Exploring the Limits of Synthetic Creation of Solar EUV Images via Image-to-Image Translation

Valentina Salvatelli [ID],[1,2,3] Luiz F. G. dos Santos [ID],[4,5] Souvik Bose [ID],[6,7,8,9] Brad Neuberg,[2,3,10] Mark C. M. Cheung [ID],[8] Miho Janvier [ID],[11] Meng Jin [ID],[3,8] Yarin Gal [ID],[12] and Atilim Güneş Baydin [ID][13]

[1]*Microsoft Research, Cambridge CB12FB, UK*
[2]*Frontier Development Lab, Mountain View, CA 94043, USA*
[3]*SETI Institute, Mountain View, CA 94043, USA*
[4]*Shell Global Solutions International B.V., Grasweg 31, 1031 HW Amsterdam, The Netherlands*
[5]*nextSource Inc, New York, NY 10018, USA*
[6]*Rosseland Center for Solar Physics, University of Oslo,P.O. Box 1029 Blindern, NO-0315 Oslo, Norway*
[7]*Institute of Theoretical Astrophysics, University of Oslo,P.O. Box 1029 Blindern, NO-0315 Oslo, Norway*
[8]*Lockheed Martin Solar & Astrophysics Laboratory, Palo Alto, CA 94304, USA*
[9]*Bay Area Environmental Research Institute, NASA Research Park, Moffett Field, CA 94035, USA*
[10]*Planet, San Francisco, CA 94107, USA*
[11]*Université Paris-Saclay, CNRS, Institut d'astrophysique spatiale, Orsay, France*
[12]*OATML Group, Department of Computer Science, University of Oxford, UK*
[13]*Department of Computer Science, University of Oxford, Oxford OX1 3QD, UK*

ABSTRACT

The Solar Dynamics Observatory (SDO), a NASA multi-spectral decade-long mission that has been daily producing terabytes of observational data from the Sun, has been recently used as a use-case to demonstrate the potential of machine learning methodologies and to pave the way for future deep-space mission planning. In particular, the idea of using image-to-image translation to virtually produce extreme ultra-violet channels has been proposed in several recent studies, as a way to both enhance missions with less available channels and to alleviate the challenges due to the low downlink rate in deep space. This paper investigates the potential and the limitations of such a deep learning approach by focusing on the permutation of four channels and an encoder–decoder based architecture, with particular attention to how morphological traits and brightness of the solar surface affect the neural network predictions. In this work we want to answer the question: can synthetic images of the solar corona produced via image-to-image translation be used for scientific studies of the Sun? The analysis highlights that the neural network produces high-quality images over three orders of magnitude in count rate (pixel intensity) and can generally reproduce the covariance across channels within a 1% error. However the model performance drastically diminishes in correspondence of extremely high energetic events like flares, and we argue that the reason is related to the rareness of such events posing a challenge to model training.

Keywords: Sun: activity, UV radiation, and general - Techniques: image processing, GPU computing - Methods: data analysis, telescopes - Open-source software

1. INTRODUCTION

Since its launch in 2010, NASA's Solar Dynamics Observatory (SDO; Pesnell et al. 2012) has monitored the evolution of the Sun. SDO data has enabled researchers to track the evolution of the Sun's interior plasma flows over solar cycle 24 and beyond. It has also continuously monitored the evolution of the solar corona, capturing dynamical evolution at time-scales of seconds and minutes. This capability is due to the suite of four telescopes on the Atmospheric Imaging Assembly (AIA; Lemen et al. 2012) instrument, which captures full-Sun images at two ultraviolet (UV) bands, seven extreme UV (EUV) bands, and one visible band.

Corresponding author: Valentina Salvatelli
vsalvatelli@microsoft.com

The seven EUV channels are designed to capture photons from emission lines in highly ionized metals in plasmas at transition region (TR; 10^5 K $\lesssim T \lesssim 10^6$ K) and coronal temperatures ($10 \gtrsim 10^6$ K). This combination of channels with sensitivity to different temperatures allows researchers to track how transition regions and coronal plasmas heat and cool (e.g., Cheung et al. 2015), and to use these thermal histories to test theories of coronal heating and of flares.

The high spatial resolution ($\sim$ **1.5″**, 4096×4096 pixels), high cadence (12 s for EUV channels) full-disk observing capability is possible because of SDO's ground system providing a sustained downlink rate of $\sim$ 67 Mbps. The collection of continuous data, over more than one solar cycle, provides not only numerous opportunities to perform data-driven scientific studies but also research with the potential to help optimize future solar physics missions.

For instance, the idea of using SDO images for image-to-image translation has been explored in several papers, most notably by Díaz Baso & Asensio Ramos (2018); Galvez et al. (2019); Szenicer et al. (2019); Park et al. (2019); Salvatelli et al. (2019). Image-to-image translation can potentially provide a way to enhance the capabilities of solar telescopes with fewer channels or less telemetry than is available to SDO. The *SDO image translation problem* can be defined as follows: given a set of N (nearly) contemporaneous images taken in different EUV channels, can a model be developed which maps the N input images to the image of a missing (not in input) EUV channel?

Notably, Lim et al. (2021) adopted a widely used image translation method (Pix2Pix, Isola et al. 2017) to tackle the SDO image-translation problem and to understand which subset of channels can better translate other channels. They trained and evaluated models for all combinations of input channels for both $N = 2$ and $N = 3$ variants of the problem, and compared global image quality metrics to pick out the channel combinations that perform the best. For some channel combinations, the reported pixel-to-pixel correlation coefficient approaches unity.

In this paper, we build on the method presented in Salvatelli et al. (2019) for one single channel and we delve deeper into the opportunities and the limitations of applicability of such "virtual telescopes". We focus on a permutation of a subset of channels (4 out of 10) and we explore in greater detail what is the quality of this synthetic generation on a number of scientifically-motivated metrics (figures of merit) and in relation to periods and regions of different level of activity of the Sun.

Together with this paper we also open source the code we used for the analysis Salvatelli et al. (2022)[1] and that can be used by the community to train and evaluate similar models on the publicly available SDO dataset released by Galvez et al. (2019) .

2. DATA

The work presented in this project is based on data from SDO's AIA. The AIA instrument takes full-disk, 4096×4096 pixel, imaging observations of the solar photosphere, chromosphere and corona in two UV channels and in seven extreme UV (EUV) channels. The original SDO dataset was processed in Galvez et al. (2019) into a machine-learning ready dataset of $\sim$ 6.6 TB (hereafter SDOML) that we leveraged for the current work.

The SDOML dataset is a subset of the original SDO data ranging from 2010 to 2018. Images are spatially co-registered, have identical angular resolutions, are corrected for the instrumental degradation over time and have exposure corrections applied. All the instruments are temporally aligned. AIA images in the SDOML dataset are available at a sampling rate of 6 min. The 512×512 pixel full-disk images have a pixel size of $\sim$ 4″8.

The images are saved in single-precision floating point to preserve the high dynamic range ($\gtrsim$ 14 bits per channel per pixel). For numerical performance purposes, the images of each channel are re-scaled by a per-channel constant factor which is approximately the average count rate for that channel. The per-channel constant factors can be found at Tab.6.

3. METHODOLOGY

Our approach of synthesizing solar EUV images is to perform image translation from multiple input channels to one single output channel. For the development of this work we focused on the permutations of four channels (94, 171, 193, 211 Å). These channels are sensitive to coronal plasmas at different temperatures (Cheung et al. 2015).

To perform the image translation we used a deep neural network (DNN, Goodfellow et al. 2016), more specifically we adopted a U-Net architecture (Ronneberger et al. 2015), an encoder–decoder with skip connections that was first designed for image segmentation on medical images. We used Adam optimizer (Kingma & Ba 2014) and Leaky ReLU (Maas et al. 2013) activations, and implement the code using the open source library PyTorch (Paszke et al. 2017). The full details of the adopted architecture is given in Fig. 1. We limit the number of channels to four for computational resources constraints. For the training and inference of the architecture presented above we used 4× NVIDIA Tesla T4s. We trained each model for 600 epochs.

---

[1] Zenodo: ML pipeline for Solar Dynamics Observatory (SDO) data
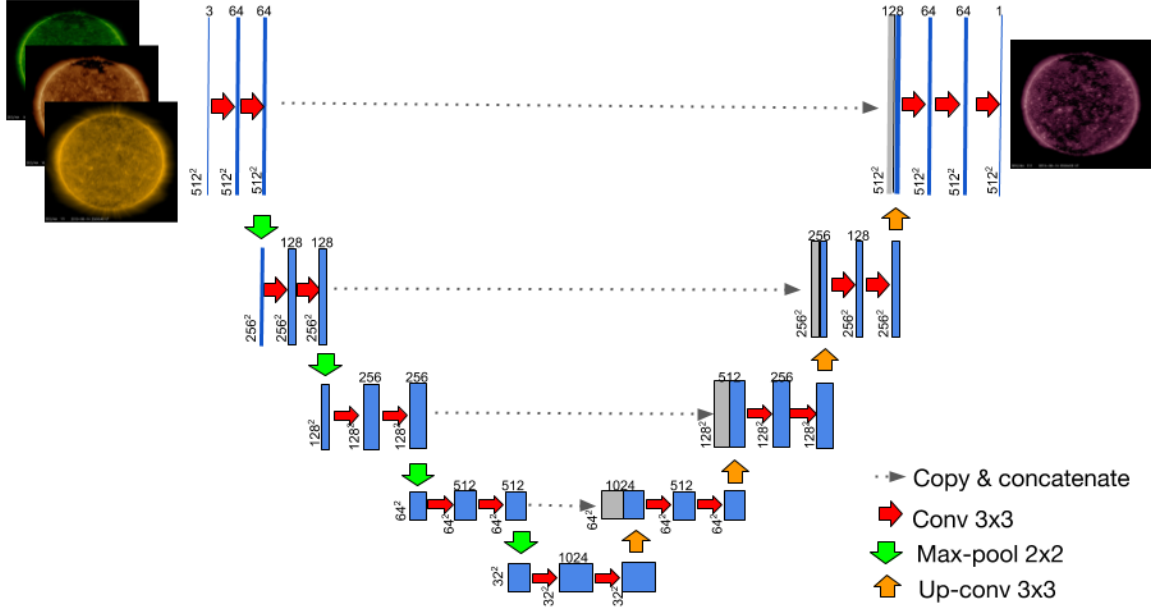
Figure 1. U-Net based architecture used to synthesize solar EUV images. Each box corresponds to a multi-channel feature map. Grey boxes are copied maps. The number of channels is shown on top of the box. Resolution in pixels is indicated on the left of the box. Arrows represent operations. For images of size 512×512, the trainable parameters are $34, 513, 857$. Figure taken from (Salvatelli et al. 2019).

For comparison we experimented also with a simpler baseline model, described by the following equation:

$$Y_{\text{pred}} = \alpha X_1 + \beta X_2 + \gamma X_3 + \delta \tag{1}$$

where Y_{pred} is the reconstructed pixel of the output channel, X_i are the pixel values of the input channels; α, β, γ are the weights and δ the bias of the linear combination of the channels. $\alpha, \beta, \gamma, \delta$ are trainable parameters of the model.

The metrics we use to evaluate the accuracy of our results for each permutation are:

- The difference between predicted and ground truth images in the form of normalized mean squared error (NMSE; Eq. 2) and normalized root mean squared error (NRMSE; Eq. 3).

$$\text{NMSE}(y, \hat{y}) = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} y_i^2} \tag{2}$$

$$\text{RNMSE}(y, \hat{y}) = \frac{\sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}}{\overline{y}} \tag{3}$$

- The structural similarity index (SSIM; Wang et al. 2004), a metric commonly used in computer vision to compute similarity between images, measuring the difference in terms of visually perceived texture and morphology. Identical images have SSIM equal to 1.

- The average of NRMSE and SSIM, as described in Eq. 4. Lower values mean better performance in this metric.

$$\text{Err}(y, \hat{y}) = \frac{\text{NRMSE}(y, \hat{y}) + [1 - |\text{SSIM}(y, \hat{y})|]}{2} \tag{4}$$

- The average pixel-to-pixel Pearson correlation coefficient.

In order to assess how much the DNN is able to learn the physical correlations between channels and to correctly reproduce them in the synthetic images, we also evaluate the difference between the real and the synthetic covariance of the channels. With

Deep Neural Network	211_sqr	211	193_sqr	193	171_sqr	171	94_sqr	94
NMSE	0.010024	0.008748	0.013414	0.013015	0.015270	0.010151	0.009482	0.013643
NRMSE	0.195127	0.182286	0.225717	0.222332	0.240829	0.196360	0.189773	0.227641
|1 - SSIM |	0.040844	0.046189	0.022866	0.024522	0.030636	0.034892	0.114447	0.138455
(NRMSE + |1 - SSIM|)/2	0.117985	0.114237	0.124292	0.123427	0.135732	0.115626	0.152110	0.183048

Table 1. Performance of the DNN on different permutations of input/output channels in the set (94, 171, 193, 211 Å) and for different scaling of the input data. In every column the input channels are all but the one indicated in the column name that corresponds to the output channel. Each value is the mean over the whole test dataset. For each metric in this table lower is better. For 94 Å the similarity index is higher than for the others channels, this can be explained by the fact the average value in this channel is higher and the metric is affected by the absolute values. See Sec. 3 for explanation of the metrics.

Deep Neural Network	Model output			
Scaling	211 Å	193 Å	171 Å	94 Å
Non Root	0.994 ± 0.004	0.991 ± 0.006	0.993 ± 0.003	0.991 ± 0.003
Root	0.993 ± 0.004	0.996 ± 0.004	0.990 ± 0.005	0.994 ± 0.004

Table 2. DNN model. Average Pearson correlation coefficient pixel-to-pixel, mean and standard deviation over the full test dataset for permutations of input/output channels in the set (94, 171, 193, 211 Å). For each channel combination the average Pearson correlation coefficient pixel-to-pixel was calculated for both trained models, with and without root scaling. The results observed are impressive and in all cases the performance is superior to 0.99.

the aim of better understanding the error, in addition to the standard covariance we compute the neighborhood covariance. In this case the output is a map of the same size of the input images where each value in the map corresponds to the covariance on a squared patch centered in the pixel and of size 20×20 pixels as described in Eq. 5.

$$cov_{patch} = \frac{\sum_i^N [(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})]}{N - 1} \tag{5}$$

where N is the total number of pixels in the patch.

Each model has been trained on $6,444$ images ($1,611$ timestamps, one image per channel for each timestamp) in the intervals January 1st 2011 to July 31st 2011 and January 1st 2012 to July 31st 2012. For testing $2,668$ images (667 timestamps) have been used, taken in the intervals August 1^{st}2011 to October 31^{st}2011 and August 1^{st}2012 to October 31st 2012. Each timestamp is at least 61 hours apart from the closest ones. These time ranges have been selected to ensure we were testing on images significantly different from the training ones. Only timestamps for which all the channels of interest were available have been included in the above datasets.

4. EXPERIMENTS

For this analysis we trained eight models using the data and architecture described in Sec. 2 and Sec. 3, two models for each of the four channels permutation. For each channel permutation we trained (1) a model where the input data was scaled by a constant factor (cf. Tab. 6) and (2) a model where the square root of the input data was taken, in addition to the constant scaling. The second scaling technique is to explore the impact of pixels with extreme ranges on the training. Each model has been evaluated by studying both the aggregated performance on the full test data and the performance on specific timestamps. Namely timestamps in the neighborhood of Valentine's Day flare (2011-2-15:1:50:00 UT) and in a quiet day of the same month (2011-02-10 00:00:00). The focus of these experiments is to evaluate the robustness of the image-to-image translation approaches in normal and extreme conditions of the Sun's activity. For comparison, we trained also four linear models, one model for each of the four channels permutation, using Eq. 1and input scaled by a constant factor.

5. RESULTS

In Tab. 1 we explore the permutations of three input channels and one output channel and the effect of applying a root scaling transformation to the input images. In addition in Tab. 2 we show the correlation pixel by pixel for each of the permutations. We found that the same architecture produces similar reconstruction errors and correlation values over all the channels with a NMSE

Linear Model	211	193	171	094
NMSE	0.749594	0.742833	0.741476	0.875264
NRMSE	1.687336	1.679708	1.678174	1.823300
1 - SSIM	0.588910	0.441623	0.490644	0.976495
(NRMSE + |1 - SSIM|)/2	1.138123	1.060665	1.084409	1.399897

Table 3. For comparison with Tab. 1, performance of the linear model on different permutations of input/output channels in the set (94, 171, 193, 211 Å) for standard (no square root) scaling. The DNN consistently improves results of one order of magnitude in each of these metrics. The comparison demonstrates non-linear patterns between channels are important for a correct reconstruction of the images.

of about 0.01. We observe the similarity index of 94 Å is worse of an order of magnitude with respect to the other channels, this can be explained by the fact SSIM is a not normalized metric and the average test value for this channel is higher than for the others (see Appendix, Tab. 7). The results are remarkable, for example for 94 Å the peak emission lies at a considerably higher temperature than the input channels (see Fig. 1 of Cheung et al. 2015) that makes the reconstruction task a particularly challenging one. These results are in agreement with the results in Salvatelli et al. (2019) and Lim et al. (2021). Please note that the values reported in Tab.1 of Salvatelli et al. (2019) are not normalized. The squared-root scaling model shows roughly equivalent performance with the model with no squared-root applied to input data except for the channel 94 Å.

It is interesting to compare the results in Tab. 1 with those in Tab. 3 where the same set of metrics are computed for the linear model. The DNN consistently improves by one order of magnitude over the linear model performance. This result clearly displays the value of using a DNN over a simpler model for the synthesis of the image. The comparison also demonstrates the strength of non-linearity between EUV channels and the fact it cannot be neglected for a meaningful reconstruction.

In order to further evaluate the performance of both models, we calculate in Tab. 2 the average pixel-to-pixel Pearson correlation for pixels inside the solar disk for each channel combination. Agreeing with Tab. 1 results, the average pixel-to-pixel correlation shows both models have a remarkable performance where none of the channel combinations had a performance lower than 0.99. These results outperform all the channel combinations presented in Lim et al. (2021), which tries several combinations of EUV channels translations using the DL method "Model B" from Park et al. (2019) and Isola et al. (2017).

Notably Lim et al. (2021) did not report on other metrics we can use to compare the quality of the corresponding synthetic images. We demonstrate in the following analysis that the elevate visual quality of the images and the excellent pixel-to-pixel Pearson correlation values are not enough to guarantee the absence of artifacts which may impact the scientific utility of the synthetic images. This is illustrated in Fig. 2, Tab. 4 and Fig. 6. Whether the discrepancies between real and synthetic images are sufficiently small to neglect clearly depends on the science case. For this reason, we argue that metrics such as covariance between real and synthetic image and accuracy by intensity should be standard metrics to be considered when reporting on models for the synthesis solar images.

While useful to evaluate the overall performance of the algorithm, the aggregated metrics do not provide insights about the range of validity of the algorithm and the reasons behind its errors. Firstly, to understand how to possibly improve the model, and secondly, to clarify what could be a concrete use of the algorithm in future missions, it is helpful to evaluate the prediction uncertainty at different intensities. For all the permutations, in Fig.2 we show the uncertainty on the predicted count rate (top) and the pixel distributions (bottom) as a function of the real count rate. These plots highlight three important factors:

- The algorithm does well over about three orders of magnitude of true count rate (intensity) and it largely increases its error when trying to predict the highest and lowest count rates. It means the global metrics would be much more favorable if removing these extreme pixels. This behavior also implies the algorithm could be used with confidence for applications that do not require accuracy on the most extreme values of count rates.

- The difficulty in predicting the pixels with the highest and the lowest count rate is not surprising if looking at the count rate distributions (histograms in Fig.2). The tails of the distributions, where the model's accuracy and uncertainty increase, are severely underrepresented in the distribution. This implies the image-to-image translation algorithm has not been trained or trained in a very limited way on pixels having these count rate values. This observation also provides a clear indication of which strategies can improve the algorithm performance, i.e., techniques to compensate the magnitude imbalance rather than larger architectures.

- Applying root scaling to the input images during the training tends to improve the results for low count rate pixels and reduces the uncertainty on the prediction. Some channels (193, 211 Å) are more positively impacted than others by this
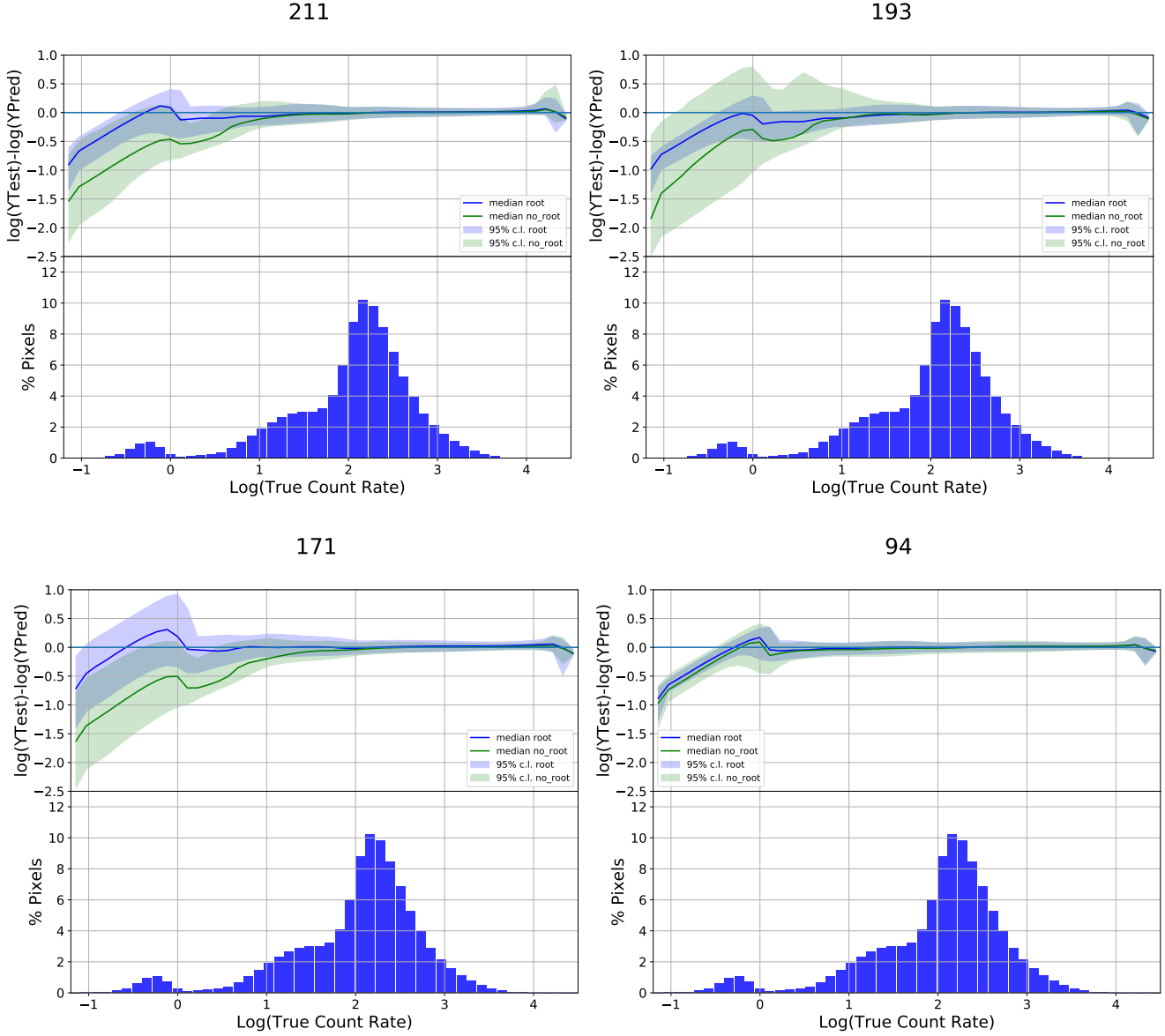
Figure 2. Predicted Intensity vs Real Intensity for each of the four channels, for all the pixels contained in the 667 images on the test set. From top to bottom: 211, 193, 171, 94 Å channels. For each channel: the **top** plot shows the error on the predicted count rate as a function of the real count rate in $log10$. The error band represents the standard deviation, the line corresponds to the median. In green the standard U-net model, in blue the same architecture with square root scaling applied to the input images; the **bottom** plot shows the histogram of the pixel count rate distributions over the test set. The model performs well over three orders of magnitude but its accuracy degrades quickly in the extreme regions where fewer pixels are available.

change. This behavior is explained by the fact root scaling improves the sensitivity to small values during the training. We hypothesize that further exploration of different scaling strategies for the training can also be a way to extend the accuracy of the algorithm over more orders of magnitude.

Examples of the resulting recovered images when adopting the DNN architecture described in Sec. 3 and a model with root scaled input, is given in Fig. 3 and Fig. 4. The root scaling is reverted in the illustrated images. The first are example of reconstructions on a quiet day, where the Sun shows less activity, while the second are during the well known Valentine's Day flare. In these figures, the first column corresponds to the original images, while the second column corresponds to the ones generated by the DNN. Based on visual inspection, the synthetic image reproduces the morphology of coronal loops in the ground truth image for channels 211 and 171 Å, and the prediction is instead a bit less realistic for 193 Å for both quiet and active
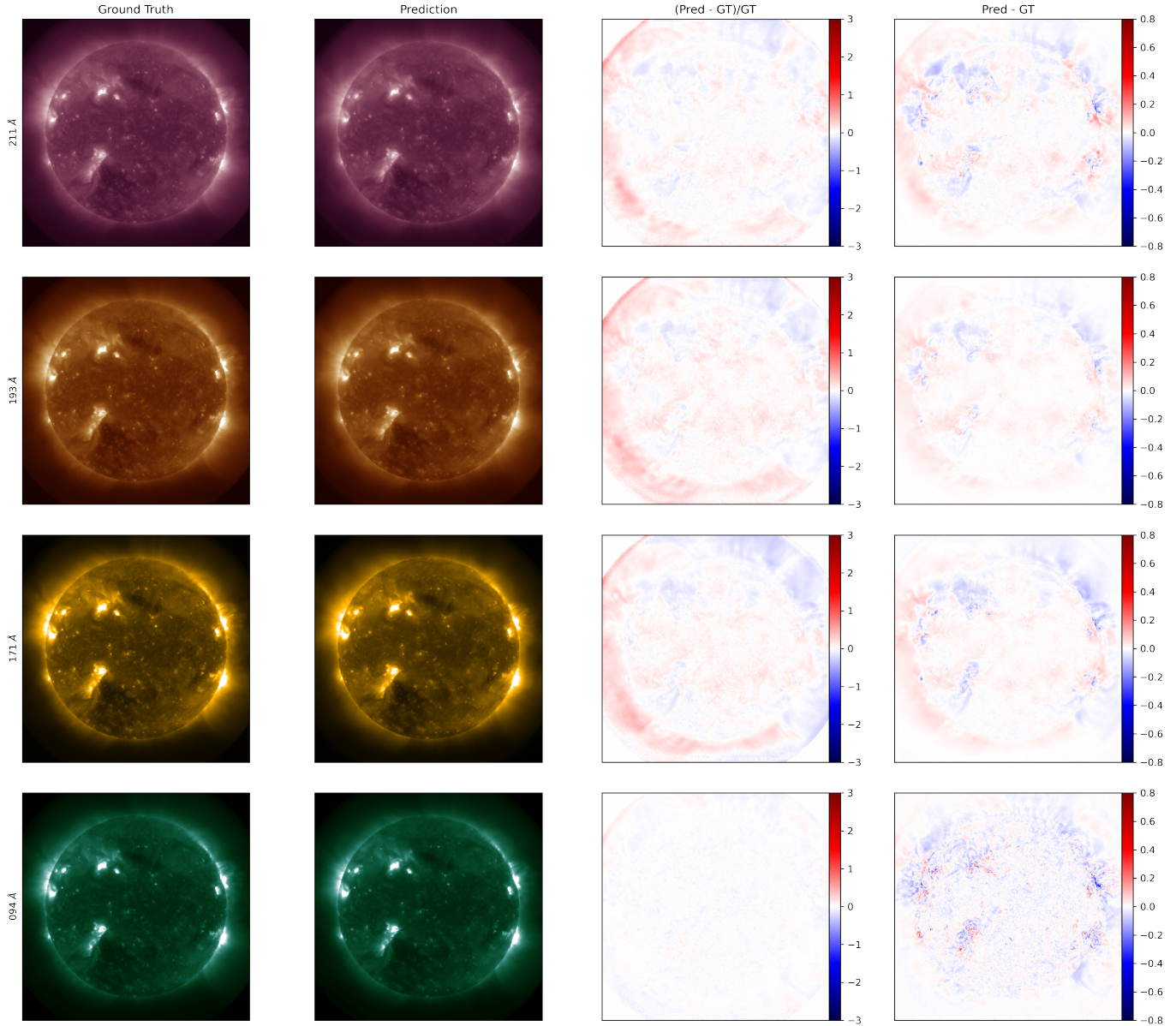
Figure 3. Real versus synthetic images on a quiet timestamp (2011-02-10 00:00:00 UT) when using model with root scaling. From left to right: real image, image synthesized by looking at the other 3 channels, residuals relative to the GT value

and difference between the two images. From top to bottom: 211, 193, 171, 94 Å channels.

days. Clearly during the quiet day the all three channels have better performance than in the Valentine's day. It also interesting to observe that 94 Å is the best performing channel during the quiet day, but the worst performing channel during the active day. This aligns to the results showed at Tab.1 and 2. It is unsurprising since the input AIA channels 94, 171 and 193 Å channels have sensitivity to the plasma observed in the 211 Å channel. This outperforms previous results in Park et al. (2019), where a conditional generative adversarial network (CGAN) had been trained to translate HMI magnetograms to AIA images.

In the third column of Fig. 3 and Fig. 4 we included the residuals relative to the real image and in the fourth column of the same figures we display the differences between the real and generated images. Dark blue and bright red correspond to the regions where the differences are the largest, and can be seen to be located where the active regions (shown as the brightest regions in the original and generated images) are.

Interestingly, the model well reconstructs coronal holes (CH) in both the active and quiet Sun cases described above, despite the low signal in these regions. This could be due to the fact that the physics of these regions is easier to model than active region
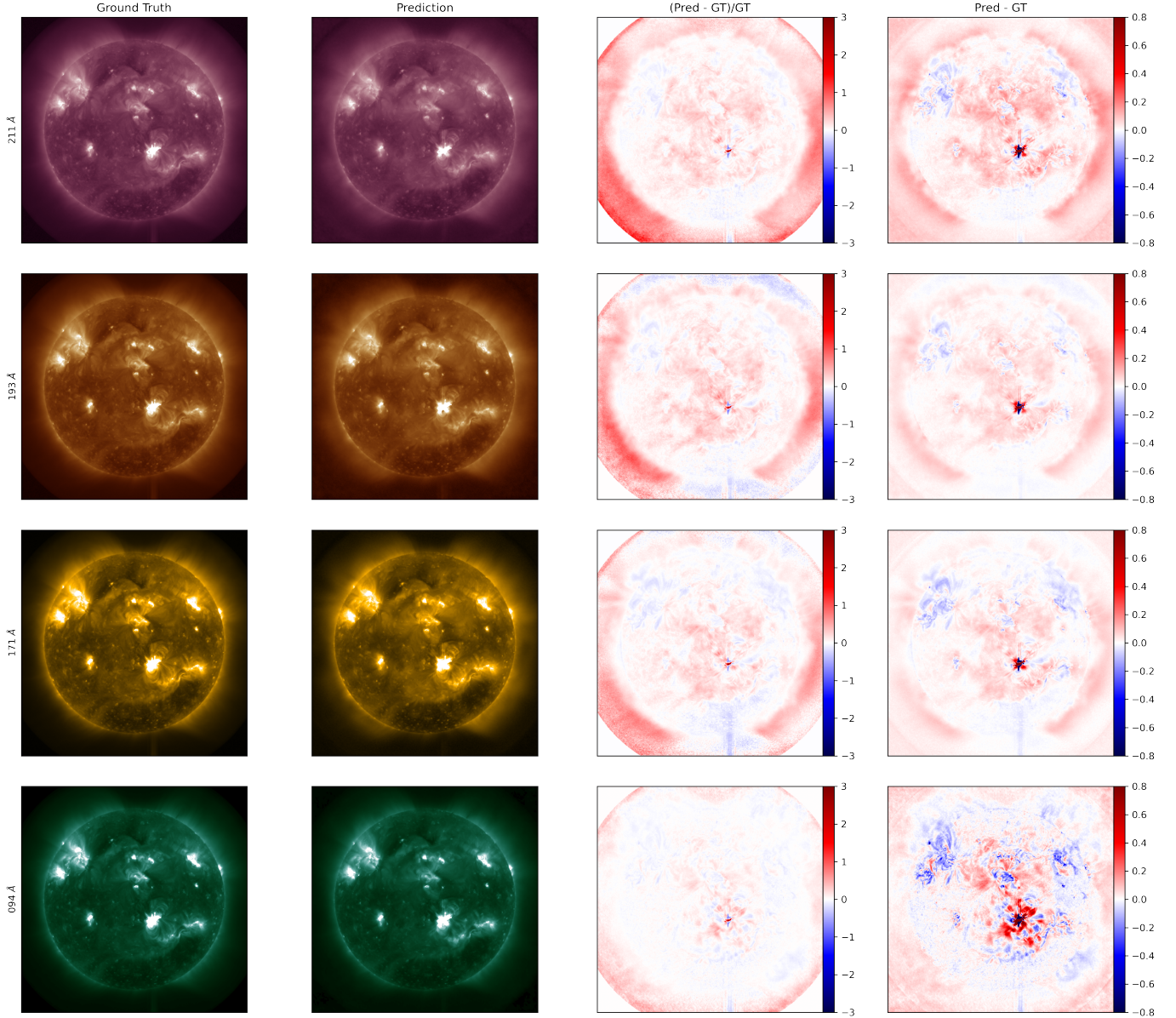
Figure 4. Real versus synthetic images during a flare (2011-02-15 02:00:00 UT) when using model with root scaling. From left to right: real image, image synthesized by looking at the other 3 channels, residuals relative to the GT value and

difference between the two images. From top to bottom: 211, 193, 171, 94 Å channels.

coronal loops as the field lines are open and have relatively simpler configuration. A quantitative comparison between CH and full-disk is shown in Fig. 5 for channel 193 Å (for the quiet Sun data represented in Fig. 3), where CHs are most distinctly visible due to their contrast. The segmentation mask identifies the CH regions based on the simple but robust adaptive intensity threshold technique (similar to the technique employed in Rotter et al. 2012, 2015), and the histograms show the difference between the ground truth and the predicted intensities (on a pixel-by-pixel basis) for pixels both within the CH boundaries and the full-disk. It is to be noted that the segmentation mask is constructed for both the predicted and ground-truth images independently using the same intensity threshold criterion. Clearly, the predicted AIA intensities are well constrained not just over the full disk but also on the relatively quieter CH areas.

In Tab. 4 and Tab. 5 we report the reconstruction error on the covariance between channels, over four hours, for the case 94, 171, 193 Å to 211 Å in correspondence of a flare and on a normally quiet day. Not surprisingly, in light of the results above, the reconstructed covariance has great accuracy (less than 1% of error) on a quiet day but its error increases in several orders
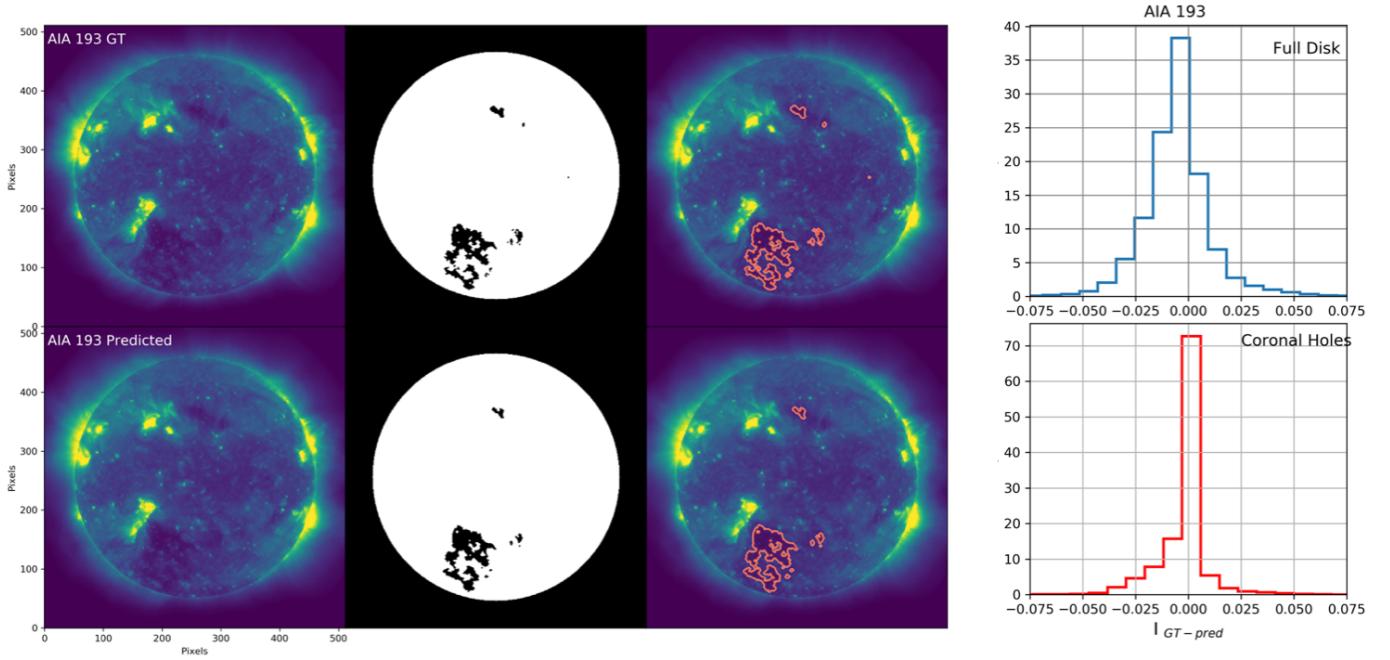
Figure 5. Coronal Holes for channel 193 Å. On the left the segmentation mask obtained by thresholding , on the right histograms showing the difference between the ground truth and the predicted intensities (on a pixel-by-pixel basis) for both the pixels within the CHs and for the full-disk.

Timestamp	Channel	True Cov	Pred Cov	Diff	%Diff
2011-2-15-0-0	94	0.278	0.256	0.022	7.9
2011-2-15-1-0	94	0.262	0.246	0.016	5.9
2011-2-15-2-0	94	13.9	92.3	-78.5	-565
2011-2-15-3-0	94	1.69	1.54	0.150	8.9
2011-2-15-4-0	94	0.392	0.375	0.017	4.4
2011-2-15-0-0	171	0.117	0.115	0.002	2.1
2011-2-15-1-0	171	0.114	0.112	0.002	1.9
2011-2-15-2-0	171	1.29	13.1	-11.8	-913
2011-2-15-3-0	171	0.186	0.178	0.008	4.3
2011-2-15-4-0	171	0.139	0.136	0.003	2.3
2011-2-15-0-0	193	0.048	0.047	0.001	1.4
2011-2-15-1-0	193	0.047	0.047	0.001	1.3
2011-2-15-2-0	193	0.191	0.605	-0.414	-216
2011-2-15-3-0	193	0.065	0.063	0.003	4.0
2011-2-15-4-0	193	0.055	0.054	0.001	2.1

Table 4. Errors in reconstructing the covariance between 211 Å and the other 3 channels when using the synthetically produced image for 211 Å in correspondence of a highly energetic event (Valentine's Day flare on 2011-2-15:1:50:00 UT). Interestingly the reconstructed covariance has a much higher error than what seen in a quiet period, cf. Tab. 5, at least 1h before the flare has been detected.

Timestamp	Channel	True Cov	Pred Cov	Diff	%Diff
2011-2-13-0-0	94	0.1506	0.1504	0.0002	0.1
2011-2-13-1-0	94	0.1672	0.1654	0.0018	1.1
2011-2-13-2-0	94	0.1601	0.1588	0.0013	0.8
2011-2-13-3-0	94	0.1713	0.1718	-0.0004	-0.3
2011-2-13-4-0	94	0.1652	0.1650	0.0002	0.1
2011-2-13-0-0	171	0.1213	0.1210	0.0002	0.2
2011-2-13-1-0	171	0.1261	0.1254	0.0007	0.5
2011-2-13-2-0	171	0.1227	0.1223	0.0004	0.3
2011-2-13-3-0	171	0.1241	0.1244	-0.0002	-0.2
2011-2-13-4-0	171	0.1226	0.1219	0.0007	0.6
2011-2-13-0-0	193	0.0449	0.0448	0.0000	0.1
2011-2-13-1-0	193	0.0470	0.0468	0.0002	0.4
2011-2-13-2-0	193	0.0439	0.0439	-0.0000	-0.1
2011-2-13-3-0	193	0.0465	0.0468	-0.0003	-0.7
2011-2-13-4-0	193	0.0471	0.0470	0.0001	0.2

Table 5. Errors in reconstructing the covariance between 211 Å and the other 3 channels when using the synthetically produced image for 211 Å in correspondence of a quiet period few days before Valentine's Day flare. The percentage difference is below 1% for all the channels.

of magnitude in correspondence of the extreme event. The results reported in Tab. 4 and Tab. 5 are obtained using the model without square root scaling, the most sensitive to extreme values. They should therefore be interpreted as an upper bound on the error that a similar image translation would have. With the aim of better understanding the source of error, in addition to the standard covariance, we compute a covariance map with spatial mean on a rolling squared window of 20×20 pixels, see Eq. 5 for definition. The resulting covariance map in correspondence of a flare is shown in Fig. 6. The map clearly shows the error of the model is localised in the area of the flare and it does not affect the rest of the map, in agreement with the localized reconstruction error shown in Fig. 4. This result confirms the results of the "virtual telescope" would be accurate for most of the pixels, also in presence of an extremely energetic event, but for the specific area where the event happens. Similar results hold for the covariance in other channel permutations.

Incidentally, the above covariance result suggests an increase in its reconstruction error could also be used as a method for early detection of flares as the error starts to increase before the actual flare's event. Variations in reconstruction errors are commonly used in machine learning as anomaly detection methods (e.g. An & Cho (2015); Zhou & Paffenroth (2017). While directly detecting an increase in the data count could be found to be more effective, the sensitivity to non-linearity of the reconstruction task could produce a stronger or complementary signal that we think is interesting to consider in future work.

6. CONCLUDING REMARKS

In this study, we analyzed the performance of an image-to-image translation DNN model in accurately reconstructing extreme ultra-violet images from a solar telescope, focusing on the permutations of four channels. We found that the reconstruction error is extremely accurate over three orders of magnitude in pixel intensity (count rate) and it rapidly increases when considering extremely low and high range of intensities. This behavior is explained by the pixel count rate distribution in the training set, the rarer the value the more difficult for the DNN to provide an accurate prediction. Similarly, when looking at the reconstruction error on the covariance at different times, we found the model can synthetically predict the covariance with less than 1% of error on quiet days but its performance is severely affected in correspondence of flares, in the active regions.

The results show that a virtual telescope would produce accurate estimations on a range of intensities but, if built following the methodology here described, would not be able to accurately reproduce extremely energetic events like flares. How and in which limit the reconstruction error for such specific events could be improved is an area of research that we leave for future work. The rareness of flare events poses a challenge in training machine learning algorithms to accurately reproduce such events. Based on the results above, we think adopting oversampling techniques and different scaling strategies would improve at least in
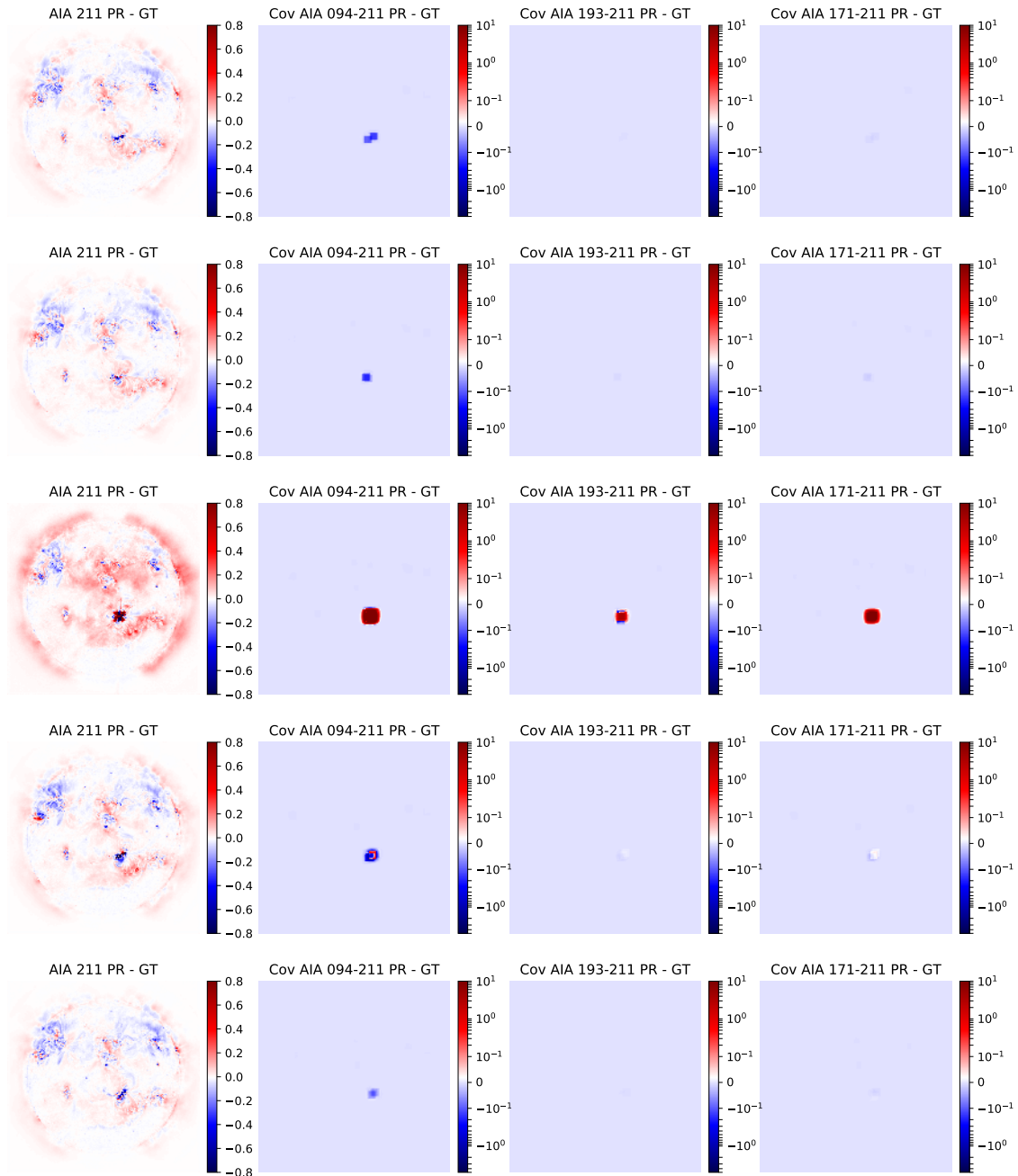
Figure 6. Reconstruction error on the covariance in correspondence of Valentine's Day flare. From left to right: Difference between the ground truth and the predicted images. Differences between the real and predicted covariance maps between 211 Å - the predicted channel - and each of the input channels. From top to bottom: each row corresponds to a different timestamp at interval of 1h. The 3rd line is the closest to the time of the flare.

some measure the performance. To overcome this challenge, other strategies like automatic detection of anomalies could also be adopted in combination with image-to-image translation, in the design of a virtual solar telescope.

In this paper, we did not explore the dependence of model performance from spatial resolution. In principle smaller subpixel scales could have information that improve the global performance of image synthesis and we think this is an important question to be addressed in future work. Importantly, we expect the deterioration of the synthetic accuracy for rare events to happen regardless of the adopted scale because it is caused by the scarcity of examples for training.

APPENDIX

A. SCALING UNITS FOR EACH AIA CHANNEL

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

Table 6. Table of AIA channel scaling units.

AIA channel (Å)	$\overline{Y_{test}}$
94	26
171	0.13
193	0.087
211	0.26

Table 7. Table of average values over the test set after scaling by channel

B. CODE DESCRIPTION

In this appendix we describe the modular software used to produce the analysis and made freely available online on GitHub under GPL licence. Users are invited to consult the code documentation for additional detail.

- *src/sdo* - contains all the modules required to run the pipeline plus additional functionalities that can be used as standalone library to interact with the SDO-ML dataset v1.

- *config* - contains some configuration templates.

- *scripts* - contains some analysis scripts specific to the paper, they can be used to reproduce the results.

- *notebooks* - contains some notebooks specific to the paper that can be used to reproduce some of the plots in the paper and some examples to show how to use some functionalities (e.g. how to use the dataloader to load timestamps of interest).

The most relevant modules under *src* are:

- *src/sdo/datasets/sdo_dataset.py* this module contains the *SDO_Dataset* class, a custom Dataset class compatible with *torch.utils.data.DataLoader*. It can be used to flexibly load a train or test dataset from the SDO local folder. Data can be selected according to the 3 criteria:

  * asking for a specific range of years and a specific frequency in months, days, hours, minutes

  * passing a file that contains all the timestamps of interest

  * passing two timestamps ranges and a desired step

  This class assumes a pre-computed inventory of the SDO dataset exists.

- *src/sdo/pipelines/virtual_telescope_pipeline.py* this module contains the *VirtualTelescopePipeline* class, the class that contains all the training and test logic of the modeling approach. This class also handles the metrics logging and the files saving. Beyond being used for reproducing the results of this work, this class can be used as example of how to integrate the dataloader above with other PyTorch models for a different set of experiments.

- *src/sdo/parse_args.py* this module contains the description of all the parameters that can be passed as input to the pipeline and their default values.

C. ADDITIONAL FIGURES

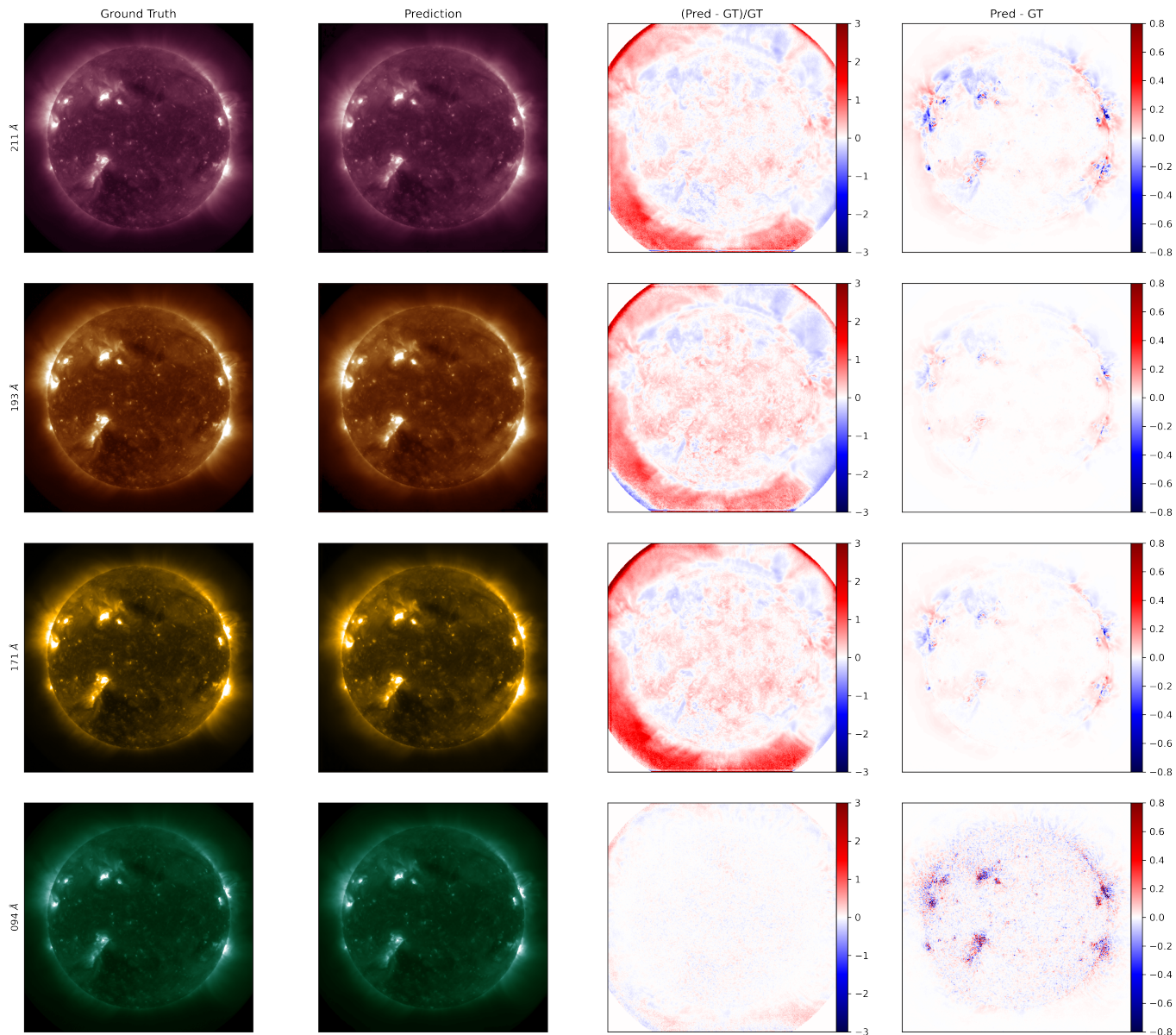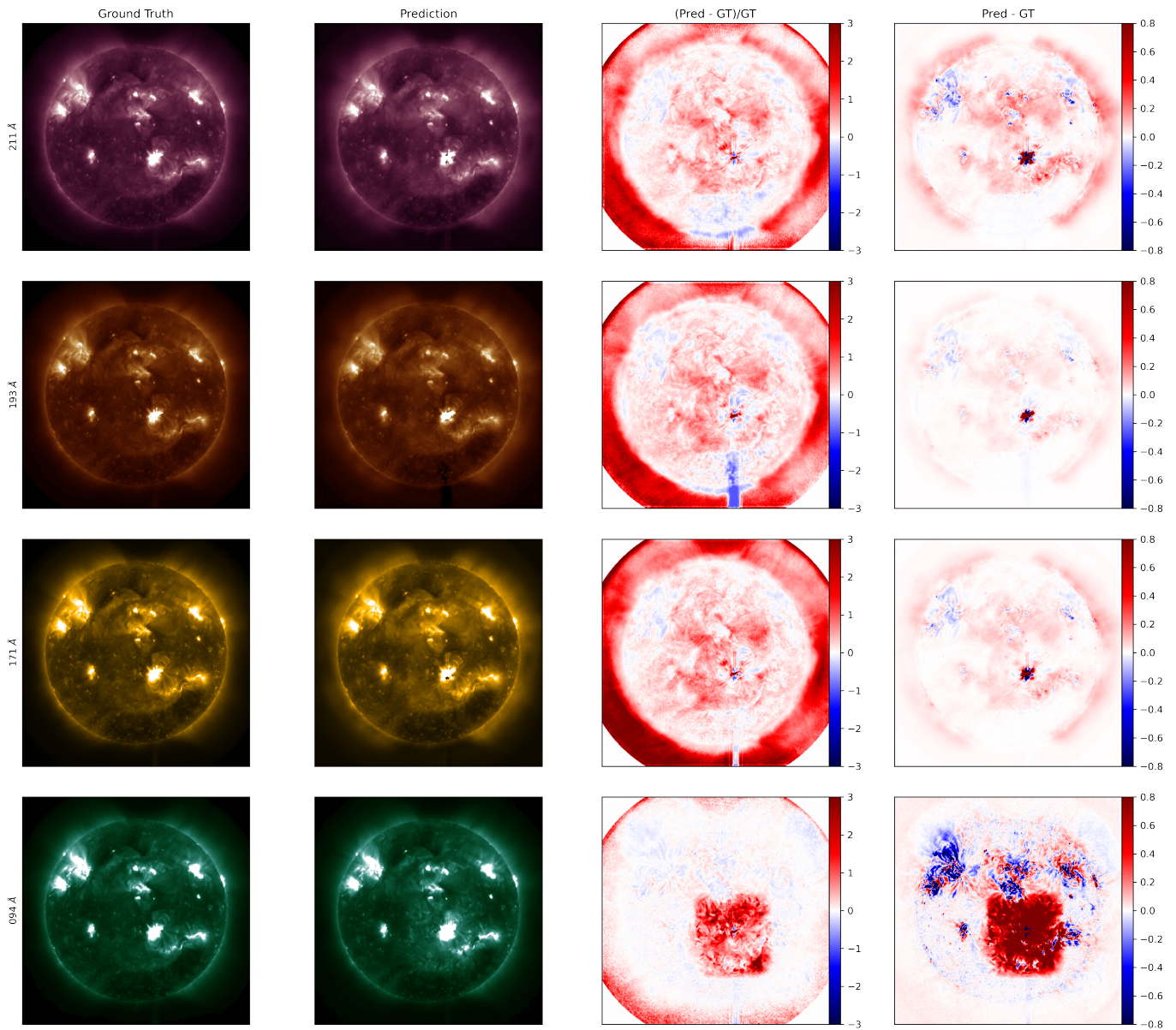In this appendix we report some additional results not included in the main text.

Figure 7. Real versus synthetic images on a quiet timestamp (2011-02-10 00:00:00) when using model without root scaling. From left to right: real image, image synthesized by looking at the other 3 channels, residuals relative to the GT value and difference between the two images. From top to bottom 211, 193, 171, 94 Å channels.

Figure 8. Real versus synthetic images during a flare (2011-02-15 02:00:00) when using model without root scaling. From left to right: real image, image synthesized by looking at the other 3 channels, residuals relative to the GT value and difference between the two images. From top to bottom 211, 193, 171, 94 Å channels.

REFERENCES

An, J., & Cho, S. 2015

Chetlur, S., Woolley, C., Vandermersch, P., et al. 2014, arXiv
e-prints, arXiv:1410.0759. https://arxiv.org/abs/1410.0759

Cheung, M. C. M., Boerner, P., Schrijver, C. J., et al. 2015, The
Astrophysical Journal, 807, 143,
doi: 10.1088/0004-637X/807/2/143

Díaz Baso, C. J., & Asensio Ramos, A. 2018, Astronomy &
Astrophysics, 614, A5, doi: 10.1051/0004-6361/201731344

Galvez, R., Fouhey, D. F., Jin, M., et al. 2019, The Astrophysical
Journal Supplement Series, 242, 7,
doi: 10.3847/1538-4365/ab1005

Goodfellow, I., Bengio, Y., & Courville, A. 2016, Deep learning
(MIT press)

Hunter, J. D. 2007, Computing in Science and Engineering, 9, 90,
doi: 10.1109/MCSE.2007.55

Isola, P., Zhu, J., Zhou, T., & Efros, A. A. 2017, in 2017 IEEE
Conference on Computer Vision and Pattern Recognition, CVPR
2017, Honolulu, HI, USA, July 21-26, 2017 (IEEE Computer
Society), 5967–5976, doi: 10.1109/CVPR.2017.632

Kingma, D. P., & Ba, J. 2014, arXiv e-prints, arXiv:1412.6980.
https://arxiv.org/abs/1412.6980

Lemen, J. R., Title, A. M., Akin, D. J., et al. 2012, Solar Physics,
275, 17, doi: 10.1007/s11207-011-9776-8

Lim, D., Moon, Y.-J., Park, E., & Lee, J.-Y. 2021, The
Astrophysical Journal Letters, 915, L31,
doi: 10.3847/2041-8213/ac0d54

Maas, A. L., Hannun, A. Y., & Ng, A. Y. 2013, in ICML Workshop
on Deep Learning for Audio, Speech and Language Processing

Park, E., Moon, Y.-J., Lee, J.-Y., et al. 2019, The Astrophysical
Journal, 884, L23, doi: 10.3847/2041-8213/ab46bb

Paszke, A., Gross, S., Chintala, S., et al. 2017, in NeurIPS Autodiff
Workshop

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach.
Learn. Res., 12, 2825.
http://jmlr.org/papers/v12/pedregosa11a.html

Pesnell, W., Thompson, B., & Chamberlin, P. 2012, solphys, 275,
3, doi: 10.1007/s11207-011-9841-3

Ronneberger, O., Fischer, P., & Brox, T. 2015, in Medical Image
Computing and Computer-Assisted Intervention – MICCAI
2015, ed. N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi
(Cham: Springer International Publishing), 234–241,
doi: 10.1007/978-3-319-24574-4_28

Rotter, T., Veronig, A. M., Temmer, M., & Vršnak, B. 2012, SoPh,
281, 793, doi: 10.1007/s11207-012-0101-y

—. 2015, SoPh, 290, 1355, doi: 10.1007/s11207-015-0680-5

Salvatelli, V., Bose, S., Neuberg, B., et al. 2019, arXiv e-prints,
arXiv:1911.04006. https://arxiv.org/abs/1911.04006

Salvatelli, V., Neuberg, B., Dos Santos, L. F. G., et al. 2022, ML
pipeline for Solar Dynamics Observatory (SDO) data,
v0.3-alpha, Zenodo, doi: 10.5281/zenodo.6954828

Szenicer, A., Fouhey, D. F., Munoz-Jaramillo, A., et al. 2019,
Science Advances, 5, doi: 10.1126/sciadv.aaw6548

van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, Comput. in
Sci. Eng., 13, 22, doi: 10.1109/MCSE.2011.37

Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, Nature
Methods, 17, 261,
doi: https://doi.org/10.1038/s41592-019-0686-2

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. 2004,
IEEE Transactions On Image Processing, 13, 600

Wes McKinney. 2010, in Proceedings of the 9th Python in Science
Conference, ed. Stéfan van der Walt & Jarrod Millman, 56 – 61,
doi: 10.25080/Majora-92bf1922-00a

Zhou, C., & Paffenroth, R. C. 2017, in Proceedings of the 23rd
ACM SIGKDD International Conference on Knowledge
Discovery and Data Mining, KDD '17 (New York, NY, USA:
Association for Computing Machinery), 665–674,
doi: 10.1145/3097983.3098052