

Multichannel autocalibration for the Atmospheric Imaging Assembly using machine learning

Luiz F. G. Dos Santos^{1,2}, Souvik Bose^{3,4}, Valentina Salvatelli^{5,6}, Brad Neuberg^{5,6}, Mark C. M. Cheung⁷, Miho Janvier⁸, Meng Jin^{6,7}, Yarin Gal⁹, Paul Boerner⁷, and Atılım Güneş Baydin^{10,11}

¹ Heliophysics Science Division, NASA, Goddard Space Flight Center, Greenbelt, MD 20771, USA

² The Catholic University of America, Washington, DC 20064, USA

e-mail: 51guedesdosantos@cua.edu

³ Rosseland Center for Solar Physics, University of Oslo, PO Box 1029 Blindern, 0315 Oslo, Norway

⁴ Institute of Theoretical Astrophysics, University of Oslo, PO Box 1029 Blindern, 0315 Oslo, Norway

⁵ Frontier Development Lab, Mountain View, CA 94043, USA

⁶ SETI Institute, Mountain View, CA 94043, USA

⁷ Lockheed Martin Solar & Astrophysics Laboratory (LMSAL), Palo Alto, CA 94304, USA

⁸ Université Paris-Saclay, CNRS, Institut d'astrophysique spatiale, Orsay, France

⁹ OATML, Department of Computer Science, University of Oxford, Oxford OX1 3QD, UK

¹⁰ Department of Computer Science, University of Oxford, Oxford OX1 3QD, UK

¹¹ Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK

Received 2 December 2020 / Accepted 29 January 2021

ABSTRACT

Context. Solar activity plays a quintessential role in affecting the interplanetary medium and space weather around Earth. Remote-sensing instruments on board heliophysics space missions provide a pool of information about solar activity by measuring the solar magnetic field and the emission of light from the multilayered, multithermal, and dynamic solar atmosphere. Extreme-UV (EUV) wavelength observations from space help in understanding the subtleties of the outer layers of the Sun, that is, the chromosphere and the corona. Unfortunately, instruments such as the Atmospheric Imaging Assembly (AIA) on board the NASA Solar Dynamics Observatory (SDO), suffer from time-dependent degradation that reduces their sensitivity. The current best calibration techniques rely on flights of sounding rockets to maintain absolute calibration. These flights are infrequent, complex, and limited to a single vantage point, however.

Aims. We aim to develop a novel method based on machine learning (ML) that exploits spatial patterns on the solar surface across multiwavelength observations to autocalibrate the instrument degradation.

Methods. We established two convolutional neural network (CNN) architectures that take either single-channel or multichannel input and trained the models using the SDOML dataset. The dataset was further augmented by randomly degrading images at each epoch, with the training dataset spanning nonoverlapping months with the test dataset. We also developed a non-ML baseline model to assess the gain of the CNN models. With the best trained models, we reconstructed the AIA multichannel degradation curves of 2010–2020 and compared them with the degradation curves based on sounding-rocket data.

Results. Our results indicate that the CNN-based models significantly outperform the non-ML baseline model in calibrating instrument degradation. Moreover, multichannel CNN outperforms the single-channel CNN, which suggests that cross-channel relations between different EUV channels are important to recover the degradation profiles. The CNN-based models reproduce the degradation corrections derived from the sounding-rocket cross-calibration measurements within the experimental measurement uncertainty, indicating that it performs equally well as current techniques.

Conclusions. Our approach establishes the framework for a novel technique based on CNNs to calibrate EUV instruments. We envision that this technique can be adapted to other imaging or spectral instruments operating at other wavelengths.

Key words. Sun: UV radiation – techniques: image processing – telescopes – Sun: activity – methods: data analysis

1. Introduction

Solar activity plays a significant role in affecting the interplanetary medium and space weather around Earth and all the other planets of the Solar System (Schwenn 2006). Remote-sensing instruments on board heliophysics missions can provide a wealth of information about solar activity, primarily by capturing the emission of light from the multilayered solar atmosphere, thereby leading to the inference of various physical quantities such as magnetic fields, plasma velocities, temperature, and emission measure.

NASA currently manages the Heliophysics System Observatory (HSO), which consists of a group of satellites that constantly monitor the Sun, its extended atmosphere, and the space environments around Earth and other planets of the Solar System (Clarke 2016). One of the flagship missions of HSO is the Solar Dynamics Observatory (SDO, Pesnell et al. 2012). Launched in 2010, SDO has been instrumental in monitoring solar activity and providing a high volume of valuable scientific data every day with a high temporal and spatial resolution. It has three instruments on board: the Atmospheric Imaging Assembly (AIA, Lemen et al. 2012), which records images with

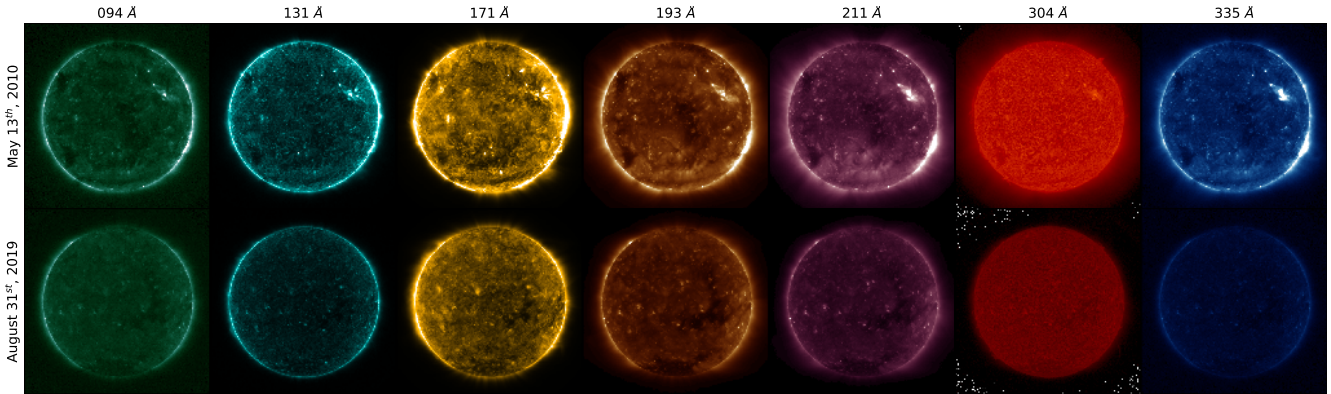


Fig. 1. Set of images to exemplify how degradation affects the AIA channels. The two sets are composed of seven images from different EUV channels. *From left to right:* AIA 94 Å, AIA 131 Å, AIA 171 Å, AIA 193 Å, AIA 211 Å, AIA 304 Å, and AIA 335 Å. *Top row:* images from 13 May 2010, and *bottom row:* images from 31 August 2019, without correction for degradation. The 304 Å channel images are in log-scale because the degradation is severe.

high spatial and temporal resolution of the Sun in the ultraviolet (UV) and extreme UV (EUV); the Helioseismic and Magnetic Imager (HMI, Schou et al. 2012), which provides maps of the photospheric magnetic field, solar surface velocity, and continuum filtergrams; and the EUV Variability Experiment (EVE, Woods et al. 2012), which measures the solar EUV spectral irradiance.

Over the past decade, SDO has played a central role in advancing our understanding of the fundamental plasma processes governing the Sun and space weather. This success can mainly be attributed to its open-data policy and a consistent high data-rate of approximately two terabytes of scientific data per day. The large volume of data accumulated over the past decade (over 12 petabytes) provides a fertile ground for developing and applying novel machine learning (ML) based data-processing methods. Recent studies, such as the prediction of solar flares from HMI vector magnetic fields (Bobra & Couvidat 2015), creation of high-fidelity virtual observations of the solar corona (Salvatelli et al. 2019 and Cheung et al. 2019), a forecast of far-side magnetograms from the Solar Terrestrial Relations Observatory (STEREO, Kaiser et al. 2008), EUV images (Kim et al. 2019), super-resolution of magnetograms (Jungbluth et al. 2019), and a map of EUV images from AIA to spectral irradiance measurements (Szenicer et al. 2019) have demonstrated the immense potential of ML applications in solar and heliophysics. In this paper, we use the availability of such high-quality continuous observations from SDO and apply ML techniques to address the instrument calibration problem.

One of the crucial issues that limit the diagnostic capabilities of the SDO-AIA mission is the degradation of sensitivity over time. Sample images from the seven AIA EUV channels in Fig. 1 show an example of this deterioration. The top row shows the images observed during the early days of the mission, from 13 May 2010, and the bottom row shows the corresponding images observed more recently on 31 August 2019, scaled within the same intensity range. The images in the bottom row clearly appear to be significantly dimmer than their top row counterparts. In some channels, especially 304 Å and 335 Å the effect is pronounced.

The dimming effect observed in the channels is due to the temporal degradation of EUV instruments in space that is also known to diminish the overall instrument sensitivity with time (e.g., BenMoussa et al. 2013). The possible causes include either the outgassing of organic materials in the telescope struc-

ture, which may deposit on the optical elements (Jiao et al. 2019), or the decrease in detector sensitivity due to exposure to EUV radiation from the Sun.

In general, first-principle models predicting the sensitivity degradation as functions of time and wavelength are not sufficiently well constrained to maintain the scientific calibration of these instruments. To circumvent this problem, instrument scientists have traditionally relied on empirical techniques, such as considering sources with known fluxes, the so-called standard candles. However, no standard candles exist in the solar atmosphere at these wavelengths because the solar corona is continuously driven and structured by evolving magnetic fields, which cause localized and intermittent heating. This causes even the quiet-Sun brightness in the EUV channels to vary significantly, depending on the configuration of the small-scale magnetic fields (Shakeri et al. 2015, and the references therein). On the one hand, the Sun may not be bright enough to appear in the hotter EUV channels such as AIA 94 Å. On the other hand, the EUV fluxes of active regions (ARs) can vary by several orders of magnitude, depending on whether the AR is in an emerging, flaring, or decaying state. Moreover, the brightness depends on the complexity of the AR magnetic field (van Driel-Gesztelyi & Green 2015). Finally, ARs in the solar corona can evolve on timescales ranging from a few minutes to several hours, leading to obvious difficulties in obtaining a standard flux for the purpose of calibration.

The current state-of-art methods for compensating for this degradation rely on cross-calibration between AIA and EVE instruments. The calibrated measurement of the full-disk solar spectral irradiance from EVE is passed through the AIA wavelength (filter) response function to predict the integrated AIA signal over the full field of view. The predicted band irradiance is later compared with the actual AIA observations (Boerner et al. 2014). The absolute calibration of SDO-EVE is maintained through periodic sounding-rocket experiments (Wieman et al. 2016) that use a near-replica of the instrument on board SDO to gather a calibrated observation that spans the short interval of the suborbital flight (lasting a few minutes). A comparison of the sounding-rocket observation with the satellite instrument observation provides an updated calibration, revealing long-term trends in the sensitivities of EVE and thus of AIA.

Sounding rockets are undoubtedly crucial; however, the sparse temporal coverage (there are flights roughly every two years) and the complexities of intercalibration are also potential

sources of uncertainty in the interinstrument calibration. Moreover, the intercalibration analysis has long latencies of months and sometimes years between the flights and depending on the times at which the calibration can be updated based on the data analysis of the data obtained during the flight; this type of calibration is also limited to observations from Earth and thus cannot easily be used to calibrate missions in deep space (e.g., STEREO).

In this paper, we focus on automating the correction of the sensitivity degradation of different AIA wavebands by exclusively using AIA information and adopting a deep neural network (DNN, Goodfellow et al. 2006) approach, which exploits the spatial patterns and cross-spectral correlations of the observed solar features in multiwavelength observations of AIA. We compare our approach with a non-ML method motivated by solar physics heuristics, which we call the baseline model. We evaluate the predicted degradation curves with those obtained through the sounding rocket cross-calibration described above. To the best of our knowledge, this is the first attempt to develop a calibration method of this type¹. The approach developed in this work may be able to remove a major impediment for developing future HSO missions that can deliver solar observations from different vantage points beyond the Earth orbit.

The paper is structured as follows: in Sect. 2 we present and describe our dataset. In Sect. 3 we illustrate the technique and how it has been developed. In Sect. 3.1 we state the hypothesis and propose a formulation of the problem, in Sect. 3.2 we present the CNN models, in Sect. 3.3 we describe the training process and the evaluation, in Sect. 3.4 we probe the multichannel relation, and in Sect. 3.5 we reconstruct the temporal degradation curve. Furthermore, in Sect. 4 we present the baseline, followed by Sect. 5, in which we present and discuss the results. Concluding remarks are given in Sect. 6.

2. Data description and preprocessing

We used the preprocessed SDO-AIA dataset from Galvez et al. (2019, hereafter referred to as SDOML). This dataset is ML-ready to be used for any kind of application related to the AIA and HMI data, and it consists of a subset of the original SDO data that covers from 2010 to 2018. It comprises the seven EUV channels, two UV channels from AIA, and vector magnetograms from HMI. The data from the two SDO instruments are temporally aligned, with cadences of 6 min for AIA (instead of the original 12 s) and EVE and 12 min for HMI. The full-disk images are downsampled from 4096×4096 to 512×512 pixels and have an identical spatial sampling of $\sim 4''/8$ per pixel.

In SDOML, the AIA images are compensated for the exposure time and corrected for instrumental degradation over time using piecewise-linear fits to the V8 corrections released by the AIA team in November 2017². These corrections are based on cross-calibration with SDO-EVE, where the EVE calibration is maintained by periodic sounding rocket under flight (including,

¹ We presented an early-stage result of this work as an extended abstract at the NeurIPS workshop on ML and Physical Sciences 2019 (which has no formal proceedings) (NeurIPS 2019, Neuberg et al. 2019), where we described some preliminary results in this direction. In this paper, we extend the abstract with full analyses and discussion of several important issues, such as the performance on the real degradation curve and the limitations of the presented models, which are both crucial for evaluating the applicability of this ML-based technique.

² Available at https://aiapy.readthedocs.io/en/stable/generated/gallery/instrument_degradation.html#sphx-glr-generated-gallery-instrument-degradation-py

in the case of the V8 corrections, a flight on 1 June 2016). Consequently, the resulting dataset offers images where changes in pixel brightness are directly related to the state of the Sun rather than instrument performance.

We applied a few additional preprocessing steps. First, we downsampled the SDOML dataset to 256×256 pixels from 512×512 pixels. We established that 256×256 is a sufficient resolution for the predictive task of interest (inference of a single coefficient), and the reduced size enabled quicker processing and more efficient use of the computational resources. Second, we masked the off-limb signal ($r > R_{\odot}$) to avoid possible contamination due to the telescope vignetting. Finally, we rescaled the brightness intensity of each AIA channel by dividing the image intensity by a channel-wise constant factor. These factors represent the approximate average AIA data counts in each channel and in the period from 2011 to 2018 (derived from Galvez et al. 2019). This rescaling is implemented to set the mean pixel values close to unity in order to improve the numerical stability and the training convergence of the CNN. Data normalization such as this is standard practice in NNs (Goodfellow et al. 2006). The specific values for each channel are reported in Appendix A.

3. Method

3.1. Formulation of the problem

It is known that some bright structures in the Sun are observed at different wavelengths. Figure 2 shows a good example from 07 April 2015 of a bright structure in the center of all seven EUV channels from AIA. Based on this cross-channel structure, we established a hypothesis divided into two parts. First, that the morphological features and the brightness of solar structures in a single channel are related (e.g., typically, dense and hot loops over ARs). Second, that this relation of the morphological features and the brightness of solar structures can be found in multiple AIA channels. We hypothesize that these two relations can be used to estimate the dimming factors, and that a deep-learning model can automatically learn these inter- and cross-channel patterns and exploit them to accurately predict the dimming factor of each channel.

To test our hypothesis, we considered a vector $\mathbf{C} = \{C_i, i \in [1, \dots, n]\}$ of multichannel synchronous SDO/AIA images, where C_i denotes the i -th channel image in the vector, and a vector $\alpha = \{\alpha_i, i \in [1, \dots, n]\}$, where α_i is the dimming factor that is independently sampled from the continuous uniform distribution between $[0.01, 1.0]$. We chose an upper bound value of $\alpha_i = 1$ because we only considered dimming of the images and not enhancements. Furthermore, we created a corresponding vector of dimmed images as $\mathbf{D} = \{\alpha_i C_i, i \in [1, \dots, n]\}$, where \mathbf{D} is the corresponding dimmed vector. It is also to be noted that the dimming factors α_i were applied uniformly per channel and are not spatially dependent. The spatial dependence of the degradation is assumed to be accounted for by regularly updated flat-field corrections applied to AIA images. Our goal in this paper is to find a deep learning model $M : \mathbf{D} \rightarrow \alpha$ that retrieves the vector of multichannel dimming factors α from the observed SDO-AIA vector \mathbf{D} .

3.2. Convolutional neural network model

Deep learning is a highly active subfield of ML that focuses on specific models called deep neural networks (DNNs). A DNN is a composition of multiple layers of linear transformations and nonlinear element-wise functions (Goodfellow et al. 2006). One

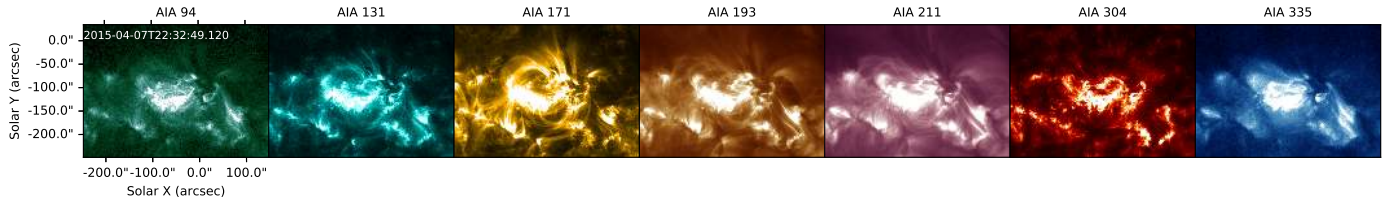


Fig. 2. Colocated set of images of the seven EUV channels of AIA to exemplify structures that are observed at different wavelengths. *From left to right:* AIA 94 Å, AIA 131 Å, AIA 171 Å, AIA 193 Å, AIA 211 Å, AIA 304 Å, and AIA 335 Å.

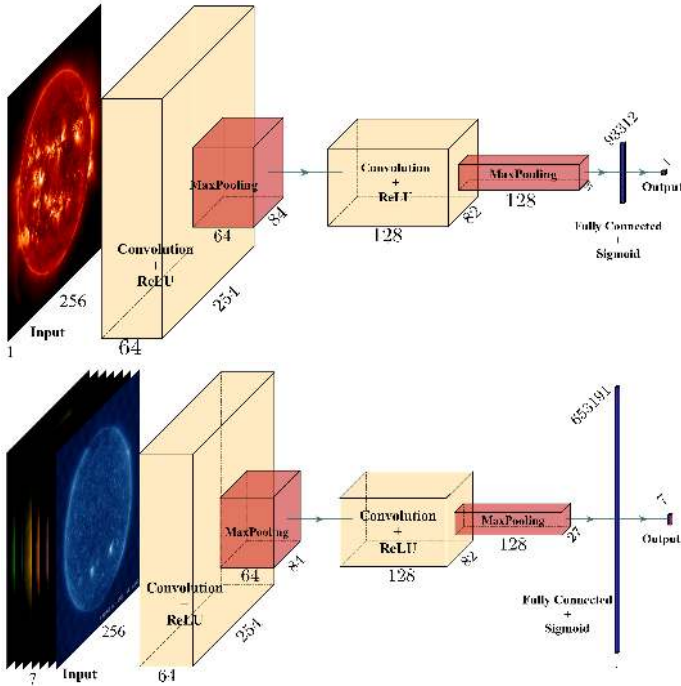


Fig. 3. CNN architectures. The single-channel architecture with a single wavelength input that is composed of two blocks of a convolutional layer is shown at the *top*, with the ReLU activation function and max pooling layer, followed by a fully connected (FC) layer and a final sigmoid activation function. The multichannel architecture with a multi-wavelength input that is composed of two blocks of a convolutional layer and max pooling layer, followed by an FC layer and a final sigmoid activation function. Figures constructed following Iqbal (2018).

of the main advantages of deep learning is that it can learn the best feature representation for a given task from the data without the need to manually engineer these features. DNNs have produced the state-of-art results in many complex tasks, including object detection in images (He et al. 2016), speech recognition (Amodei et al. 2016) and synthesis (Oord et al. 2016), and translation between languages (Wu et al. 2016). A DNN expresses a differentiable function $F_{\theta} : \mathcal{X} \rightarrow \mathcal{Y}$ that can be trained to perform complex nonlinear transformations by tuning parameters θ using gradient-based optimization of a loss function (also known as objective or error) $L(\theta) = \sum_i l(F_{\theta}(x_i), y_i)$ for a given set of inputs and desired outputs $\{x_i, y_i\}$.

For the degradation problem summarized in Sect. 3.1, we considered two CNN architectures (LeCun & Bengio 1995). The first architecture does not exploit the spatial dependence of multichannel AIA images and therefore ignores any possible relation that different AIA channels might have, and it is designed to explore only the relation in different structures in a single chan-

nel. This architecture is a test of the first hypothesis in Sect. 3.1. The second architecture is instead designed to exploit possible cross-channel relations while training, and it tests our second hypothesis: solar surface features that appear in the different channels will make a multichannel CNN architecture more effective than a single-channel CNN that only exploits interchannel structure correlations. The first model considers a single channel as input in the form of a tensor with shape $1 \times 256 \times 256$ and has a single degradation factor α as output. The second model takes in multiple AIA channel images simultaneously as input with shape $n \times 256 \times 256$ and output n degradation factors $\alpha = \{\alpha_i, i \in [1, \dots, n]\}$, where n is the number of channels, as indicated in Fig. 3.

The single- and multichannel architectures are described in Fig. 3. They both consist of two blocks of a convolutional layer followed by a rectified linear unit (ReLU) activation function (Nair & Hinton 2010) and a max pooling layer. These are followed by a fully connected (FC) layer and a final sigmoid activation function that is used to output the dimming factors. The first convolution block has 64 filters, and the second convolution block has 128 filters. In both convolution layers, the kernel size is 3, meaning that the filters applied on the image are 3×3 pixels, and the stride is 1, meaning that the kernel slides through the image one pixel per step. No padding is applied (i.e., no additional pixels are added at the border of the image to avoid a change in size). The resulting total learnable parameters (LP) are 167 809 for the single-channel model and 731 143 for the multichannel model. The final configurations of the model architectures were obtained through a grid search of different hyperparameters and layer configurations. More details of the architectures can be found in Appendix B.

We used the open-source software library PyTorch (Paszke et al. 2017) to implement the training and inference code for the CNN. The source code that we used to produce this paper is publicly available³.

3.3. Training process

The actual degraded factors $\alpha_i(t)$ (where t is the time since the beginning of the SDO mission, and i is the channel) trace a single trajectory in an n -dimensional space starting with $\alpha_i(t=0) = 1 \forall i \in [1, \dots, n]$ at the beginning of the mission. During training, we intentionally excluded this time-dependence from the model. This was done (1) using the SDOML dataset, which has already been corrected for degradation effects, (2) without assuming any relation between t and α and avoiding to use t as an input feature, and (3) temporally shuffling the data used for training. As presented in Sect. 3.1, we degraded each set of multichannel images \mathcal{C} by a unique $\alpha = \{\alpha_i, i \in [1, \dots, n]\}$. We then devised a strategy such that from one training epoch to the next, the same

³ Salvatelli et al. (2021) and https://github.com/vale-salvatelli/sdo-autocal_pub

set of multichannel images could be dimmed by a completely independent set of α dimming factors. This data augmentation and regularization procedure allows the model to generalize and perform well in recovering dimming factors over a wide range of solar conditions.

The training set comprises multichannel images \mathbf{C} obtained during January to July from 2010 to 2013 obtained every six hours, amounting to a total of 18 970 images in 2710 time stamps. The model was trained using 64 samples per minibatch, and the training was performed for 1000 epochs. We did not use the full dataset to calculate the gradient descent and propagated back to update the network parameters or weights in the minibatch concept. Instead, we calculated the gradient descent and corrected the weights while the model was still processing the data. This procedure allowed decreasing the computation cost while still obtaining a lower variance. As a consequence of our data augmentation strategy, after 1000 epochs the model was trained with 2 710 000 unique sets of (input, output) pairs because we used a different set of α each epoch. We used the Adam optimizer (Kingma & Ba 2014) in our training with an initial learning rate of 0.001 and the mean squared error (MSE) of the predicted degradation factor (α_P), and the ground-truth value (α_{GT}) was used as the training objective (loss).

The test dataset, that is, the sample of data we used to provide an unbiased evaluation of a model fit on the training dataset, holds images obtained during August to October between 2010 and 2013, again every six hours per day, totaling 9422 images over 1346 time stamps. The split by month between the training and test data has two objectives: (1) it prevents the bias due to the variation in the solar cycle, thereby allowing the model to be deployed in future deep-space missions forecasting α for future time steps, and (2) it ensures that the same image is never present in both datasets (any two images adjacent in time will approximately be the same), leading to a more precise and comprehensive evaluation metric.

3.4. Toy model formulation to probe the multichannel relation

Using the described CNN model, we tested the hypothesis using a toy dataset, which is simpler than the SDOML dataset. We tested whether the physical relation between the morphology and brightness of solar structures (e.g., ARs, coronal holes) across multiple AIA channels would help the model prediction. For this purpose, we created artificial solar images, in which a 2D Gaussian profile was used (Eq. (1)) to mimic the Sun as an idealized bright disk with some center-to-limb variation,

$$C_i(x, y) = A_i \exp(-[x^2 + y^2]\sigma^{-2}), \quad (1)$$

where A is the amplitude centered at $(0, 0)$, the characteristic width is σ , and x and y are the coordinates at the image. σ is sampled from a uniform distribution between 0 and 1. These images are not meant to be a realistic representation of the Sun. However, as formulated in Eq. (1), they include two qualities we posit to be essential for allowing our autocalibration approach to be effective. The first is the correlation of intensities in the wavelength channels (i.e., ARs tend to be bright in multiple channels). The second is the existence of a relation between the spatial morphology of EUV structures with their brightness. This toy dataset was designed so that we were able to independently test the effect of the presence of (a) a relation between brightness A_i and size σ , and (b) a relation between A_i for various channels; and the presence of both (a) and (b) on the performance. To evaluate this test, we used the MSE loss and expect the presence of both (a) and (b) to minimize this loss.

Table 1. MSE for all combinations proposed in Sect. 3.4.

		Brightness and size correlation	
		Yes	No
Cross-channel correlation	Yes	0.017	0.023
	No	0.027	0.065

Notes. The top left cell shows the scenario of a cross-channel correlation and a relation between brightness and size of the artificial Sun. The top right cell shows the loss with a cross-channel correlation, but not the relation between brightness and size. The bottom left cell shows the loss when there is no cross-channel correlation, but it has a relation between brightness and size. The bottom right cell presents the loss when the parameters are chosen freely.

The test result of the multichannel model with artificial solar images is shown in Table 1. When $A_0 \propto \sigma$ (linear relation between size and brightness) and $A_i = A_0^i$ (i.e., dependence across channels; here the superscript i denotes A_0 to the i -th power), the CNN solution delivered minimum MSE loss (top left cell). When the interchannel relation (i.e., each A_i was randomly chosen) or the relation between brightness A_i and size σ were removed, the performance was poorer, which increased the MSE loss. Ultimately, when both A_i and σ_i were randomly sampled for all channels, the model performed equivalently to randomly guessing or regressing (bottom right cell) and having the greatest loss of all tests. These experiments confirmed our hypothesis and indicate that a multichannel input solution outperforms a single-channel input model in the presence of relations of the morphology of solar structures and their brightness across multiple AIA channels.

3.5. Reconstruction of the degradation curve using the CNN models

In order to evaluate the model on a different dataset from the dataset we used in the training process, we used both single- and multichannel CNN architectures to recover the instrumental degradation in the entire period of SDO (from 2010 to 2020). To produce the degradation curve for the two CNN models, we used a dataset that was equivalent to the SDOML dataset, but we did not correct the images for degradation⁴ (Dos Santos et al. 2021). This dataset included data from 2010 to 2020. All other preprocessing steps, including masking the solar limb, rescaling the intensity, and so on, remained unchanged. The CNN degradation estimates were then compared to the degradation estimates obtained from cross-calibration with irradiance measurements that were computed by the AIA team using the technique described in Boerner et al. (2014).

The cross-calibration degradation curve relies on the daily ratio of the AIA observed signal to the AIA signal predicted by SDO-EVE measurements until the end of EVE MEGS-A operations in May 2014. From May 2014 onward, the ratio is computed using the FISM model (Chamberlin et al. 2020) in place of the EVE spectra. FISM is tuned to SDO-EVE, so that the degradation derived from FISM agrees with the degradation derived from EVE through 2014. However, the uncertainty in the correction derived from FISM is greater than that derived from EVE observations, primarily because of the reduced spectral resolution and fidelity of FISM compared to SDO-EVE.

⁴ The SDOML dataset not corrected for degradation overtime is available at https://zenodo.org/record/4430801#.X_xuP0lKhmE

While the EVE-to-AIA cross-calibration introduced errors of only a few percent (in addition to the calibration uncertainty intrinsic to EVE itself), the FISM-to-AIA cross-calibration errors are considerably larger.

We examined V8 and V9 of the cross-calibration degradation curve. The main change from V8 calibration (released in November 2017, with linear extrapolations extending the observed trend after this date) to V9 (July 2020) is based on the analysis of the EVE calibration sounding rocket that was flown on 18 June 2018. The analysis of this rocket flight resulted in an adjustment of the trend of all channels during the interval covered by the FISM model (from May 2014 onward), as well as a 20% shift in the 171 Å channel normalization early in the mission. These changes become clearer in Fig. 6 in Sect. 5. The uncertainty of the degradation correction during the period prior to May 2014, and on the date of the most recent EVE rocket flight, is dominated by the ~10% uncertainty of the EVE measurements themselves. For periods outside of this time (particularly periods after the most recent rocket flight), the uncertainty is a combination of the rocket uncertainty and the errors in FISM in the AIA bands (approximately 28%).

Moreover, we obtained and briefly analyzed the feature maps from the second max pooling layer from the multichannel model. A feature map is the output of one mathematical filter applied to the input. The feature maps expand our understanding of the model operation. This process helps to understand the image processing and provides insight into the internal representations that combine and transform information from the seven different EUV channels into the seven dimming factors.

4. Baseline model

We compared our DNN approach to a baseline motivated by the assumption that the EUV intensity outside magnetically ARs, that is, the quiet Sun, is invariant in time (a similar approach was also considered for the in-flight calibration of some UV instruments, e.g., Schühle et al. 1998). A similar assumption in measuring the instrument sensitivity of the Solar and Heliospheric Observatory (SOHO, Domingo et al. 1995) CDS was also adopted by Del Zanna et al. (2010), who assumed that the irradiance variation in the EUV wavelengths is mainly due to the presence of ARs on the solar surface and that the mean irradiance of the quiet Sun is essentially constant throughout the solar cycle. There is evidence of small-scale variations in the intensity of the quiet Sun when observed in the transition region (Shakeri et al. 2015), but their contribution is insignificant compared to that of their AR counterparts. We used this idea for our baseline model as described in this section.

It is important to remark that we used exactly the same data preprocessing and splitting approach as we used for the neural network model described in Sect. 3.3. From the processed dataset, a set of reference images per channel, C_{ref} , were selected at time $t = t_{\text{ref}}$. Because the level of solar activity evolves continuously in time, we only selected the regions of the Sun that correspond to low activity, as discussed in the preceding paragraph. Furthermore, the activity level was decided based on coaligned (with AIA) magnetic field maps from HMI. To define these regions, we first made a square selection with a diagonal of length $2R_{\odot}$ centered at $R = 0$ of the solar images so as to avoid line-of-sight (LOS) projection effects toward the limb. We then applied an absolute global threshold value of 5 Mx cm^{-2} on the coaligned HMI LOS magnetic field maps corresponding to $t = t_{\text{ref}}$, such that only those pixels in which B_{LOS} was lower than the threshold were extracted. This resulted in a binary

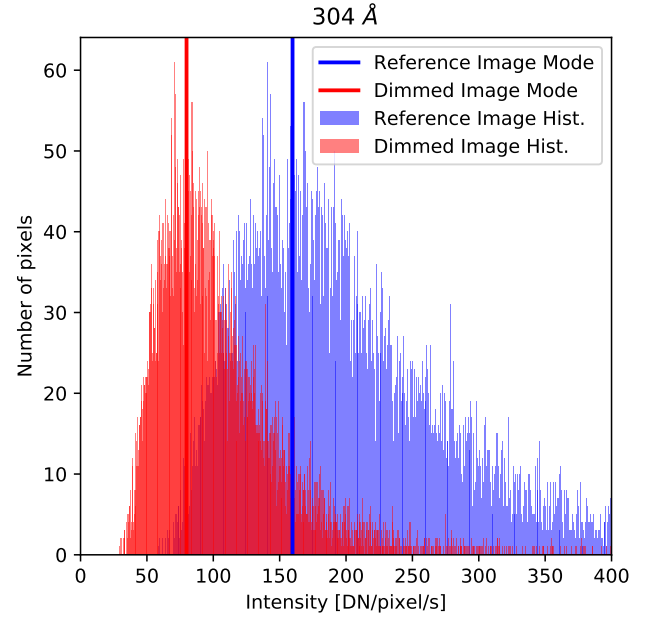


Fig. 4. Histograms of the pixel values for the 304 Å channel. In blue we show the histogram for the reference image, and in red the histogram for the dimmed image. The y-axis is the number of pixels, and the x-axis is the pixel intensity [DN/pixel/s]. The modes are marked with blue and red lines for the reference and dimmed images, respectively.

mask in which 1 corresponded to the pixels of interest, and 0 the rest. This minimum chosen value of the magnetic flux density is close to the noise level of the HMI_720s magnetograms (Liu et al. 2012; Bose & Nagaraju 2018). Finally, we used this mask to extract the cospatial quiet-Sun (less active) pixels from each AIA channel and computed the respective 1D histograms of the intensity values as shown in Fig. 4. Based on the assumption that the intensity of the quiet-Sun area does not change significantly over time (as discussed in the preceding section), we chose to artificially dim these regions by multiplying them with a constant random factor between 0 and 1. Values close to 0 will make the images progressively dimmer. The histograms for the dimmed and the original (undimmed) quiet-Sun intensities for the AIA 304 Å channel are shown in Fig. 4. The idea is to develop a non-ML approach that could be used to retrieve this dimming factor.

Based on Fig. 4, we find that both the dimmed and undimmed 1D histograms have a skewed shape, with a dominant peak at lower intensities and extended tails at higher intensities. A skewed distribution for the quiet-Sun intensities like this has been reported by various previous studies (see Shakeri et al. 2015), where they were modeled either as a sum of two Gaussians (Reeves et al. 1976) or as a single log-normal distribution (Griffiths et al. 1999; Fontenla et al. 2007). Despite an increased number of free parameters in a double-Gaussian fitting, Pauluhn et al. (2000) showed that the observed quiet-Sun intensity distribution could be fitted significantly better with a single log-normal distribution. The skewed representation, such as the one we show for the 304 Å channel, was also observed for all the other EUV channels, indicating that the criterion for masking the quiet-Sun pixels described here is justified.

We then computed the mode (most probable value) of the dimmed and undimmed log-normal distributions and indicate them by $I_{i,\text{ref}}^{\text{mp}}$ (where i implies the AIA channel under consideration, and mp stands for the modal value for the undimmed

Table 2. Results of the baseline and CNN models applied to all the EUV AIA channels.

Channel	Baseline					Single-channel model					Multichannel model				
	0.05	5%	10%	15%	20%	0.05	5%	10%	15%	20%	0.05	5%	10%	15%	20%
94 Å	32%	08%	18%	28%	40%	70%	37%	61%	78%	87%	82%	48%	73%	85%	92%
131 Å	76%	50%	73%	86%	96%	94%	72%	92%	98%	99%	99%	76%	94%	97%	99%
171 Å	58%	27%	48%	66%	85%	93%	70%	93%	97%	99%	84%	48%	72%	86%	93%
193 Å	38%	13%	27%	44%	53%	73%	41%	69%	85%	93%	90%	59%	85%	94%	98%
211 Å	31%	11%	21%	29%	39%	63%	30%	53%	71%	84%	76%	41%	68%	82%	92%
304 Å	86%	66%	89%	95%	100%	90%	65%	89%	97%	99%	94%	62%	86%	93%	96%
335 Å	38%	13%	29%	42%	51%	62%	31%	54%	69%	80%	73%	39%	65%	82%	91%
Mean	51%	27%	43%	56%	66%	78%	50%	73%	85%	92%	85%	53%	77%	89%	94%

Notes. The table is divided into three sections: baseline, single-channel, and multichannel model. From the left, we list the channel number, the success rates for the baseline, the success rates for the single-channel CNN model, and the success rates for the multichannel CNN model. Each model performance is considered at different tolerance levels. At the bottom, we list the mean of the success rate across all the channels. Green shows success rates greater than 90%, yellow a success rate of 80%–90%, and red shows a success rate lower than 80%.

images), and I_i^{mp} represents the modal intensity value for the corresponding images dimmed with a dimming factor (say α_i). These are indicated by vertical blue and red lines in Fig. 4. Subsequently, the dimming factor was obtained by computing the ratio of the two most probable intensity values according to the following equation:

$$\alpha_i := \frac{I_i^{\text{mp}}}{I_{i,\text{ref}}^{\text{mp}}} \quad (2)$$

Because both distributions are essentially similar except for the dimming factor, we suggest that this ratio is efficient enough to retrieve α_i reliably, forming a baseline against which the neural network models are compared. The efficiency of the baseline in recovering the dimming factor was then evaluated according to the success rate metric, and the results for all channels are tabulated in Table 2.

5. Results and discussions

5.1. Comparing the performances of the baseline model with different CNN architectures

The results of the learning algorithm were binarized using five different thresholds: the absolute value of 0.05, and relative values of 5%, 10%, 15%, and 20%. When the absolute difference between the predicted degradation factor (α_p) and the ground-truth degradation factor (α_{GT}) was lower than the threshold, it was considered a success α_p ; otherwise, it was not a success. We then evaluated the binarized results using the success rate, which is the ratio of success α_p and the total amount of α_p . We chose different success rate thresholds to gauge the model, all of which are lower than the uncertainty of the AIA calibration (estimated as $\sim 10\%$ earlier than May 2014 and $\sim 28\%$ later).

The baseline, single-channel, and multichannel model results are summarized in Table 2.

Table 2 reveals that for an absolute tolerance value of 0.05, the best results for the baseline are 86% (304 Å) and 76% (131 Å), and the mean success rate is $\sim 51\%$ in all channels. When the relative tolerance levels are increased, the mean success rate increases from 27% (for 5% relative tolerance) to 66% (with 20% relative tolerance) and with a 39% success rate in the worst-performing channel (211 Å).

Investigating the performance of the CNN architecture with a single input channel and an absolute tolerance level of 0.05, we find that this model performed significantly better than our baseline, with much higher values of the metric for all the channels. The most significant improvement was shown in the 94 Å channel with an increase from 32% in the baseline model to about 70% in the single input CNN model, with an absolute tolerance of 0.05. The average success rate increased from 51% in the baseline to 78% in the single-channel model. The worst metric for the single-channel CNN architecture was recorded for the 211 Å channel, with a success rate of just 63%, which is still significantly better than its baseline counterpart (31%). Furthermore, with a relative tolerance value of 15%, we find that the mean success rate is 85% for the single-channel model, which increases to more than 90% for a 20% tolerance level. This is a promising result considering that the error associated with the current best calibration techniques (sounding rockets) is $\sim 28\%$.

Finally, we report the results from the multichannel CNN architecture in the last section of Table 2. As expected, the performance in this case is the best of the models, with significant improvements for almost all the EUV channels. Clearly, the success rates belonging to the red category are far lower than the former models, implying that the mean success rate is the highest at all tolerance levels. The multichannel architecture recovers the degradation (dimming) factor for all channels with a success rate of at least 91% for a relative tolerance level of 20% and a mean success rate of $\sim 94\%$. It is also evident that this model outperforms the baseline and single-channel model for all levels of relative tolerances. For any given level of tolerance, the mean across all channels increased significantly. For example, with an absolute tolerance of 0.05, the mean increases from 78% to 85%, and even changed its color classification. In addition, the success rate is consistently lowest for the 335 Å and 211 Å channels at all tolerances, whereas the performance of the 131 Å channel is the best.

In specific channels, 304 Å performs consistently well in all the models with little variation, which was not expected. 171 Å performs well in the baseline and multichannel model, but surprisingly, the best performance is in the single-channel model at all tolerances, and a remarkable 94% success rate with a tolerance of 0.05. In contrast to 171 Å, channels 211 Å and 335 Å have a poor performance in the baseline and single-channel models, and they improve significantly in the multichannel model, as expected and hypothesized here.

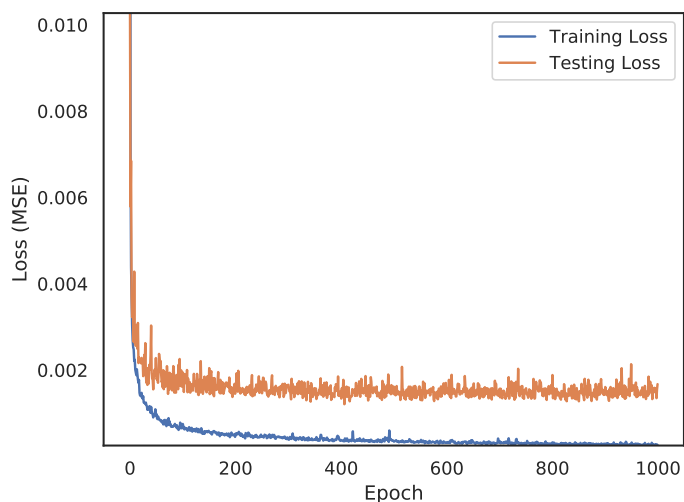


Fig. 5. Graphic of the evolution of training and testing MSE loss through the epochs.

Figure 5 shows that the training and test MSE loss curve evolves by epoch. Based on the results from Table 2 and comparing the training and test loss curves in Fig. 5, we can see that the model does not heavily overfit in the range of epochs we used, and it presents stable generalization performance on test results. We stopped the training before epoch 1000; the improvements achieved in the test set over many epochs were only marginal.

Overall, the result shows higher success rates for the CNN models, particularly for the multichannel model. This was predicted by the toy problem, and for higher tolerances.

5.2. Modeling the channel degradation over time

In this section we discuss the results obtained when the AIA degradation curves *V8* and *V9* were compared with the single- and multichannel CNN models. We used a dataset equivalent to the SDOML for this, but did not correct for degradation and data period from 2010 to 2020. This tested both models for the real degradation suffered by AIA from 2010 to 2020.

Figure 6 presents the results of our analysis for all the seven AIA EUV channels. In each panel, we show four quantities: the degradation curve *V9* (solid black line), the degradation curve *V8* (solid gray line), the predicted degradation from the single-channel model (dashed colored line), and multichannel model (solid colored line). The shaded gray band depicts the region covering the error ($\sim 10\%$ earlier than May 2014 and $\sim 28\%$ later) associated with the *V9* degradation curve, and the colored shaded areas are the standard deviation of the single- and multichannel models. The dashed vertical line coincides with the day (25 May 2014) that was the last day of EVE MEGS-A instrument data. It is important to note that MEGS-A was used earlier for sounding-rocket calibration purposes, the loss of which caused the *V8* and *V9* degradation curves to become noisier. Szenicer et al. (2019) used deep learning to facilitate a virtual replacement for MEGS-A.

The different panels of Fig. 6 show that even though we trained the single and multichannel models with the SDOML dataset, which was produced and corrected using the *V8* degradation curve, the two CNN models predict the degradation curves for each channel quite accurately over time, except for the 94 Å and 211 Å channel. However, the deviations of the predicted values for these two channels fall well within the shaded

area associated with the *V9* calibration curve. The CNN predictions agree even better with *V9* than the *V8* calibration for most of the channels. This indicates that the CNN detects some actual information that is perhaps even more responsive to degradation than FISM. The latest degradation curve (*V9*) has been updated in July 2020, and the change from *V8* to *V9* might have easily caused an impact while training the models. Moreover, the more significant deviation of the 94 Å channel in the early stages of the mission arises because we limited our degradation factor to be lower than one.

The predicted calibration curves computed from the single- and multichannel models overlap significantly throughout the entire period of observation. The single-channel model predictions, however, vary more significantly for channels 211 Å, 193 Å, and 171 Å. For a systematic evaluation and a comparison of the results of the two models in the channels, we calculated some goodness-of-fit metrics. The results are shown in Table 3.

Table 3 contains two different metrics for evaluating the goodness of fit of each CNN model with the *V9* degradation curve. The first is the two-sample Kolmogorov-Smirnov test (KS), which determines whether two samples come from the same distribution (Massey 1951), and the null hypothesis assumes that the two distributions are identical. The KS test has the advantage that the statistics distribution does not depend on the cumulative distribution function that is tested. The second metric is the fast dynamic time warping (DTW, Salvador & Chan 2007), which measures the similarity between two temporal sequences that may not be of the same length. This last is important because statistical methods can be too sensitive when both time series are compared. The DTW gives the distance between the series as an output, and as a reference, the DTW for the different EUV channels between the *V8* and *V9* degradation curves are 94 Å–72.17, 131 Å–13.03, 171 Å–9.82, 193 Å–30.05, 211 Å–16.86, 304 Å–7.02, and 335 Å–5.69.

Similar to Fig. 6, Table 3 shows that the predictions from the single- and multichannel models overlap significantly in terms of the metric and time evolution. Except for the 94 Å channel, all others have very similar metric values, well within a given level of tolerance. A low value of the KS test metric suggests that the predictions have a similar distribution as the observed *V9* calibration curve, which also indicates the robustness of our CNN architecture. The KS test agrees well with the DTW, where the obtained values are lower than the reference values (as indicated earlier) of the *V8* and the *V9* calibration curves. Overall, the metric analysis for the goodness of fit between the predictions and the actual calibration curve (*V9*) shows that the CNN models perform remarkably well in predicting the degradation curves, although they were only trained on the first three years of observations.

5.3. Feature maps

As mentioned in Sect. 3.5, the feature maps are the result of applying the filters to an input image. That is, at each layer, the feature map is the output of that layer. In Fig. 7 we present such maps obtained from the output of the last convolutional layer of our CNN. The top row shows the reference input image observed at 193 Å we used in this analysis, with its intensity scaled between 0 – 1 pixel units, and the bottom row shows 4 representative feature maps (out of a total of 128) with their corresponding weights. These maps were obtained after the final convolutional layer of the multichannel model, and they represent the result of combining all seven EUV channels as input.

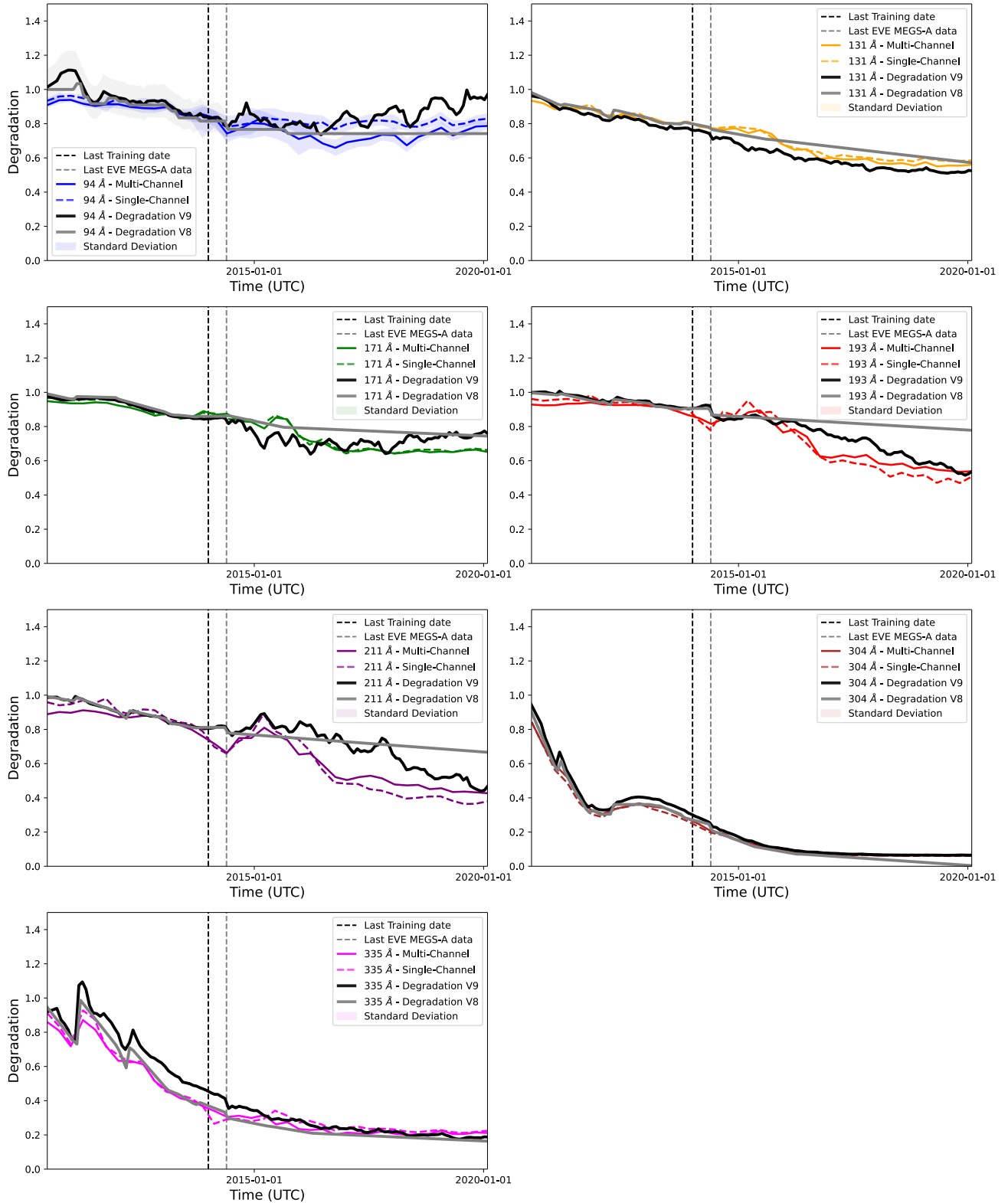


Fig. 6. Channel degradation over time. *From top to bottom:* Channel 94 Å (blue), 131 Å (yellow), 171 Å (green), 193 Å (red), 211 Å (purple), 304 Å (brown), and 335 Å (magenta). The solid black (gray) curve is the degradation profile of AIA calibration release V9 (V8). The gray shaded area correspond to the error (10% earlier than May 2014 and 28% error later) of the degradation curve V9. The colored shaded areas are the standard deviation of the CNN models. The vertical dashed black line is the last available observation from EVE MEGS-A data, and the vertical dashed gray line is the last training date.

The predicted α dimming factors from the model are given by the sigmoid activation function applied to a linear combination of these features. This mapping allows us to show that the net-

work learned to identify the different features of full-disk solar images such as the limb, the quiet-Sun features, and the ARs. The reason for visualizing a feature map for specific AIA images

Table 3. Goodness-of-fit metrics for single- and multichannel models with reference to the V9 degradation curve.

Channel	Single channel		Multichannel	
	KS	DTW	KS	DTW
94 Å	0.485	7.120	0.568	9.624
131 Å	0.346	2.711	0.275	1.624
171 Å	0.298	3.074	0.329	3.549
193 Å	0.211	1.829	0.244	2.080
211 Å	0.305	2.850	0.242	2.807
304 Å	0.282	1.412	0.100	1.311
335 Å	0.212	2.539	0.141	2.839

Notes. The first metric is the KS, and the second metric is the DTW.

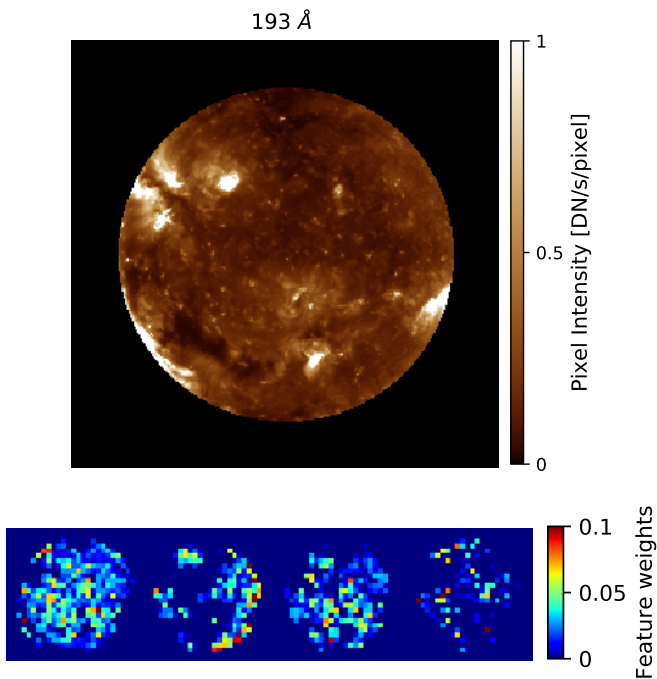


Fig. 7. Feature maps obtained from the last layer of the CNN of our model. *Top row:* a sample input in the AIA 193 Å channel, and *bottom row:* 4 representative feature maps out of 128 different feature maps from the final convolutional layer of the multichannel CNN model.

is to gain an understanding of what features that a model detects are ultimately useful in recovering the degradation or the dimming factors.

6. Concluding remarks

This paper reports a novel ML-based approach to autocalibration and advances our understanding of the cross-channel relation of different EUV channels by introducing a robust novel method to correct for the EUV instrument time degradation. We began with formulating the problem and setting up a toy model to test our hypothesis. We then established two CNN architectures that consider multiple wavelengths as input to autocorrect for on-orbit degradation of the AIA instrument on board SDO. We trained the models using the SDOML dataset and further augmented the training set by randomly degrading images at each epoch. This approach ensured that the CNN model suitably generalizes to

data that are not seen during the training, and we also developed a non-ML baseline to test and compare its performance with the CNN models. With the best-trained CNN models, we reconstructed the AIA multichannel degradation curves of 2010–2020 and compared them with the sounding-rocket-based degradation curves V8 and V9.

Our results indicate that the CNN models significantly outperform the non-ML baseline model (85% versus 51% in terms of the success rate metric) for a tolerance level of 0.05. In addition, the multichannel CNN also outperforms the single-channel CNN with a 78% success rate with an absolute 0.05 threshold. This result is consistent with the expectation that correlations between structures in different channels, size (morphology) of structures, and brightness can be used to compensate for the degradation. To further understand the correlation between different channels, we used the concept of feature maps to shed light on this aspect and determine how the filters of the CNNs were being activated. We showed that the CNNs learned representations that make use of the different features within solar images, but further work needs to be done on this aspect to establish a more detailed interpretation.

We also found that the CNN models reproduce the most recent sounding-rocket-based degradation curves (V8 and V9) very closely and within their uncertainty levels. This is particularly promising, given that no time information has been used in training the models. For some specific channels, such as 335 Å, the model reproduced the V8 curve instead of V9 because the SDOML corrected using the former. The single-channel model could perform as well as the multichannel model even though the multichannel model presented a more robust performance when evaluated on the basis of their success rates.

This paper finally presents a unique possibility of auto-calibrating deep-space instruments such as those onboard the STEREO spacecraft and the recently launched remote-sensing instrument called Extreme Ultraviolet Imager (Rochus et al. 2020) on board the Solar Orbiter satellite (Müller et al. 2020), which are too far away from Earth to be calibrated using a traditional method such as sounding-rockets. The autocalibration model could be trained using the first months of data from the mission, assuming the instrument is calibrated at the beginning of the mission. The data volume might be a problem, and different types of data augmentation could be used to overcome this problem, such as synthetic degradation and image rotation. We further envision that the technique presented here may also be adapted to imaging instruments or spectrographs operating at other wavelengths (e.g., hyperspectral Earth-oriented imagers) observed from different space-based instruments such as IRIS (De Pontieu et al. 2014).

Acknowledgements. This project was partially conducted during the 2019 Frontier Development Lab (FDL) program, a co-operative agreement between NASA and the SETI Institute. We wish to thank IBM for providing computing power through access to the Accelerated Computing Cloud, as well as NASA, Google Cloud and Lockheed Martin for supporting this project. L.F.G.S was supported by the National Science Foundation under Grant No. AGS-1433086. M.C.M.C. and M.J. acknowledge support from NASA’s SDO/AIA (NNG04EA00C) contract to the LMSAL. S.B. acknowledges the support from the Research Council of Norway, project number 250810, and through its Centers of Excellence scheme, project number 262622. This project was also partially performed with funding from Google Cloud Platform research credits program. We thank the NASA’s Living With a Star Program, which SDO is part of, with AIA, and HMI instruments on-board. CHIANTI is a collaborative project involving George Mason University, the University of Michigan (USA), University of Cambridge (UK) and NASA Goddard Space Flight Center (USA). A.G.B. is supported by EPSRC/MURI grant EP/N019474/1 and by Lawrence Berkeley National Lab. The authors thank the anonymous referee for the comments.

Software: We acknowledge for CUDA processing cuDNN (Chetlur et al. 2014), for data analysis and processing we used Sunpy (Mumford et al. 2020), Numpy (van der Walt et al. 2011), Pandas (McKinney et al. 2010), SciPy (Virtanen et al. 2020), scikit-image (van der Walt et al. 2014) and scikit-learn (Pedregosa et al. 2011). Finally all plots were done using Matplotlib (Hunter 2007) and Astropy (Astropy Collaboration 2018).

References

- Amodei, D., Ananthanarayanan, S., Anubhai, R., et al. 2016, in *International Conference on Machine Learning*, 173
- Astropy Collaboration (Price-Whelan, A. M., et al.) 2018, *AJ*, **156**, 123
- BenMoussa, A., Gissot, S., Schühle, U., et al. 2013, *Sol. Phys.*, **288**, 389
- Bobra, M. G., & Couvidat, S. 2015, *ApJ*, **798**, 135
- Boerner, P. F., Testa, P., Warren, H., Weber, M. A., & Schrijver, C. J. 2014, *Sol. Phys.*, **289**, 2377
- Bose, S., & Nagaraju, K. 2018, *ApJ*, **862**, 35
- Chamberlin, R. V., Mujica, V., Izvekov, S., & Larentzos, J. P. 2020, *Phys. A Stat. Mech. App.*, 540
- Chetlur, S., Woolley, C., Vandermersch, P., et al. 2014, ArXiv e-prints [arXiv:1410.0759]
- Cheung, C. M. M., Jin, M., Dos Santos, L. F. G., et al. 2019, *AGU Fall Meeting Abstracts*, 2019, NG31A-0836
- Clarke, S. 2016, *EGU General Assembly Conference Abstracts*, 18, EPSC2016-18529
- Del Zanna, G., Andretta, V., Chamberlin, P. C., Woods, T. N., & Thompson, W. T. 2010, *A&A*, **518**, A49
- De Pontieu, B., Title, A. M., Lemen, J. R., et al. 2014, *Sol. Phys.*, **289**, 2733
- Domingo, V., Fleck, B., & Poland, A. I. 1995, *Sol. Phys.*, **162**, 1
- Dos Santos, L. F. G., Bose, S., Salvatelli, V., et al. 2021, *SDOML Dataset Not Corrected for Degradation Over Time*
- Fontenla, J. M., Curdt, W., Avrett, E. H., & Harder, J. 2007, *A&A*, **468**, 695
- Galvez, R., Fouhey, D. F., Jin, M., et al. 2019, *ApJS*, **242**, 7
- Goodfellow, I., Bengio, Y., & Courville, A. 2006, *Deep learning* (MIT press)
- Griffiths, N. W., Fisher, G. H., Woods, D. T., & Siegmund, O. H. W. 1999, *ApJ*, **512**, 992
- He, K., Zhang, X., Ren, S., & Sun, J. 2016, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770
- Hunter, J. D. 2007, *Comput. Sci. Eng.*, **9**, 90
- Iqbal, H. 2018, <https://doi.org/10.5281/zenodo.2526396>
- Jiao, Z., Jiang, L., Sun, J., Huang, J., & Zhu, Y. 2019, *IOP Conf. Ser. Mater. Sci. Eng.*, **611**
- Jungbluth, A., Gitiaux, X., Maloney, S., et al. 2019, *Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019)*, Vancouver, Canada
- Kaiser, M. L., Kucera, T. A., Davila, J. M., et al. 2008, *Space Sci. Rev.*, **136**, 5
- Kim, T., Park, E., Lee, H., et al. 2019, *Nat. Astron.*, **3**, 397
- Kingma, D. P., & Ba, J. 2014, ArXiv e-prints [arXiv:1412.6980]
- LeCun, Y., & Bengio, Y. 1995, *The Handbook of Brain Theory and Neural Networks*, 3361, 1995
- Lemen, J. R., Title, A. M., Akin, D. J., et al. 2012, *Sol. Phys.*, **275**, 17
- Liu, Y., Hoeksema, J. T., Scherrer, P. H., et al. 2012, *Sol. Phys.*, **279**, 295
- Massey, F. J., Jr 1951, *J. Am. Stat. Assoc.*, **46**, 68
- McKinney, W. 2010, in *Proceedings of the 9th Python in Science Conference*, eds. S. van der Walt, & J. Millman, 56
- Müller, D., St. Cyr, O. C., Zouganelis, I., et al. 2020, *A&A*, **642**, A1
- Mumford, S. J., Freij, N., Christe, S., et al. 2020, <https://doi.org/10.5281/zenodo.591887>
- Nair, V., & Hinton, G. E. 2010, *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10 (USA: Omnipress)*, 807
- Neuberg, B., Bose, S., Salvatelli, V., et al. 2019, ArXiv e-prints [arXiv:1911.04008]
- Oord, A. V. d., Dieleman, S., Zen, H., et al. 2016, ArXiv e-prints [arXiv:1609.03499]
- Paszke, A., Gross, S., Chintala, S., et al. 2017, *NeurIPS Autodiff Workshop*
- Pauluhn, A., Solanki, S. K., Rüedi, I., Landi, E., & Schühle, U. 2000, *A&A*, **362**, 737
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, *J. Mach. Learn. Res.*, **12**, 2825
- Pesnell, W., Thompson, B., & Chamberlin, P. 2012, *Sol. Phys.*, **275**, 3
- Reeves, E. M., Vernazza, J. E., & Withbroe, G. L. 1976, *Philos. Trans. R. Soc. London Ser. A*, **281**, 319
- Rochus, P., Auchère, F., Berghmans, D., et al. 2020, *A&A*, **642**, A8
- Salvador, S., & Chan, P. 2007, *Intell. Data Anal.*, **11**, 561
- Salvatelli, V., Bose, S., Neuberg, B., et al. 2019, ArXiv e-prints [arXiv:1911.04006]
- Salvatelli, V., Neuberg, B., Dos Santos, L. F. G., et al. 2021, *ML pipeline for Solar Dynamics Observatory (SDO) data*
- Schou, J., Scherrer, P. H., Bush, R. I., et al. 2012, *Sol. Phys.*, **275**, 229
- Schühle, U., Brekke, P., Curdt, W., et al. 1998, *Appl. Opt.*, **37**, 2646
- Schwenn, R. 2006, *Liv. Rev. Sol. Phys.*, **3**, 2
- Shakeri, F., Teriaca, L., & Solanki, S. K. 2015, *A&A*, **581**, A51
- Szenicer, A., Fouhey, D. F., Munoz-Jaramillo, A., et al. 2019, *Sci. Adv.*, **5**, <https://advances.sciencemag.org/content/5/10/eaaw6548.full.pdf>
- van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, *Comput. in Sci. Eng.*, **13**, 22
- van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., et al. 2014, *PeerJ*, **2**, e336v2
- van Driel-Gesztelyi, L., & Green, L. M. 2015, *Liv. Rev. Sol. Phys.*, **12**, 1
- Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, *Nat. Meth.*, **17**, 261
- Wieman, S., Didkovsky, L., Woods, T., Jones, A., & Moore, C. 2016, *Sol. Phys.*, **291**, 3567
- Woods, T. N., Eparvier, F. G., Hock, R., et al. 2012, *Sol. Phys.*, **275**, 115
- Wu, Y., Schuster, M., Chen, Z., et al. 2016, ArXiv e-prints [arXiv:1609.08144]

Appendix A: Scaling units for each AIA channel

Table A.1. AIA channel scaling units.

AIA channel (Å)	Scaling unit [DN/s/pixel]
94	10
131	80
171	2000
193	3000
211	1000
304	500
335	80

Appendix B: Detailed model architectures

Table B.1. Detailed single-channel architecture.

#	Layer	Output Shape	Learnable Parameters
1	Input	[1, 256, 256]	0
2	Convolution	[64, 254, 254]	640
3	ReLU	[64, 254, 254]	0
4	MaxPooling	[64, 84, 84]	0
5	Convolution	[128, 82, 82]	73 856
6	ReLU	[128, 82, 82]	0
7	MaxPooling	[128, 27, 27]	0
8	Fully Connected	[1]	93 313
9	Sigmoid	[1]	2
Total Learnable Parameters			167 809

Notes. From left to right columns, we list the layer number, layer type, output shape, and the total number of learnable parameters of the layer. The architecture is composed of two blocks of a convolutional layer, ReLU activation function, and max pooling layer, followed by an FC layer and a final sigmoid activation function.

Tables B.1 and B.2 present more detailed information of the CNN architecture. From left to right, they show the layer number, the type of layer, the output shape of each layer, and the learnable parameters each layer has. During the training pro-

Table B.2. Detailed multichannel architecture.

#	Layer	Output Shape	Learnable Parameters
1	Input	[7, 256, 256]	0
2	Convolution	[64, 254, 254]	4096
3	ReLU	[64, 254, 254]	0
4	MaxPooling	[64, 84, 84]	0
5	Convolution	[128, 82, 82]	73 856
6	ReLU	[128, 82, 82]	0
7	MaxPooling	[128, 27, 27]	0
8	Fully Connected	[7]	653 191
9	Sigmoid	[7]	56
Total Learnable Parameters			731 143

Notes. From left to right columns, we list the layer number, layer type, output shape, and the total number of learnable parameters of the layer. The architecture is composed of two blocks of a convolutional layer, ReLU activation function, and max pooling layer, followed by an FC layer and a final sigmoid activation function.

cess, the model works to learn and optimize the weights and biases in a neural network. These weights and biases are indeed the learnable parameters. Any parameter within our model that is learned or updated during the training process is a learnable parameter.

Learnable parameters are calculated differently for each layer. For convolutional layers we use Eq. (B.1),

$$LP_{\text{Convolution}} = ((H \times W \times f) + 1) \times k, \quad (\text{B.1})$$

where $LP_{\text{Convolution}}$ are the learnable parameters of the convolution layer, H is the height of the input, W is the width, f is the number of filters from the previous layer, 1 is the bias, and k is the number of filters in the convolution. For the fully connected layers and sigmoids, we use Eq. (B.2),

$$LP_{\text{Connected}} = ((C \times P) + 1 \times c), \quad (\text{B.2})$$

where $LP_{\text{Connected}}$ are the learnable parameters of the fully connected layer C is the number of current layer neurons, P is the number of previous layers neurons, and 1 is the bias. ReLU and max pooling layers have zero learnable parameters because they do not have weights to be updated as the neural network is trained.