.A.K.-K; writing—review and editing, M.A.K.-K., V.T., A.P.G. and P.S.G.; visualization, M.A.K.-K., V.T. and A.P.G.; supervision, P.S.G.; project administration, P.S.G.; funding acquisition, P.S.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The *Archon* dataset will become available at the project's site http://archonproject.eu/english.html. The *SKIPP'D* dataset is available in the publicly accessible repository at https://purl.stanford.edu/jj716hx9049.

Acknowledgments: The authors acknowledge that one of the datasets used in this study was obtained from the *Archon* project, and we thank the project team for making this data available.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ASI	All Sky Imager
CNN	Convolutional Neural Network
FOV	Field Of View
GAN	Generative Adversarial Network
kNN	k-Nearest Neighbors

 doi:10.20944/preprints202604.1794.v1

MSE	Mean Square Error
PV	Photovoltaic
RES	Renewable Energy Sources
RMSE	Root Mean Square Error
RP	Random Pair
SMOTE	Synthetic Minority Oversampling Technique
SUNSET	Stanford University Neural network for Solar Electricity Trend
VT	Vision Transformer

References

1. Dritsaki, M.; Dritsaki, C.; Argyriou, V.; Sarigiannidis, P. Impact of Renewable and Non-Renewable Generation on Economic Growth in Greece. *Electr. J.* **2024**, *37*, 107421, doi: 10.1016/j.tej.2024.107421.
2. Fotis, G.; Maris, T.I.; Mladenov, V. Risks, Obstacles and Challenges of the Electrical Energy Transition in Europe: Greece as a Case Study. *Sustainability* **2025**, *17*, 5325, doi: 10.3390/su17125325.
3. Cavus, M. Advancing Power Systems with Renewable Energy and Intelligent Technologies: A Comprehensive Review on Grid Transformation and Integration. *Electronics.* **2025**, *14*, 1159, doi: 10.3390/electronics14061159.
4. Paletta, Q.; Terrén-Serrano, G.; Nie, Y.; Li, B.; Bieker, J.; Zhang, W.; Dubus, L.; Dev, S.; Feng, C. Advances in Solar Forecasting: Computer Vision with Deep Learning. *Adv. Appl. Energy* **2023**, *11*, 100150, doi: https://doi.org/10.1016/j.adapen.2023.100150.
5. Lin, F.; Zhang, Y.; Wang, J. Recent Advances in Intra-Hour Solar Forecasting: A Review of Ground-Based Sky Image Methods. *Int. J. Forecast.* **2023**, *39*, 244–265, doi: https://doi.org/10.1016/j.ijforecast.2021.11.002.
6. Sun, Y.; Venugopal, V.; Brandt, A.R. Short-Term Solar Power Forecast with Deep Learning: Exploring Optimal Input and Output Configuration. *Sol. Energy* **2019**, *188*, 730–741, doi: 10.1016/j.solener.2019.06.041.
7. Sun, Y.; Szűcs, G.; Brandt, A.R. Solar PV Output Prediction from Video Streams Using Convolutional Neural Networks. *Energy Environ. Sci.* **2018**, *11*, 1811–1818, doi: 10.1039/C7EE03420B.
8. Papatheofanous, E.A.; Kalekis, V.; Venitourakis, G.; Tziolos, F.; Reisis, D. Deep Learning-Based Image Regression for Short-Term Solar Irradiance Forecasting on the Edge. *Electronics* **2022**, *11*, 3794, doi: 10.3390/electronics11223794.
9. Venitourakis, G.; Vasilakis, C.; Tsagkaropoulos, A.; Amrou, T.; Konstantoulakis, G.; Golemis, P.; Reisis, D. Neural Network-Based Solar Irradiance Forecast for Edge Computing Devices. *Information* **2023**, *14*, 617, doi: 10.3390/info14110617.
10. Habibi, M.R.; Golestan, S.; Guerrero, J.M.; Vasquez, J.C. Deep Learning for Forecasting-Based Applications in Cyber–Physical Microgrids: Recent Advances and Future Directions. *Electronics* **2023**, *12*, 1685, doi: 10.3390/electronics12071685.
11. Nie, Y.; Zamzam, A.S.; Brandt, A. Resampling and Data Augmentation for Short-Term PV Output Prediction Based on an Imbalanced Sky Images Dataset Using Convolutional Neural Networks. *Sol. Energy* **2021**, *224*, 341–354, doi: 10.1016/j.solener.2021.05.095.
12. Nie, Y.; Paletta, Q.; Scott, A.; Pomares, L.M.; Arbod, G.; Sgouridis, S.; Lasenby, J.; Brandt, A. Sky Image-Based Solar Forecasting Using Deep Learning with Heterogeneous Multi-Location Data: Dataset Fusion versus Transfer Learning. *Appl. Energy* **2024**, *369*, 123467, doi: 10.1016/j.apenergy.2024.123467.
13. Nie, Y.; Sun, Y.; Chen, Y.; Orsini, R.; Brandt, A. PV Power Output Prediction from Sky Images Using Convolutional Neural Network: The Comparison of Sky-Condition-Specific Sub-Models and an End-to-End Model. *J. Renew. Sustain. Energy* **2020**, *12*, 046101, doi: 10.1063/5.0014016.
14. Kousounadis-Knousen, M.A.; Catthoor, F.; Bakovasilis, A.; Georgilakis, P.S. Automatic Multiclass Classification of Unlabeled Ground-Based Sky Images for Minute-Scale PV Energy Yield Forecasting. *IEEE Access* **2025**, *13*, 120547–120562, doi: 10.1109/ACCESS.2025.3587059.
15. Meddahi, A.; Tuomiranta, A.; Guillon, S. Skill-Driven Data Sampling and Deep Learning Framework for Minute-Scale Solar Forecasting with Sky Images. *Solar RRL* **2025**, *9*, doi: 10.1002/solr.202400664.
16. Schizas, S.P.; Kousounadis-Knousen, M.A.; Catthoor, F.; Georgilakis, P.S. Multi-Step Sky Image Prediction Using Cluster-Specific Convolutional Neural Networks for Solar Forecasting Applications. *Energies* **2025**, *18*, 5860, doi: 10.3390/en18215860.

 doi:10.20944/preprints202604.1794.v1

17. Moreno-Barea, F.J.; Strazzera, F.; Jerez, J.M.; Urda, D.; Franco, L. Forward Noise Adjustment Scheme for Data Augmentation. In Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI); IEEE, November 2018; pp. 728–734.
18. Wu, R.; Yan, S.; Shan, Y.; Dang, Q.; Sun, G. Deep Image: Scaling up Image Recognition. *arXiv* **2015**, doi: https://doi.org/10.48550/arXiv.1501.02876
19. Shorten, C.; Khoshgoftaar, T.M. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60, doi: 10.1186/s40537-019-0197-0.
20. Chawla, N. V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357, doi: 10.1613/jair.953.
21. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond Empirical Risk Minimization. *arXiv* **2017**, doi: https://arxiv.org/abs/1710.09412.
22. Chen, Y.; Yang, X.-H.; Wei, Z.; Heidari, A.A.; Zheng, N.; Li, Z.; Chen, H.; Hu, H.; Zhou, Q.; Guan, Q. Generative Adversarial Networks in Medical Image Augmentation: A Review. *Comput. Biol. Med.* **2022**, *144*, 105382, doi: 10.1016/j.compbiomed.2022.105382.
23. Go, S.-E.; Kim, J.-H.; Chuluunsaikhan, T.; Choi, W.-S.; Choi, S.-H.; Nasridinov, A. Unified Generative Data Augmentation for Efficient Solar Panel Soiling Localization. *Electronics* **2024**, *13*, 4859, doi: 10.3390/electronics13244859.
24. Paletta, Q.; Hu, A.; Arbod, G.; Blanc, P.; Lasenby, J. SPIN: Simplifying Polar Invariance for Neural Networks Application to Vision-Based Irradiance Forecasting. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); IEEE, June 2022; pp. 5178–5187.
25. Gentner, T.; Knoell, M.; Adam, J.; Theissler, A.; Klaiber, M. Sky Is the Limit: Exploring Solar Photovoltaic Nowcasting with Sky Images, Transfer Learning, and Data Augmentation. *Procedia Comput. Sci.* **2025**, *270*, 1649–1658, doi: 10.1016/j.procs.2025.09.285.
26. Piechocki, M.; Kraft, M. A Systematic Synthesis of Sky Image Enhancement Techniques for Ground-Based Solar Irradiance Forecasting. *Appl. Energy* **2026**, *410*, 127533, doi: 10.1016/j.apenergy.2026.127533.
27. Archon Project. Available online: http://archonproject.eu/english.html (accessed on 20 February 2026).
28. Nie, Y.; Li, X.; Scott, A.; Sun, Y.; Venugopal, V.; Brandt, A. 2019 Sky Images and Photovoltaic Power Generation Dataset for Short-Term Solar Forecasting (Stanford Raw). Stanford Digital Repository 2022. Available at https://Purl.Stanford.Edu/Jj716hx9049.
29. Kärkkäinen, T.J.; Hänninen, J. Additive Autoencoder for Dimension Estimation. *Neurocomputing* **2023**, *551*, 126520.
30. MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, January 1967, pp. 281–297.
31. Kingma, D.P. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, doi: https://doi.org/10.48550/arXiv.1412.6980.