













Physically Motivated Deep Learning to Superresolve and Cross Calibrate Solar Magnetograms

Andrés Muñoz-Jaramillo^{1,10} , Anna Jungbluth^{2,10} , Xavier Gitiaux^{3,10} , Paul J. Wright^{4,5,10} , Carl Shneider⁶ ,
Shane A. Maloney^{4,7} , Atılım Güneş Baydin⁸ , Yarin Gal⁸ , Michel Deudon⁹ , and Freddie Kalaitzis⁸ 

¹ Southwest Research Institute, 1050 Walnut Street, Ste 300, Boulder, CO 80020, USA; amunozj@boulder.swri.edu

² University of Oxford, Parks Road, Oxford, OX1 3PJ, UK

³ George Mason University, 4400 University Drive, Fairfax, VA 22030, USA

⁴ Astronomy & Astrophysics Section, Dublin Institute for Advanced Studies, Dublin, D02 XF86, Ireland

⁵ W.W. Hansen Experimental Physics Laboratory, Stanford University, Stanford, CA 94305, USA

⁶ Dutch National Center for Mathematics and Computer Science (CWI), Science Park 123, Amsterdam, 1098, The Netherlands

⁷ School of Physics, Trinity College Dublin, College Green, Dublin 2, Ireland

⁸ University of Oxford, Parks Road, Oxford, OX1 3QG, UK

⁹ Element AI, London, UK

Received 2023 July 10; revised 2023 November 6; accepted 2023 December 4; published 2024 March 26

Abstract

Superresolution (SR) aims to increase the resolution of images by recovering detail. Compared to standard interpolation, deep learning-based approaches learn features and their relationships to leverage prior knowledge of what low-resolution patterns look like in higher resolution. Deep neural networks can also perform image cross-calibration by learning the systematic properties of the target images. While SR for natural images aims to create perceptually convincing results, SR of scientific data requires careful quantitative evaluation. In this work, we demonstrate that deep learning can increase the resolution and calibrate solar imagers belonging to different instrumental generations. We convert solar magnetic field images taken by the Michelson Doppler Imager (resolution $\sim 2'' \text{ pixel}^{-1}$; space based) and the Global Oscillation Network Group (resolution $\sim 2'' 5 \text{ pixel}^{-1}$; ground based) to the characteristics of the Helioseismic and Magnetic Imager (resolution $\sim 0'' 5 \text{ pixel}^{-1}$; space based). We also establish a set of performance measurements to benchmark deep-learning-based SR and calibration for scientific applications.

Unified Astronomy Thesaurus concepts: [Solar magnetic fields \(1503\)](#); [The Sun \(1693\)](#); [Solar physics \(1476\)](#); [Solar active regions \(1974\)](#)

1. Introduction

Over the last 50 yr, space- and ground-based instruments have mapped the solar surface magnetic field (Figure 1). These images, known as magnetograms, have significantly advanced our understanding of solar magnetism (Hathaway 2010), understanding of the solar corona (Linker et al. 1999), and prediction of space weather events (Tóth et al. 2005). Magnetograms are constructed from measurements of spectral polarization (Borrero & Ichimoto 2011), which are compared to models of the solar atmosphere to find an optimal fit between an estimated local magnetic field and the observed spectral properties.

Despite the wealth of archival data, differences in resolution, spectral inversion techniques, instrument noise levels, or other instrument properties prevent us from easily combining data across instruments to study magnetic field structures over multiple solar cycles (Figure 1) (e.g., Díaz Baso & Asensio Ramos 2018). Compared to traditional cross-calibration techniques such as pixel-to-pixel comparison (Liu et al. 2012), histogram equalization (Riley et al. 2014), or harmonic scaling (Virtanen & Mursula 2019), machine learning (ML) has

previously been shown to successfully calibrate magnetograms (Guo et al. 2021; Higgins et al. 2022).

Superresolution (SR) is an image-processing technique that aims to increase the resolution of images by recovering subpixel detail (Shukla et al. 2020). The information used for recovering detail can come from subpixel shifts provided by sequences of images (frequency domain), or by a good understanding of the degradation processes, including blurring, that cause the loss of detail (i.e., atmospheric seeing, point-spread function, etc.) (Shukla et al. 2020). In the case of applications with sufficient and representative low-resolution (LR) and high-resolution (HR) samples, context can provide an additional source of information (i.e., the knowledge that all LR images belong to a specific category). Convolutional neural networks (CNNs) are especially suited for this type of application due to their ability to empirically map the underlying connections between an image pixel and those surrounding it (Yang et al. 2019). Furthermore, neural networks can learn features and feature relationships that are inherent to the data domain as a whole. For example, previous work (Dahl et al. 2017) has shown that neural networks successfully superresolve downsampled images of human faces. Important features (i.e., a mouth, nose, and two eyes) were reconstructed correctly and their spatial relationships were preserved. The majority of applications of CNNs for SR involve natural images (i.e., images with three color channels representing red, green, and blue). These approaches use training metrics that are tailored toward SR outcomes that are optimal for human visual perception. In other words, their

¹⁰ These authors contributed equally to this work.

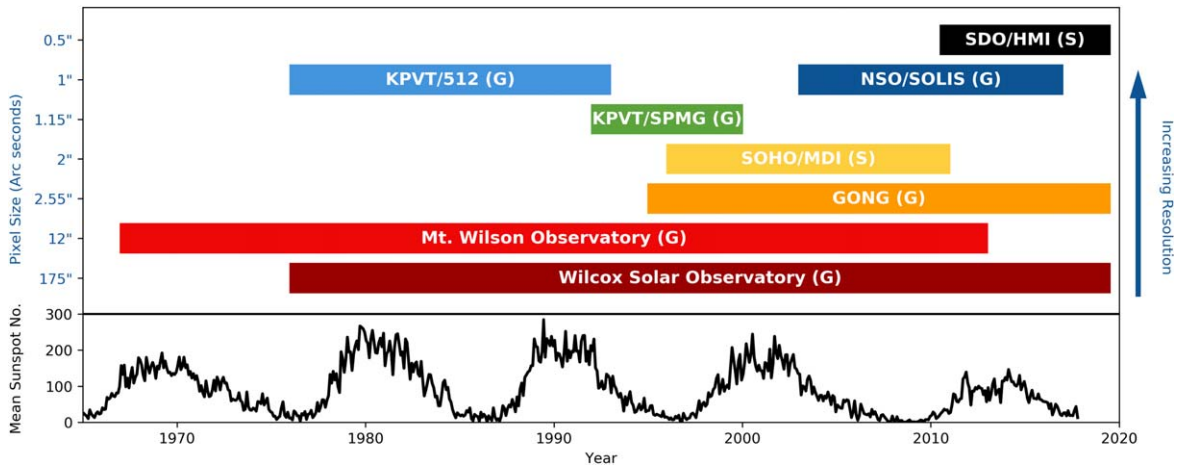


Figure 1. Overview of ground (G) and space-based (S) observations of the Sun. The top panel shows the pixel size in arcseconds, and the operation time span of eight different instruments. The bottom panel shows the variation of the mean sunspot number over the last 50 yr. To date, the HMI, on board SDO, provides the highest resolution full-disk magnetograms. This is followed by the 512 channel magnetograph of the Kitt Peak Vacuum Telescope and the Synoptic Optical Long-term Investigations of the Sun by the NSO, GONG, the Mount Wilson Observatory, and finally the Wilcox Solar Observatory.

objective is to produce images that look right to the human observer (Wang et al. 2020).

Deep-learning applications of SR for scientific data have tremendous potential due to their ability to simultaneously superresolve (recover scientifically accurate detail in images) and cross-calibrate (correct systematic differences between measurement instruments). However, scientific images have a significantly larger dynamic range than natural images. For example, pixels in magnetic field images can assume real values spanning several orders of magnitude, while pixels in natural images take discrete values over a fixed range. Another important difference is the fact that magnetic fields have two polarities that cancel out at lower resolutions, while in natural images intensity is strictly positive. Furthermore, since the Sun is a three-dimensional object projected onto a two-dimensional image, solar magnetic field images suffer from projection effects, especially close to the solar limb. Therefore, metrics in computer vision that are tailored toward estimating the perceptual quality of a natural image do not necessarily capture how well superresolved magnetograms represent the physical properties of the Sun’s magnetic field.

Our work has three main objectives: (1) to demonstrate that a deep-learning approach can leverage the information present in astronomical images to recover detail in LR images while maintaining their scientific accuracy; (2) to show how super-resolving a scientific image via deep learning homogenizes instrument properties and adds value compared to simple calibration at the same resolution, (3) to establish a set of quantitative performance measurements that can be used to benchmark the performance of different SR algorithms for astronomical images, as well as to benchmark the performance of future applications of SR to the physical sciences.

Previous SR approaches for solar magnetograms relied on physics-based models to simulate HR magnetograms as the ground truth (Díaz Baso & Asensio Ramos 2018). Other deep-learning approaches superresolved a down-scaled version of the same instrument, e.g., using generative adversarial networks (Rahman et al. 2020). The novelty in our approach is that we use deep learning to cross-calibrate and superresolve across different instruments. We cross-calibrate and super-resolve line-of-sight (LOS) magnetograms from the Michelson Doppler Imager (MDI; $\sim 2''$ pixel $^{-1}$; space-based) on board the

Solar and Heliospheric Observatory (SOHO; Scherrer et al. 1995), as well as LOS magnetograms taken by the National Solar Observatory’s (NSO) Global Oscillation Network Group (GONG; $\sim 2.5''$ pixel $^{-1}$; ground based, Harvey et al. 1988) to the $\sim 0.5''$ pixel $^{-1}$ resolution of magnetograms taken by the Helioseismic and Magnetic Imager (HMI; last generation, space based, Scherrer et al. 2012) on board the Solar Dynamics Observatory (SDO; Pesnell et al. 2012). Our results indicate that deep learning can leverage the complex information and context present in magnetograms. This allows us to encode the structure of the magnetic field in a lower dimensional latent space and then map magnetograms from one instrument to the other. Additionally, we show that superresolved magnetograms are better at capturing the physical properties of Space Weather HMI Active Region Patches (SHARPs; Bobra et al. 2014) than cross-calibration alone.

2. Data

In this work, we use solar magnetograms from NSO/GONG (Harvey et al. 1988), SOHO/MDI (Domingo et al. 1995; Scherrer et al. 1995), and SDO/HMI (Pesnell et al. 2012; Scherrer et al. 2012; Schou et al. 2012). NSO/GONG and SOHO/MDI act as our *source* instruments and SDO/HMI as our *target* instrument. NSO/GONG is a ground-based instrument currently used as the main operational magnetograph for NOAA’s Space Weather Prediction Center. SOHO/MDI is SDO/HMI’s predecessor featuring a lower cadence, sensitivity, and resolution.

2.1. Data Splits

To train our SR architecture, we leverage overlapping observation periods between MDI and HMI (2010–2011) and between GONG and HMI (2010–2019), which provides us with ~ 9000 ($\sim 19,000$) MDI-HMI (GONG-HMI) magnetogram pairs. We split the data into training/validation/test sets by randomly allocating 10 months to the training set, 1 month to the validation set, and 1 month to the test set for each overlapping year (see Table 1).

In the case of GONG-HMI, we only use even years (2010, 2012, 2014, 2016, and 2018) for this work to keep the data volume manageable. The test set comprises magnetograms

Table 1
Data Split Used for Model Training, Validation, and Testing

	Training/Validation	Testing
MDI → HMI	2010 April to 2011 April, excluding the test set	2010 June, 2011 March
GONG → HMI	2012, 2014, 2016, 2018, excluding the test set	2011 March, 2012 April 2014 December, 2016 February 2018 November

taken at a 96 minute cadence for MDI, and a 10 minute cadence for GONG. Across all experiments, we choose 2010 June and 2011 March as our test months.

Each full-disk magnetogram is split into small patches (discussed below) to ensure that model training is done on spatial scales and features that evolve over hours. This helps break any correlations that may happen from one solar rotation to the next. Because of this, we do not provide a time buffer between the training, validation, and test months to ensure data independence. For other tasks involving larger areas, or for performing full-disk conversions, the slow evolution of the global magnetic field could unintentionally leak into the test set if a sufficient time buffer is not provided.

2.1.1. Data Preprocessing

The data is preprocessed according to the following three steps:

1. Standardization of the Sun’s orientation by rotating solar north to image north.
2. Standardization of the detector’s angular resolution and solar angular radius. We use the *reproject* package¹¹ to ensure that the resolution of our source and target instruments are integer multiples of each other.
3. Splitting of full-disk magnetograms into small patches, followed by co-alignment of each source target pair.

In detail, each full-disk magnetogram is split into 1024 patches of size 32×32 pixels for the LR input, and 128×128 pixels for the HR target. For each LR source patch (32×32 pixels), a search is performed within an extended target patch window (256×256 pixels) to find the optimal 128×128 pixel area that best matches the source. This helps us account for slight displacements due to solar rotation, as well as optical aberrations. We find that this template matching leads to significantly better performance than when the approximate alignment of patches is performed.

2.2. Data Augmentation

The observations used in this work cover one solar cycle, meaning that our data set contains systematic polarity orientations for both positive and negative fields and their relative distributions across the solar disk. These polarity orientations change from cycle to cycle and from hemisphere to hemisphere, increasing the risk of our network learning unintended structures and patterns. To avoid this, we augment data through random polarity flips, as well as north–south, and east–west reflections.

3. Methodology

3.1. Neural Network Architecture

The deep-learning model used in this work was adapted from the HighRes-net model (Deudon et al. 2020)¹² (see Figure 2), as this model has shown great performance for SR of Earth observation data. Our model input consists of two channels, a magnetogram patch and a location channel. The location channel captures the normalized radial distance of each pixel to the disk center and provides the network with the necessary information for estimating projection and foreshortening effects at the solar limb. The data is encoded into 64 channels through a series of convolution operations shown in the *encode* block of Figure 2. In the decoding operation, the size of each encoded patch is increased through bilinear upsampling to a target patch size of 128×128 pixels, before passing it through a final convolutional layer.

We use reflection padding to retain constant dimensions as the different convolutional layers are applied. We also experimented with zero padding and constant padding, but found reflection padding to result in better performance.

3.2. Loss Function and Partial Grid Search for Loss Coefficients

Training a neural network involves the minimization of an objective function that quantifies how well the transformation of the input matches the target. This objective function is typically referred to as the *loss* function (\mathcal{L}). As SR is an ill-posed problem (i.e., a one-to-many operation), multiple SR outputs can explain the same LR input. For scientifically useful applications of SR, the model output should capture the physical properties of the target, and cannot just be perceptually convincing. To better respect the physical properties of the target HR magnetograms, we construct a loss function that combines four terms, each of which aims to capture a different aspect of what makes a magnetogram physically plausible. Importantly, compared to recent advances in developing *physics informed neural networks* (e.g., Raissi et al. 2019), our model does not explicitly draw upon physical laws, but instead relies on terms that are common for SR and image processing and that also capture physical quantities of interest for magnetograms.

The loss function used in this work is

$$\mathcal{L} = \mathcal{L}_{l2} + w_{\text{grad}} \mathcal{L}_{\text{grad}} + w_{\text{hist}} \mathcal{L}_{\text{hist}} + w_{\text{ssim}} \mathcal{L}_{\text{ssim}}, \quad (1)$$

with the following terms:

1. \mathcal{L}_{l2} penalizes the mean squared error (MSE) between the superresolved output and the target, and captures pixel-based differences in the signed flux.

¹¹ <https://reproject.readthedocs.io/en/stable/>

¹² <https://github.com/ElementAI/HighRes-net>

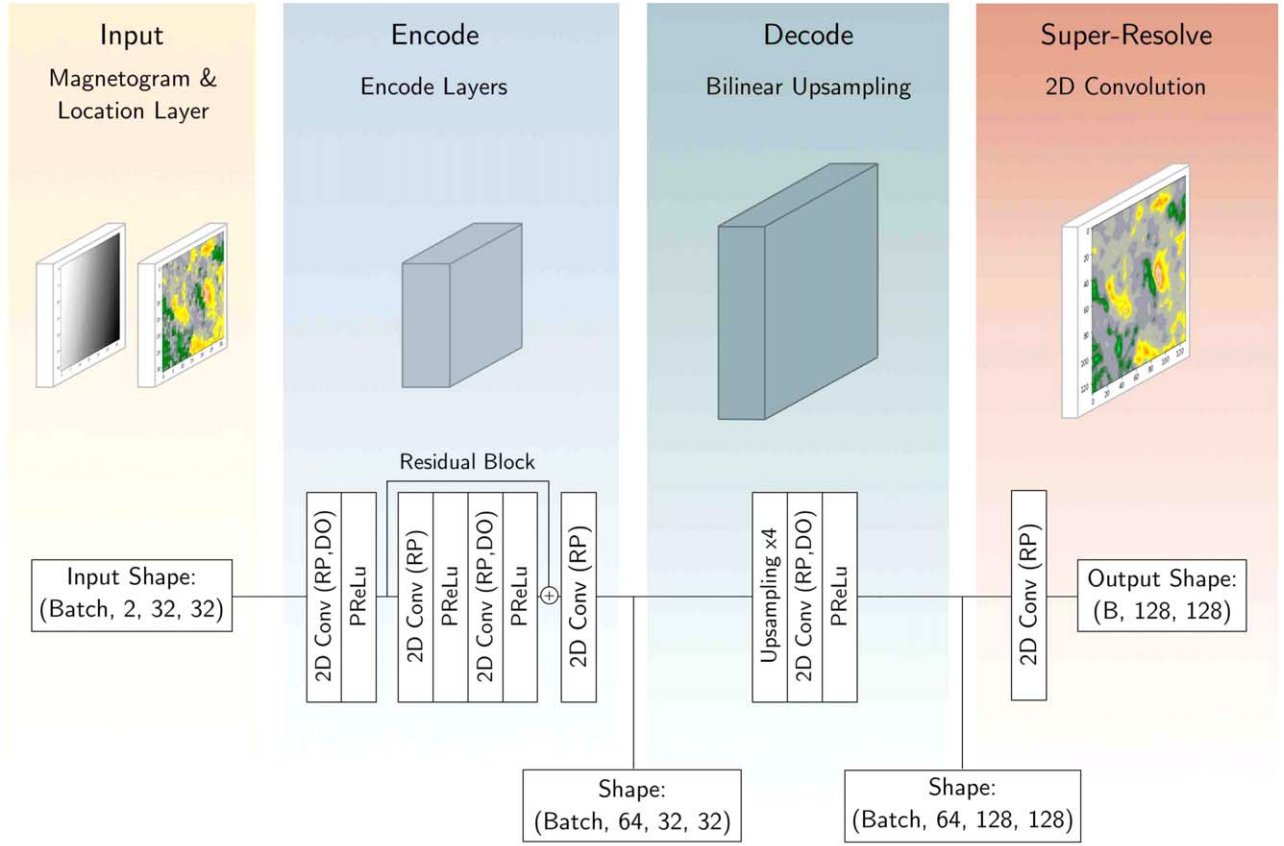


Figure 2. Diagram of the deep-learning model used in this work. The input data consists of a magnetogram and a location channel, each of size 32×32 pixels. The main operations of the model consist of an encoder and a decoder. Through bilinear upsampling by a factor of 4, the size of the image is increased from 32×32 to 128×128 pixels per patch. RP and DO denote whether a convolutional layer was trained with reflection padding or Monte Carlo dropout respectively.

- $\mathcal{L}_{\text{grad}}$ penalizes the mean squared difference between pixel gradients of the superresolved output and the target. The gradients are approximated using a Sobel operator (Pingle 1969). $\mathcal{L}_{\text{grad}}$ aims to capture the gradients present at the boundaries between positive and negative polarities.
- $\mathcal{L}_{\text{hist}}$ penalizes an approximation of the total variation distance between magnetic field distributions of the output and target magnetograms. For that, we calculate a differentiable pixel histogram using the method described in Wang et al. (2018). Formally, if the magnetic field of a magnetogram is divided into K bins, given a value of the magnetic field $B_{i,j}$ at pixel (i, j) , we assign a differentiable weight $\psi_k(B_{i,j})$ to bin $k = 1, \dots, K$, where

$$\psi_k(B_{i,j}) = \max\left\{0, 1 - \frac{1}{\gamma_k} |B_{i,j} - \mu_k|\right\},$$

where μ_k and γ_k are learnable parameters. Then, for each bin k , we compute a batch differentiable count of magnetic field values that falls within $[\mu_k - \gamma_k, \mu_k + \gamma_k]$ as

$$\sum_{n=1}^N \sum_{(i,j) \in p_n} \psi_k(B_{i,j}),$$

where N is the number of 128×128 patches p_n in the batch.

- $\mathcal{L}_{\text{ssim}}$ measures the structural similarity (SSIM; Wang et al. 2004) between regions surrounding each pixel, including similarities in contrast, unsigned flux, and variance. Formally, given a superresolved 128×128 patch $\hat{B}_n = \{\hat{B}_{i,j,n}\}$, its target $B_n = \{B_{i,j,n}\}$ and weights $\alpha_{i,j}$, the structural similarity is defined as

$$\text{SSIM}(B_n, \hat{B}_n) = \frac{(2\mu_B \mu_{\hat{B}} + C_1)(2\sigma_{B\hat{B}} + C_2)}{(\mu_B^2 + \mu_{\hat{B}}^2 + C_1)(\sigma_B^2 + \sigma_{\hat{B}}^2 + C_2)},$$

where

$$\mu_B = \sum_{(i,j)} w_{i,j} B_{i,j},$$

$$\sigma_B^2 = \sum_{(i,j)} w_{i,j} (B_{i,j} - \mu_B)^2,$$

and

$$\sigma_{B\hat{B}} = \sum_{(i,j)} w_{i,j} (B_{i,j} - \mu_B)(\hat{B}_{i,j} - \mu_{\hat{B}}).$$

$\mu_{\hat{B}}$ and $\sigma_{\hat{B}}$ are defined similarly as μ_B and σ_B . In practice, the weights are defined by an 11×11 circular Gaussian weighting scheme. And the constants C_1 and C_2 are set to 0.0001 and 0.009, respectively.

We initialized the values of the loss coefficients w to scale each term's contribution to the same order of magnitude as the

Table 2

Values of the Loss Coefficients for MDI and GONG Resulting from a Partial Grid Search

	MSE	Gradient	Histogram	SSIM
Optimal MDI run	1	5	1e-5	5e-4
Optimal GONG run	1	5	1e-6	5e-5

\mathcal{L}_2 term. We then refined the scaling factors by conducting a partial grid search to determine which w_{grad} , w_{hist} , and w_{ssim} minimize $\mathcal{L}_2 + w_{\text{grad}}\mathcal{L}_{\text{grad}}$, $\mathcal{L}_2 + w_{\text{hist}}\mathcal{L}_{\text{hist}}$, and $\mathcal{L}_2 + w_{\text{ssim}}\mathcal{L}_{\text{ssim}}$, respectively (see Table 2). We subsequently used the values of the weights in Table 2 to minimize the loss function included as Equation (1). These coefficients vary by instrument because of differences in the properties of GONG and MDI, e.g., resolution, noise, and saturation levels.

3.3. Training Hyperparameters

Table 3 shows the values of the hyperparameters used to train the models. All magnetic fields are normalized by 3500 to avoid overflow issues. We use an Adam optimizer (Kingma & Ba 2014), with a constant learning rate (10^{-4}) without an annealing schedule.

4. Results

4.1. Assessing Physical Properties of Superresolved Magnetograms

As mentioned previously, compared to natural images, which often consist of three color channels with integer pixel values, LOS estimates of the solar magnetic field can be positive or negative and span multiple orders of magnitude. Additionally, given that the Sun is a three-dimensional object projected onto a two-dimensional image, LOS measurements show projection effects that are location dependent. More specifically, the solar limb shows significantly larger projection effects than areas close to the center of the Sun. This highlights the need to consider pixel locations when evaluating the performance of SR approaches (as discussed in more detail later).

To measure the performance of any SR or cross-calibration operation of solar magnetograms, it is essential to approach them as scientific measurements rather than standard images. We propose to use the following quantities to compare the performances of SR/cross-calibration approaches. Note that these quantities are post-mortem measurements that, we believe, should be reported for any SR/cross-calibration methods of solar magnetograms and astronomical data in general.

We denote $\hat{B}_{i,j,n}$ as the superresolved magnetic field at pixel (i, j) and patch n ; and $B_{i,j,n}$ as the ground-truth target magnetic field value for the corresponding patch and at the same location. Each patch n , unless specified otherwise, refers to an area of 128×128 pixels, corresponding to $1/1024$ of a full-disk HMI magnetogram. We find that patches of 128×128 pixels are large enough to encompass a whole active region, while being small enough to allow fast matrix computations and obtain an approximate flux balance (Mackay et al. 2011).

1. *Correlations:* We follow Liu et al. (2012) and measure how superresolved magnetograms are cross-calibrated to their HR counterpart by measuring the Pearson

Table 3

Hyperparameters Used to Train the Best Model

Hyperparameter	Value
Number of epochs	20
Learning rate	10^{-4}
Batch size/GPU	64
Number of GPUs	8

correlation coefficient between $\hat{B} = \{\hat{B}_{i,j,n}\}$ and $B = \{B_{i,j,n}\}$ across all pixels (i, j) and patches n :

$$\rho = \frac{\sum_{i,j,n} (B_{i,j,n} - \bar{B})(\hat{B}_{i,j,n} - \bar{\hat{B}})}{\sqrt{\sum_{i,j,n} (B_{i,j,n} - \bar{B})^2 \sum_{i,j,n} (\hat{B}_{i,j,n} - \bar{\hat{B}})^2}}, \quad (2)$$

where \bar{B} and $\bar{\hat{B}}$ are the average ground-truth and superresolved magnetic field across all pixels and patches (ρ takes value between 0 and 1). The larger ρ the better is the cross-calibration of the superresolved magnetograms to their HR counterpart.

2. *Signed fluxes:* Magnetic fields are divergence-free, i.e., the integration of the radial magnetic field over the solar surface sums to zero. To evaluate how an SR technique conserves the signed flux, we calculate the signed flux of a pixel by converting the LOS field into the radial field and correcting for area foreshortening:

$$E_{i,j,n}^{\text{flux}} = \hat{\phi}_{i,j,n} - \phi_{i,j,n}. \quad (3)$$

3. *Extreme values:* Regions with extreme magnetic field values occupy areas that are smaller than the area covered by a pixel of a magnetogram. This is particularly true for lower-resolution instruments (e.g., MDI, GONG). We expect that the filling factor (i.e., the ratio of the area occupied by the magnetic field to the total area) is larger at HR than at LR. Therefore, particularly for the study of sunspots and active regions, it is of interest to assess the ability of an SR technique to generate extreme values that occupy an area smaller than an LR pixel. Moreover, extreme values of the magnetic field have low frequency. Therefore, they may be described less confidently by an SR technique that learns its predictions from data with a limited number of occurrences of extreme values. To measure the ability of an SR technique to reproduce the tail of the magnetic field distribution, we compute the absolute difference between the minimum/maximum magnetic field over each 128×128 patch n :

$$E_n^{\text{max}} = |\max B_n - \max \hat{B}_n| \quad (4)$$

and

$$E_n^{\text{min}} = |\min B_n - \min \hat{B}_n|. \quad (5)$$

4. *Gradients:* We expect large magnetic field values to occupy a smaller number of pixels in HR magnetograms than in lower-resolution magnetograms. Pixel-level gradients of magnetic field values can quantify variations of the magnetic field within an LR pixel. This also helps to evaluate how an SR technique captures polarity inversion and defines boundaries between positive and negative regions. We compute $E_{i,j,n}^g$ as the (i, j) pixel of the image gradient of the difference $\hat{B}_n - B_n$ between the

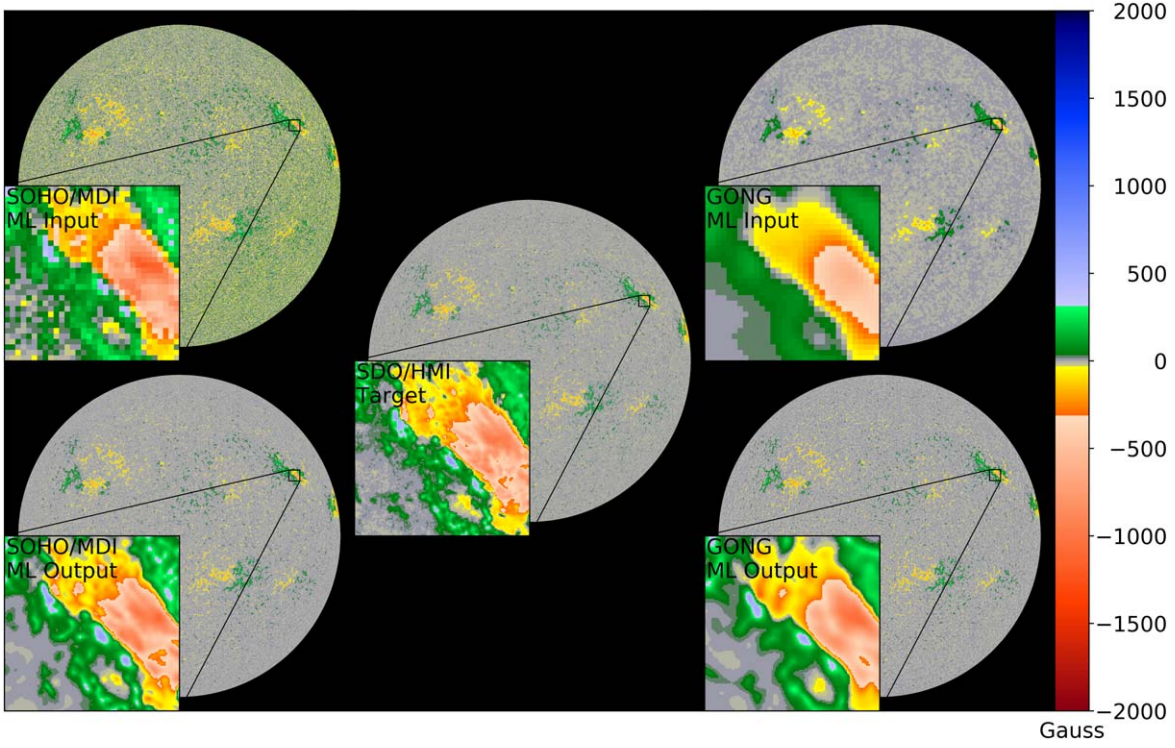


Figure 3. Example of the resolution differences between an input magnetogram (MDI/GONG), the deep-learning output, and the target (HMI). The insets show a 128×128 patch with clear resolution differences between the model inputs (MDI and GONG) and the model target (HMI). All panels show nearly simultaneous measurements made by GONG, MDI, and HMI on 2011 March 14.

predicted and true magnetogram of patch n :

$$\begin{aligned}
 E_{i,j,n}^g &= g(\hat{B}_n - B_n)_{i,j}, \quad \text{where} \\
 g(I)_{i,j} &= \sqrt{g_x(I)_{i,j}^2 + g_y(I)_{i,j}^2}, \quad \text{with} \\
 g_x(I) &= G_x * I, \quad \text{and} \\
 g_y(I) &= G_y * I.
 \end{aligned} \tag{6}$$

Here, g is the (i, j) pixel of the output image of the Sobel operator g on image I . G_x and G_y are 3×3 kernels¹³ that convolve an image to produce the smoothed finite difference on the x and y image dimensions, respectively.

To measure the performance of SR, we compute the signed flux and extreme values at small spatial scales using patches of size 4×4 , 8×8 , 16×16 , and 32×32 pixels. In addition, we also calculate the Pearson correlation as a function of magnetic field strength and location on the surface of the Sun. This allows us to understand how the performance of any SR technique applied to the solar magnetic field depends on the spatial scale and strength of the magnetic field.

4.2. Baseline Comparisons

To benchmark our deep-learning approach, we compare it against a bicubic upsampling baseline. Bicubic upsampling interpolates only the information contained in the LR image, and does not add new information to the higher-resolution counterpart.

¹³ $G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$ and $G_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$.

We follow the method presented in Liu et al. (2012) and apply a cross-calibration factor to MDI and GONG. We perform a linear regression of LR magnetic fields (MDI or GONG) against the HR magnetic field (HMI) of the form $\text{MDI/GONG} = a + b \times \text{HMI}$. We find that $b = 1.3$ ¹⁴ for MDI and $b = 0.7$ for GONG. We construct our baseline by bicubic upsampling, and then, scaling of MDI by $1/1.3$ and GONG by $1/0.7$.

In addition to our baseline, we also compare our work to results achieved with the same neural network but employing a loss function that is only based on the MSE between the model output and target. This allows us to highlight the need for including physically motivated terms in the loss function when handling scientific data.

4.3. Ablation Study—Optimization Penalty Terms

Our first trained models only contained the MSE loss (see Section 3.2). However, it became apparent that this loss was not sufficiently nuanced to capture important properties of the HR magnetograms. In this section, we show the results of the gradual addition of the loss terms described (see Section 3.2) roughly in the order in which they were added. In total, we added three additional terms that addressed the need for better polarity inversion lines ($\mathcal{L}_{\text{grad}}$, Grad), well-balanced positive and negative polarities ($\mathcal{L}_{\text{hist}}$, Hist), and more concentrated and detailed magnetic field features ($\mathcal{L}_{\text{SSIM}}$, SSIM). These were the only additional loss terms that we experimented with and they resulted in superresolved magnetograms that better match the

¹⁴ Liu et al. (2012) find a value of $b = 1.4$, but their regression uses uniform weight across all inputs. Instead, we bin the LR magnetic field and weight each input by the fraction of points that fall into the bin the input belongs to. A weighted regression balances the impact of low and high magnetic fields.

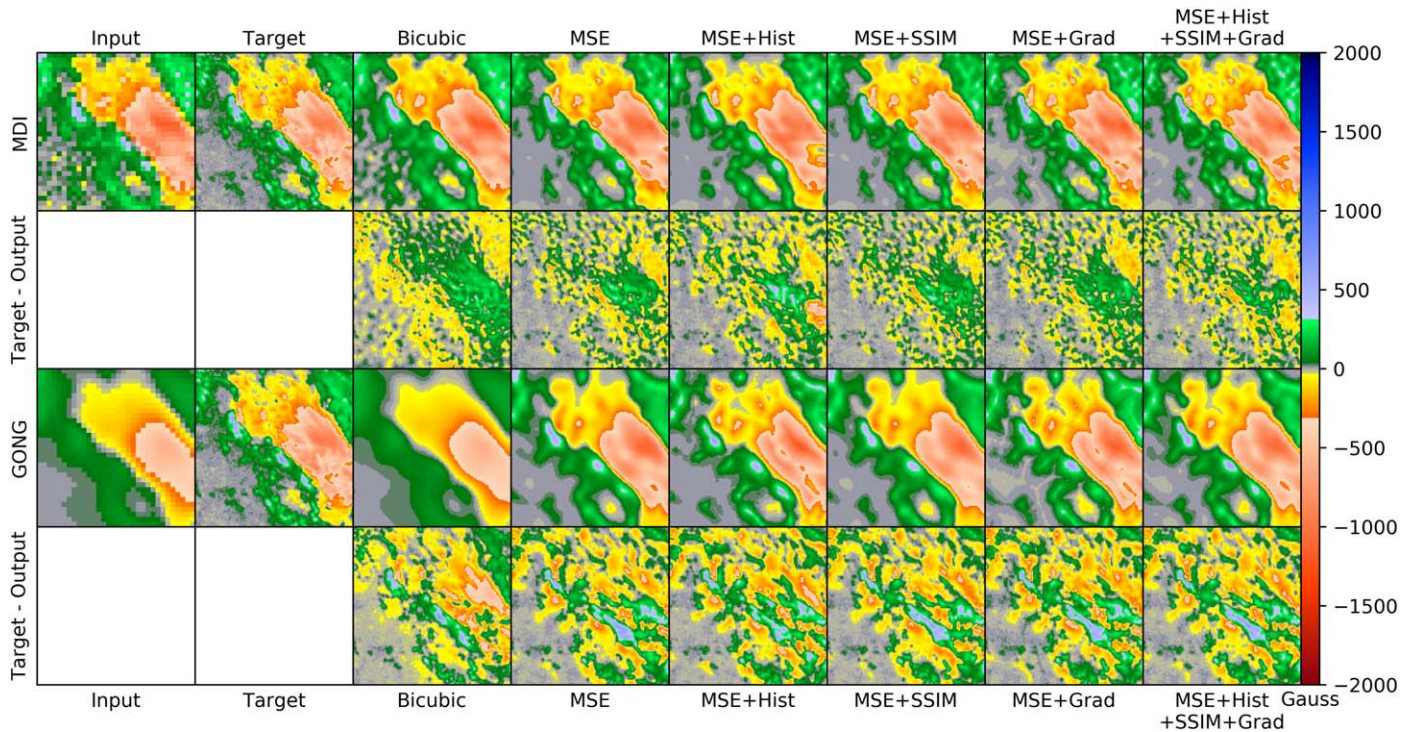


Figure 4. Comparison of the same area of a magnetogram on 2011 March 14 for the input, target, and deep-learning output with different loss functions. The top two rows show MDI magnetograms as input, the bottom two rows show a comparison for the conversion of GONG magnetograms. We chose this particular magnetogram patch as an example for its interesting structure and moderate to high magnetic field strength.

statistical properties of HMI. However, we cannot rule out the possibility that there are other even better loss terms, or combinations of them, that could further improve the results.

Figure 3 shows full-disk images of our input data (top row) and the best results of our deep-learning SR network (bottom row). The insets show one of the 1024 patches used for splitting the Sun during training. These results were achieved with a loss function that features a combination of differentiable physically motivated terms including the MSE, magnetic field gradients, pixel histograms, and self-similarity penalties, as discussed earlier. The superresolved full-disk magnetograms of MDI and GONG have noise levels, texture, and relative magnetic field intensity akin to those of the HMI target. Zooming in closer, the insets show higher-resolution structures for the model’s outputs, which better match those of the HMI target. The improvement is especially significant for GONG, with a striking difference in small-scale structures between the input and output patches.

Figure 4 compares superresolved magnetograms obtained with different loss functions to the input (MDI/GONG), the target (HMI), and our bicubic upsampling baseline. The first row (third row) in Figure 4 shows the same patch of the Sun with MDI (GONG) as the input. The second and last rows show the calculated difference between the upsampled magnetograms and the target.

Starting with our baseline, the bicubic upsampled MDI magnetogram still shows the salt-and-pepper-like noise structure that is present in the MDI input in the lower left corner of the magnetogram patch. This is because simple upsampling techniques extrapolate the magnetic field, including its noise, to the higher-resolution image. Moreover, bicubic upsampling cannot leverage the information present in the whole data set of magnetograms. Bicubic upsampling of GONG increases the

sharpness of edges around active regions, but the large patch-like features do not increase in detail.

Using our deep-learning model with a simple MSE loss removes the noise floor of the MDI input image. In addition, we start to recover small-scale features in and around active regions. Adding optimization penalty terms to the MSE loss modifies details in the HR reconstructions. It also visibly reduces the characteristic size of the structures in the difference images (Figure 4, second row). We see this as evidence that the additional loss terms allow the CNN to better capture the structure of the target magnetic field. However, a purely visual inspection of the images is not enough to find significant differences or distinguish which loss function is best at recovering HR features.

Figure 5 is an ablation study that compares the effect of each component included in the loss function on the reconstruction of the magnetic field. We compare performances by evaluating the post-mortem measurements introduced earlier and calculate (a) differences in extreme magnetic field values, (b) the Pearson correlation coefficient, (c) differences in image gradients, and (d) differences in the signed flux of the target and deep-learning output magnetograms. All metrics are calculated on a pixel-to-pixel basis across our test set, which contains approximately 1 million patches for MDI and 8 million patches for GONG. Figure 5 shows the results obtained for MDI input magnetograms.

As mentioned above, a simple MSE loss succeeds at creating visually pleasing magnetogram outputs that show a higher level of detail than the input magnetograms (see Figure 4). However, an objective function based exclusively on MSE is unable to reconstruct extreme values of the magnetic field (i.e., the strongest positive and negative magnetic fields in a patch) properly as shown in Figure 5(2-a)). Looking at how well

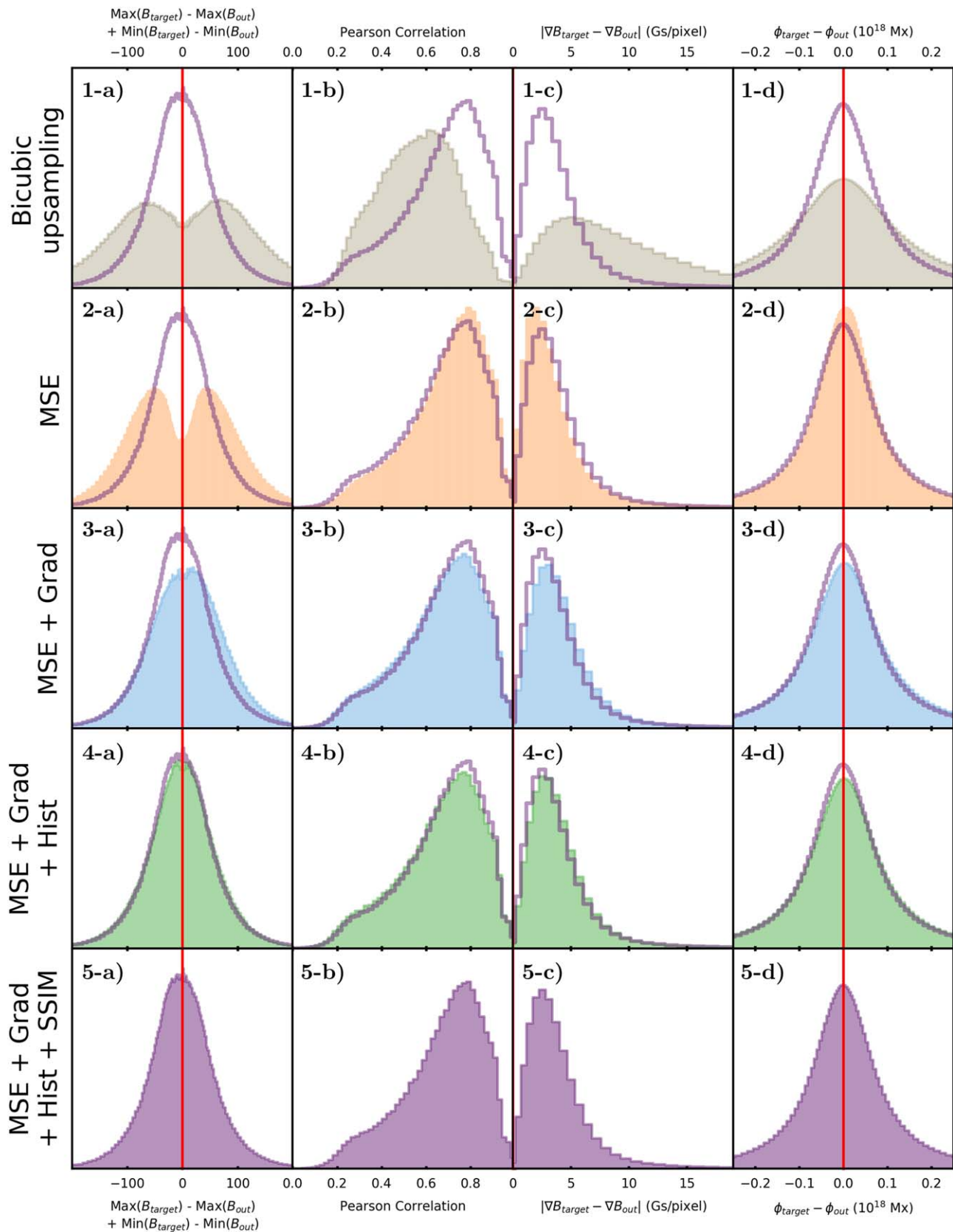


Figure 5. Quantitative comparison of the performance of different loss functions trained on MDI magnetograms. We use bicubic upsampling as a comparative baseline (row 1). All loss functions are based on the MSE term, plus up to three additional penalty terms. The shaded histograms correspond to calculations across the test set for patches within 90% of the solar disk radius. A red vertical line indicates the ideal value for the corresponding performance metric. The two leftmost columns (a) and (b) show calculations performed across patches of 32×32 pixels. The two rightmost columns (c) and (d) show calculations performed per pixel in each patch of the test set. Columns left to right show (a) the added difference between the target and output maxima and minima per patch, (b) Pearson correlation per patch, (c) the magnitude of the target and output gradient difference per pixel, and (d) the difference in target and output magnetic flux per pixel. To compare MSE+Grad+Hist+SSIM with the rest, the bottom row is superimposed (nonshaded histogram) on all other rows.

extreme values are reconstructed, we observe double peaks centered around ± 100 Gauss in the bicubic baseline (Figure 5(1-a)), and when using an MSE loss (Figure 5(2-a)). With MSE alone, the neural network consistently underestimates the magnitude of extreme values, leading to a peak centered around $+100$ Gauss for the maximum magnetic field values, and a second peak centered around -100 Gauss for the minimum magnetic field values when comparing the target and the deep-learning output. Additionally, the MSE loss produces a clear asymmetry in the signed flux (Figure 5(2-d)). This is highly problematic considering that the solar magnetic field must be divergence-free, and thus in flux balance.

Including a gradient penalty term in the loss function (indicated in the third row of Figure 5 as MSE + Grad) removes the double peaks and centers the distribution around zero (red line in Figure 5(3-a)). Taking image gradients into account is a measure often used in computer vision to improve edge detection and texture matching (Forsyth & Ponce 2002). For the application to magnetograms, edge detection aids in defining boundaries around active regions, and texture matching helps to recover detailed features in the HR image. Despite these improvements, maximum fields are still slightly underestimated, as indicated by the fact that the distribution of extreme values is asymmetrically skewed toward positive values for an MSE + gradient loss function (Figure 5(3-a)).

On average, the sum of the magnetic field values on the surface of the Sun is expected to be close to zero. Deviations from zero only occur when the leading part of an active region comes into view of the instrument, and the following cancellation of the magnetic field cannot be viewed yet. Biases in reconstructing positive or negative fields in the super-resolved magnetic field would violate Gauss' law (which states that the magnetic field must be divergence-free). The histogram penalty (MSE + Grad + Hist in the fourth row of Figure 5) manages to mostly correct the skewed distribution of extreme values (Figure 5(4-a)) while also slightly shifting the discrepancies in image gradients (Figure 5(4-c)) closer to zero.

We further improve model performance by adding a similarity penalty term (SSIM, see Section 3.2) that forces the model to learn spatial structures of the solar magnetic field (MSE + Grad + Hist + SSIM in the fifth row of Figure 5). This combination of terms also produces the solution with the best balance between positive and negative fluxes (Figure 5(4-d)).

This ablation study shows that the structure of the loss function we optimize for generates trade-offs for the physical properties of the superresolved magnetograms. While MSE alone achieves better performances in terms of Pearson correlations, adding a gradient, histogram, and SSIM penalty terms significantly improves how the model captures extreme values of the solar magnetic field, as well as achieving the flux balance critical to ensuring that the recovered magnetic field is divergence-free.

4.4. Benefits of Data Homogenization

In the following, we demonstrate the value added to using superresolved magnetograms over their LR counterparts. Specifically, we investigate small- and large-scale structures, homogenization properties, and temporal patterns of the superresolved magnetic fields.

In Figure 6, the first (third) row shows a pixel-to-pixel correlation plot between target magnetograms and SR output

for the entire test set for MDI (GONG). The test set contains ≈ 25 (≈ 125) million pixels for MDI (GONG). The orange lines highlight regression lines between the output and target. To put these results into perspective, Figure 6 also shows a comparison of the correlation between the bicubic upsampling baseline and the target magnetograms (purple graph on the right). Correlation plots aligning with the 45° diagonal or small residuals indicate good cross-calibration. Our deep-learning approach centers the correlation plots more on the 45° diagonal, thereby improving the cross-calibration between MDI (GONG) and HMI. This is clearly visible for GONG (Figure 6, third row) and less strongly observable for MDI (Figure 6, first row).

Figure 6 suggests great performance of both the bicubic baseline (purple) and our deep-learning approach (orange) when only looking at the correlation plots as both comparisons are strongly centered around $x = y$. The second and fourth rows in Figure 6 show further quantitative visualizations of both the cross-calibration and SR approaches, showing that deep learning is more effective at performing SR. We measure this improvement by investigating the relative average deviation of the output and target across a 4×4 pixel area compared to the corresponding LR pixel. In the case of MDI, the deep-learning algorithm is able to fix the well-known saturation error at strong magnetic fields. In the case of GONG, the deep-learning algorithm is able to overcome the large disagreement between GONG and HMI magnetograms present around 1000 G.

These results highlight one of the main challenges of finding suitable quantities to capture SR. When looking simply at averages, it is easy to become overly confident and misrepresent the quality of results. Our work encourages benchmarking SR techniques with a quantitative assessment that directly measures the reconstruction of small-scale structures of the magnetic field.

4.5. Comparison of Large-scale Structures

Table 4 replicates the quantitative assessment in Tables 1 and 2 of Liu et al. (2012) and compares the Pearson correlation coefficient between superresolved MDI and GONG magnetograms across different radial regions of the Sun and different values of the magnetic field. For both MDI and GONG, the Pearson coefficient is computed on our test set of magnetograms from 2011 March. Our results show that our deep-learning approach generates magnetograms that contain information present in HMI magnetograms, but not in their LR counterparts. Across all radial regions and field values, the Pearson correlation coefficient between superresolved and HMI magnetograms increases by 5%–7% relative to the correlation between lower-resolution and HMI magnetograms.

4.6. Comparison of Small-scale Structures

In Table 5, we compare statistics of the magnetic field between HMI and superresolved MDI/GONG over kernels of various pixel sizes. We benchmark our results against our baseline approach (bicubic upsampling with linear rescaling). Our results show that the difference in gradient between HMI and deep-learning output over small kernels (2–4 pixels) is 30% (for MDI) and 4% (for GONG) smaller than between HMI and the baseline outputs. Similar improvements are observed for extreme values of the magnetic field within kernels of different sizes. This confirms that our deep-learning SR

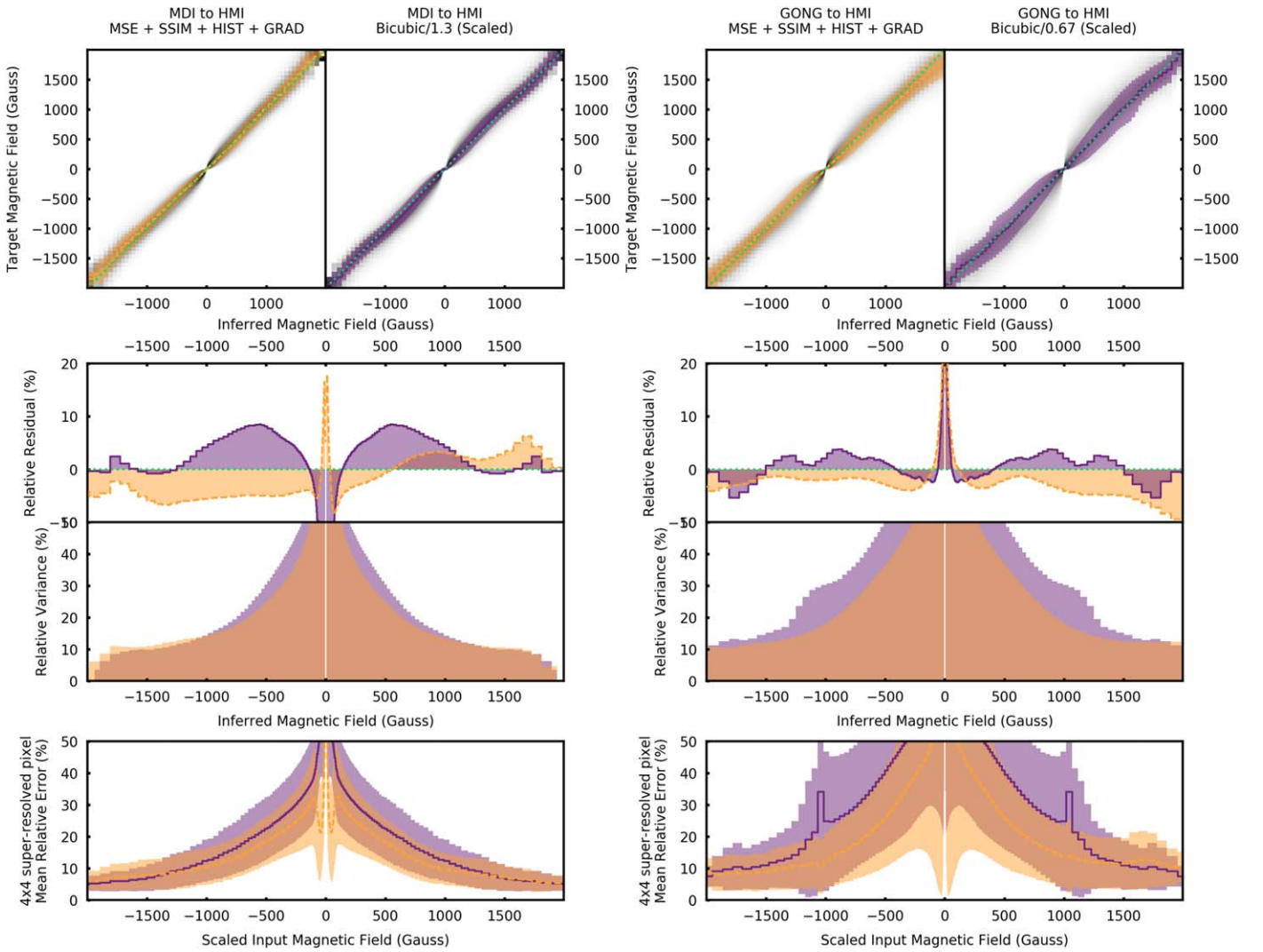


Figure 6. 2D Histogram of pixel-to-pixel comparison for test patches of the MDI \rightarrow HMI (top two rows) and GONG \rightarrow HMI (bottom two rows) transformation. The following caption equally applies to the panels belonging to each instrument. In all panels, orange indicates the deep-learning output, while purple indicates the bicubic upsampling baseline. The bold lines denote the median of the quantity and the shaded area is the 25th–75th percentile range. Top row: comparison between the inferred (x -axis) and target magnetic field (y -axis) for all individual pixels in our test set. The $y = x$ line is denoted with a green dotted line. Note that the bold median lines are difficult to see because a simple correlation plot is not ideal for evaluating SR performance. Bottom row: mean relative error in each superresolved 4×4 pixel patch (y -axis) as a function of the magnetic field of the corresponding LR input pixel (x -axis). We propose this as a far superior quantity to visually evaluate the performance of SR algorithms. The closer this quantity is to zero, the better.

Table 4

Comparison of MDI/Superresolved SR-MDI and GONG/Superresolved SR-GONG with HMI

Area	MDI	SR-MDI	GONG	SR-GONG
$0 \leq r \leq 1$	0.89	0.93	0.77	0.88
$0 \leq r \leq 2/3$	0.91	0.94	0.82	0.90
$0 \leq r \leq 1/3$	0.90	0.94	0.80	0.88
$1/3 \leq r \leq 1/2$	0.91	0.94	0.82	0.90
$1/2 \leq r \leq 3/4$	0.91	0.94	0.83	0.90
$10 \text{ G} \leq \text{field}$	0.92	0.94	0.82	0.90
$600 \text{ G} \leq \text{field}$	0.97	0.98	0.94	0.97
$0 \leq \text{field} \leq 600 \text{ G}$	0.82	0.88	0.68	0.81

Note. The radius of the Sun r is normalized to 1 at the limb. The table reports the Pearson correlation coefficient between MDI and HMI; and GONG and HMI. Higher values are better. To compare with HMI, MDI, and GONG images are upsampled and multiplied by the cross-calibration factor as described in Section 4.2. The Pearson correlation coefficient is computed as in Equation (2).

captures details that are averaged out at lower resolution. Remarkably, improvements in small-scale patterns extend to structures of size larger than the $4\times$ upscaling factor.

4.7. Quality Assessment of SR Using SHARPs

Finally, in this section, we show the value added of superresolved magnetograms to study time series of unsigned fluxes and gradients within SHARPs (Bobra et al. 2014).¹⁵ The SHARPs data series allows us to systematically track solar active regions at a 12 minute cadence over the lifetime of each region, and is among the most widely used products for connecting the evolution of individual magnetic regions and space weather events (see Asensio Ramos et al. 2023, and the references therein). Each region tracked through SHARPs has an associated HMI Active Region Patch (HARP) number (HARPNUM), and a number of associated calculated quantities

¹⁵ <http://jsoc.stanford.edu/doc/data/hmi/sharp/sharp.htm>

Table 5
Quantitative Comparison between MDI/Superresolved SR-MDI and HMI, and between GONG/Superresolved SR-GONG and HMI

Kernel Size (Pixels)	Gradient				Extreme Values			
	(G)				(G)			
	MDI	SR-MDI	GONG	SR-GONG	MDI	SR-MDI	GONG	SR-GONG
2	4.91	3.38	4.06	3.90	8.07	7.04	8.52	7.92
4	4.92	3.38	4.07	3.91	10.70	10.27	13.18	11.45
8	4.93	3.39	4.08	3.91	16.33	16.51	22.37	18.72
16	4.99	3.40	4.09	3.93	30.64	30.29	42.0	35.39
32	5.06	3.43	4.13	3.96	66.69	61.82	87.74	74.02

Note. The table reports the average over patches of size 4×4 , 8×8 , 16×16 , and 32×32 of the extreme values and gradient metrics reported earlier. To compare with HMI, the MDI and GONG images are upsampled and multiplied by the cross-calibration factor as described in Section 4.2.

used for space weather forecasts, with the average unsigned flux and average unsigned gradient being the only two that can be calculated using LOS magnetograms. The main objective of this exercise is to evaluate whether superresolving GONG and MDI magnetograms is adding value to the resulting magnetograms, or whether pure calibration, i.e., adjusting for instrumental differences without upscaling, either using our baseline approach or via deep learning, results in better end products.

Figure 7 focuses our analysis on HARP region 407 (NOAA AR 11169). This region was chosen out of the 36 HARPNUM observed during 2011 March (our test set) because it is large, long-lived, and has relatively good time coverage by both the MDI and GONG instruments. The top panels (Figure 7(a)) show five snapshots of the evolution of HARP region 407 as observed by our target instrument (HMI). Figure 7(b) shows the unsigned magnetic flux for the HMI target (red triangles), compared to ML SR or bicubic upscaling of MDI (green/beige dots) or of GONG (purple/light blue dots). The colored vertical lines indicate the timestamp of the snapshots shown in Figure 7(a). Figure 7(c) shows the ratio between the ML or bicubic upscaling outputs and the HMI targets.

We find that the magnitude of the unsigned flux, and its evolution as the region moves through the instrument’s field of view, are well reproduced in the ML outputs of both MDI and GONG. Looking at the ratio comparison with HMI, we find the calibration to be centered around the target value of 1.0, with two main characteristics. The first one is the underestimation of the unsigned flux during the emerging phase of the HARP region, which stabilizes after the region is fully developed (Figure 7(c)). This behavior arises from the fact that the ML algorithm denoises the magnetogram, thereby reducing the relative amount of unsigned flux in the ML output compared to the HMI target. As the region grows (and most unsigned magnetic flux is in the region itself), this stops being an issue. One of the main differences between the ML output (green circles and purple squares) and the bicubic upsampling baselines (tan circles and blue squares) is the ability of the ML algorithm to address known systematic issues introduced by projection effects due to area foreshortening close to the solar limb. While the bicubic baselines systematically underestimate flux as the HARP region rotates out of view, the ML output is generally stable throughout the lifetime of the HARP.

The second observable characteristic in Figure 7(c) is a clear 24 hr modulation that is a known systematic issue in HMI measurements due to HMI’s fast orbital speed (see Figure 9). Neither GONG, being ground based, nor MDI, having a stable orbit at L1, suffer from this problem. Interestingly, the ML output retains the stability of the input instruments. This likely

arises because time is not included as an input parameter, and therefore no direct information is provided for the model to learn temporal dependencies. While we have not explored this further, this hints at the possibility of using GONG and MDI to remove this outstanding systematic issue from HMI in the future.

Figure 7(d) shows a detailed comparison of the magnetic field and gradients calculated on the HMI target, and the MDI and GONG outputs using bicubic upscaling and ML-based SR at the same point in time. Each panel also contains the numerical measurement of the average unsigned magnetic flux and average unsigned gradient. The quantitative and qualitative superiority of the superresolved output over the baseline bicubic upscaling is evident.

Figure 8 further illustrates the superiority of the ML-based SR output (ML-4X; light blue squares) over both the baseline (purple dots) and an ML-based calibration approach (ML-1X; green squares). The latter was obtained by training our ML model on HMI magnetograms that have been downsampled to the resolution of MDI/GONG. In this figure, we also show results from all SHARPS regions and time stamps available during our test period.

Figures 8(a)–(d) show the time evolution of the unsigned flux and the average unsigned gradient as estimated by the different algorithms applied to HARP region 407 as it crosses the solar disk. Figures 8(e) and (f) show scatter plots of the HMI target against the different outputs for all 36 regions (and all their time stamps during 2011 March; a total of 2261 SHARPs compared). The closest match to the target HMI measurements (for both the average unsigned flux and average unsigned gradient) is always the SR output (ML-4X). While all algorithms underestimate the unsigned average gradient (Figures 8(c), (d), (g), and (h)), the improvement brought by SR is evident. This improvement is especially remarkable for GONG, which is currently the world’s main space weather operations magnetograph.

In the case of average unsigned flux, we also find the best match to be the SR outputs. Interestingly, bicubic upscaling overestimates the unsigned flux by enhancing the contribution of noise (Figures 8(a) and (b)). In comparison, our SR approach not only properly calibrates the inputs to the target instrument, but also denoises the magnetograms, ensuring an almost perfect match with the target HMI magnetograms. Table 6 in Appendix A.3 shows that the benefits of SR extend to all SHARPs. Taken all together, we see this as evidence that ML-driven SR is superior to simple upscaling by all relevant metrics and performance quantities. Considering that this is one of the first implementations of ML-based SR to solar

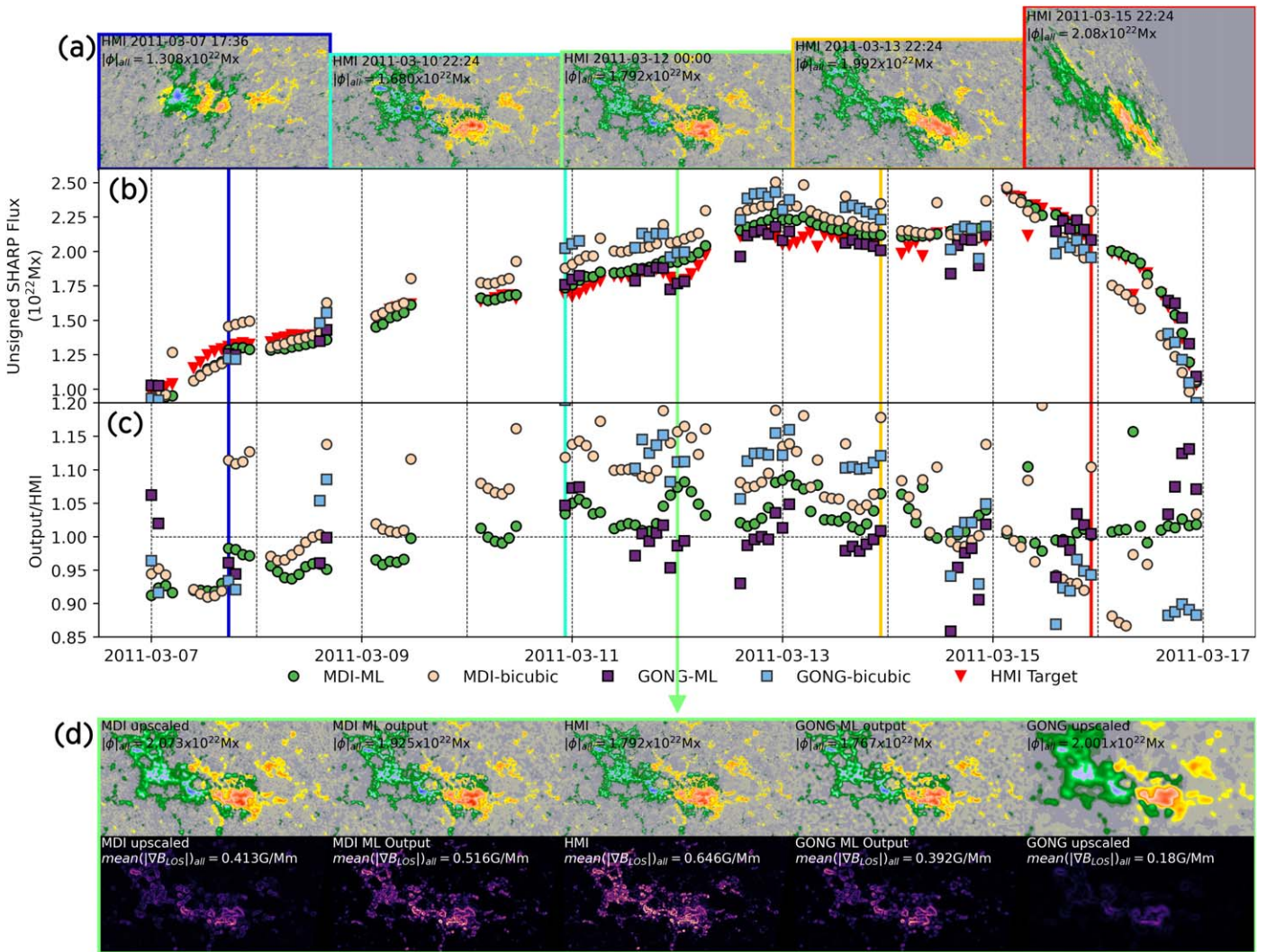


Figure 7. Time evolution of HARP 407. (a) Row showing the HMI magnetogram patch associated with HARP 407 at different stages of its lifetime. The different time stamps are indicated as vertical colored lines in the timelines of panels (b) and (c). (b) Total unsigned flux within the whole patch for the HMI target (red triangles) and the SR output of MDI (green dots) or GONG (purple squares). The bicubic baseline is also shown for MDI (beige dots) and GONG (light blue squares). The magnitude and evolution of the unsigned flux of the HMI target region is well reproduced in all cases. (c) Ratio of the unsigned flux between the SR output (or bicubic baseline) and HMI. (d) Magnetograms (top row) and gradients (bottom row) obtained for the same time step and using different upscaling techniques. The center panel is the HMI target.

magnetograms, it is reasonable to assume that these algorithms will only perform better in the future.

5. Conclusions

This work shows that deep-learning-based SR successfully upsamples and homogenizes solar magnetic field images, while adding detail that quantitatively improves the measurement of space-weather-relevant quantities in these images. Referring back to the goals set out in the introduction, our (1) demonstrates that a deep-learning approach can leverage the information present in astronomical images to recover detail in LR images while maintaining their scientific accuracy; (2) shows how superresolving a scientific image via deep learning homogenizes instrument properties and adds value compared to simple calibration at the same resolution, (3) establishes a set of quantitative performance measurements that can be used to benchmark the performance of different SR algorithms for astronomical images, as well as to benchmark the performance of future applications of SR to the physical sciences.

More specifically, we demonstrate the suitability of our approach by upsampling and cross-calibrating MDI (GONG) magnetograms to the characteristics of HMI. We show that a careful design of the loss function of the neural network improves the quality of the SR application, a conclusion that may be applicable to any deep-learning SR application in the physical sciences. Specifically, in the loss function, we include four penalty terms (i.e., the MSE, and differences in magnetic field value distributions, gradients, and self-similarity between the output and target) that constrain the deep-learning output to better match the ground truth. We further propose a set of quantities to evaluate the quality of (1) cross-calibration, and (2) SR of magnetograms that can also be applied across disciplines.

An important contribution of this work is to offer a benchmark of measurements and methods for performance comparison of future ML-based approaches that cross-calibrate/superresolve solar magnetic field images in particular, and scientific images more generally. We compare moments of the magnetic field at various spatial scales to capture how our

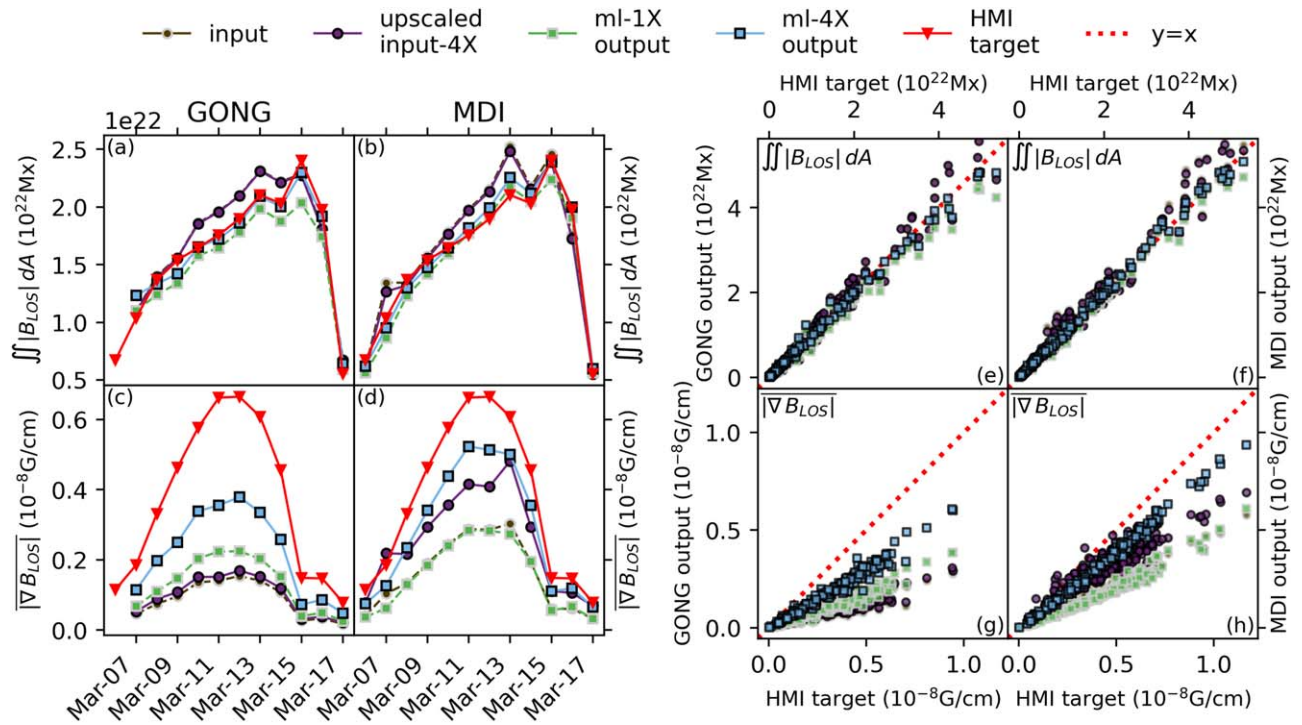


Figure 8. Benefits of SR, compared to bicubic upscaling or ML-based calibration without SR. Time evolution of the (a) and (b) unsigned flux and (c) and (d) average gradient for HARP 407 for superresolved (a)/(c) GONG and (b)/(d) MDI magnetograms. The HMI target is shown with red triangles. In all cases, the closest match is the superresolved ML-4X output. Scatter plots of the (e) and (f) average unsigned flux and (g) and (h) average gradient for all HARPs and all corresponding time stamps during 2011 March. The x-axis always corresponds to the HMI target and the y-axis to the ML output of (e)/(g) GONG and (f)/(h) MDI magnetograms. The closest match to the $y = x$ line (dotted red) is always the superresolved ML-4X output.

technique superresolves MDI and GONG magnetograms. Establishing benchmarks is necessary for the development and progress of deep-learning approaches for solar magnetic field research. Furthermore, in order to improve transparency and cross-comparability, we encourage reporting the same metrics for the same test month (2011 March) for future applications of SR and cross-calibration of solar magnetic field images.

Future work will explore including temporal information in the deep-learning architecture through multiframe SR (Deudon et al. 2020). Moreover, it is essential to investigate how the preprocessing of the solar magnetograms, including feature alignment and reprojection in a common coordinates system affect the performances of a deep-learning approach. Lastly, our current deep-learning approach does not allow us to quantify how confident the model is about its predictions, particularly for periods where there is no ground truth. It is a promising avenue for future research to implement a probabilistic ML approach that would estimate the uncertainty of its superresolved predictions, all the more as SR is an ill-posed problem with many superresolved images being consistent with the same LR input.

5.1. Data and Code Availability

The SDO/HMI, and SOHO/MDI data set used during the current study are available from the Joint Science Operations Center at <http://jsoc.stanford.edu/>. GONG magnetograms are available from the NSO website at <https://gong2.nso.edu/archive/patch.pl>. To prepare the original data for our deep-learning pipeline, we follow the preprocessing steps outlined in the Methodology section (Section 3). The code used in this work is available at <https://gitlab.com/frontierdevelopmentlab/>

[living-with-our-star/super-resolution-maps-of-solar-magnetic-field](https://github.com/muñoz-jaramillo/super-resolution-maps-of-solar-magnetic-field). All questions regarding the code should be directed to the corresponding author. A simplified version of the repository is available via Muñoz-Jaramillo et al. (2021a), containing (1) the necessarily functions to process the original data into a machine-learning-ready format, and (2) inference-based magnetogram converters using our best models. An example data set of our upscaled and calibrated magnetograms is available via Muñoz-Jaramillo et al. (2021b).

Acknowledgments

This work was initiated at the 2019 NASA Frontier Development Lab (FDL), and is based on the works of Gitiaux et al. (2019) and Jungbluth et al. (2019), which were published in workshops at the 33rd Conference and Workshop on Neural Information Processing Systems (NeurIPS 2019). NASA FDL is a public-private partnership between NASA, the SETI Institute, and private-sector partners, including Google Cloud, Intel, IBM, Lockheed Martin, NVIDIA, and Element AI.

The SDO data is courtesy of NASA/SDO and the AIA, EVE, and HMI science teams.

This work utilizes data from the MDI and the HMI distributed by the Joint Science Operations Center (JSOC) at Stanford University (<http://jsoc.stanford.edu/>). Programmatic access to these data sources is available through the `drms` Python package (Glogowski et al. 2019). We also use data from the National Solar Observatory’s Integrated Synoptic Program, which is operated by the Association of Universities for Research in Astronomy, under a cooperative agreement with the National Science Foundation and with additional financial support from the National Oceanic and Atmospheric Administration, the National Aeronautics and Space Administration,

and the United States Air Force. The GONG network of instruments is hosted by the Big Bear Solar Observatory, High Altitude Observatory, Learmonth Solar Observatory, Udaipur Solar Observatory, Instituto de Astrofísica de Canarias, and Cerro Tololo Interamerican Observatory.

This research was partially supported by NASA grants 80NSSC18K0671, 80NSSC19M0165, 80NSSC19K1207, and 80NSSC20K0602. A.M.-J. acknowledges support from IR grants 15-R6134 and 15-R6134 from the Southwest Research Institute. P.J.W. acknowledges support from NASA Contract NAS5-02139 (HMI) to Stanford University. A.G.B. is supported by EPSRC/MURI grant EP/N019474/1 and by the Lawrence Berkeley National Lab. We thank Sairam Sundaresan and Santiago Miret from Intel for their advice on this project.

Our CNN training framework was built using PyTorch (Paszke et al. 2019). Patch alignment during data preparation was performed using Scikit-image (van der Walt et al. 2014). All our analysis was performed using the AstroPy (Astropy Collaboration et al. 2018), NumPy (Harris et al. 2020), SciPy (Virtanen et al. 2020), SunPy (The SunPy Community et al. 2020), and Matplotlib (Hunter 2007) packages.

Facilities: SDO, SOHO, GONG.

Software: PyTorch (Paszke et al. 2019), Scikit-image (van der Walt et al. 2014), AstroPy (Astropy Collaboration et al. 2018), NumPy (Harris et al. 2020), SciPy (Virtanen et al. 2020), SunPy (The SunPy Community et al. 2020), drms (Glogowski et al. 2019), and Matplotlib (Hunter 2007).

Appendix Supplementary Information

A.1. 24 hr Variability in HMI Data

As an extension to Figure 7, in Figure 9 we show the HMI and SR MDI time series along with the ratio of time series, and the radial velocity of HMI. The observed oscillations arise from a Doppler shift in the spectral line due to the orbital variation of the spacecraft (Couvidat et al. 2016).

A.2. Effect of Signal-to-noise Ratio

Our results indicate that the effectiveness of the neural network to cross-calibrate and superresolve the magnetic field is sensitive to the signal-to-noise ratio within a magnetogram. The signal-to-noise ratio is affected by (1) the strength of the magnetic field itself, and (2) the proximity to the solar limb. HMI's noise level is 15 (Hoeksema et al. 2014). Figure 10 compares the Pearson correlation metric calculated for super-resolved MDI magnetograms as a function of patch location and magnetic field value. The gray-shaded histograms were calculated for patches across the full solar disk, while the blue-shaded histograms were calculated for patches that lie within 90% of the radius of the solar disk. In addition, we compare the Pearson correlation for all patches (left column), and those that have an average unsigned field larger than 15 G (right column). Looking at all magnetic field values, we can see that the distribution of Pearson correlation coefficients shows two peaks around 0.25 and 0.75 when patches across the entire solar disk are considered (Figure 10, left column, gray histogram). Discarding patches that lie outside of the central 90% of the radius of the solar disk removes the double peak and shifts the distribution to be asymmetrically centered around 0.8 (Figure 10, left column, blue histogram).

When we also disregard patches that show magnetic field strengths close to HMI's noise level, we see that the distribution of Pearson correlation coefficients becomes substantially narrower and is more symmetrically centered around 0.8. This observation supports our finding that magnetogram patches with field strengths around the noise level are harder to align and less reliable for the neural network to learn. In addition, near the solar limb, the magnetic field intensity weakens due to projection effects and a reduction of the effective resolution of the instruments. In the post-mortem evaluation of our results, we therefore focus on patches that lie within 90% of the radius of the solar disk, unless otherwise specified.

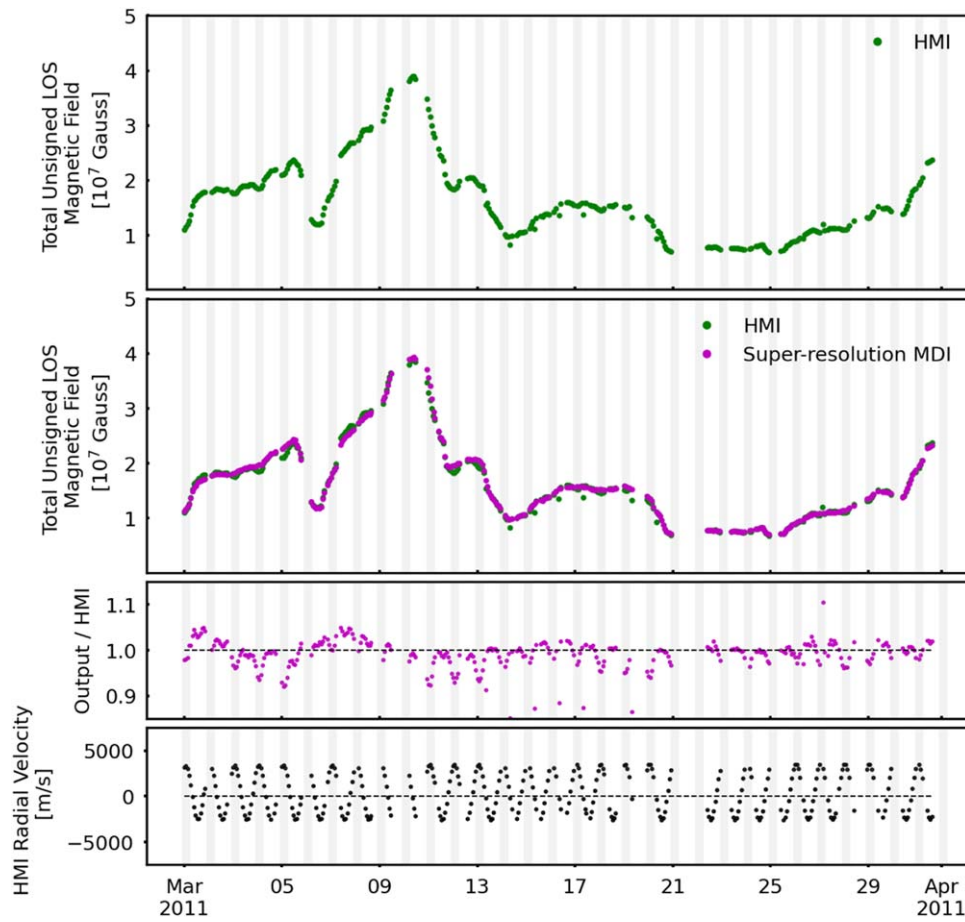


Figure 9. Demonstration of the HMI radial velocity leaking into HMI LOS magnetic field data. The top two panels show the total unsigned magnetic field from HMI and the SR output of MDI. The bottom two panels show the ratio of SR output compared to HMI, and the HMI radial velocity where positive values are away from the Sun (bottom panel). In each panel, vertical gray bars are shown with a 24 hr periodicity starting on 2011 March 1. It can be seen that this periodicity leaks into the HMI data, but is not observed for the SR output (as further discussed in the main text).

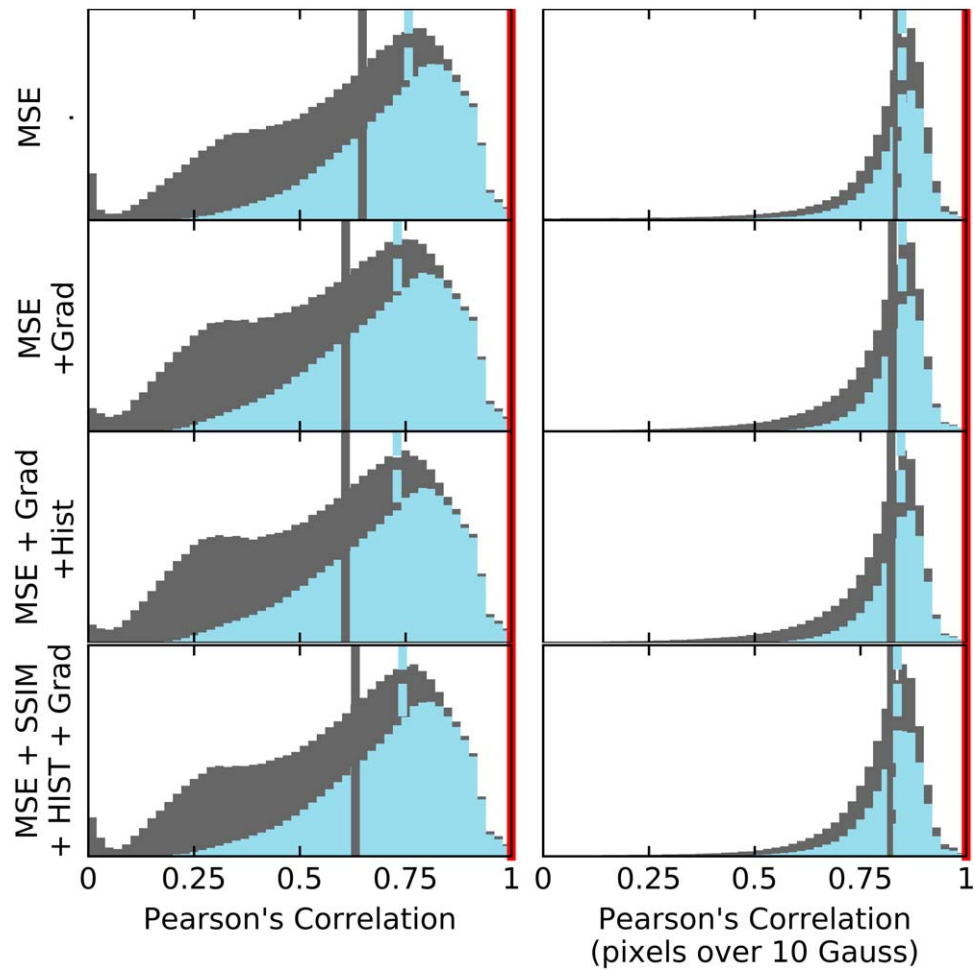


Figure 10. Quantitative comparison of the performance of different loss functions trained on MDI magnetograms. All loss functions are based on the MSE term, plus up to three additional penalty terms. The gray-shaded histograms correspond to calculations performed across the full solar disk, with the gray solid line showing the distribution average. The blue histograms were calculated for patches that fall within 90% of the solar disk radius, with the blue dashed line showing the distribution average. The first column shows the Pearson correlation calculated for all magnetic field values, while the second column shows the Pearson correlation calculated for magnetic fields with an absolute value above 10 G. In the ideal case, the Pearson correlation is 1 (indicated by the red line).

A.3. Overview of SHARPs for MDI and GONG











Table 6 in Appendix A.3 shows that the benefits of SR extend to all SHARPs.

Table 6
Baseline (Base) and SR Comparison of Different HARPs for MDI and GONG

HARPNUM	Unsigned Flux (10^{21} Mx)				Avg. Gradient (10^{-8} G cm $^{-1}$)			
	MDI		GONG		MDI		GONG	
	Base	SR	Base	SR	Base	SR	Base	SR
392	6.1	0.96	4	0.3	0.36	0.18	0.08	0.096
393	13	3.1	15	1	0.2	0.1	0.067	0.089
394	4.7	0.38	6.5	0.3	0.21	0.12	0.094	0.071
399	0.35	0.041	0.7	0.043	0.22	0.11	0.069	0.091
401	7.5	1.4	9.1	1.4	0.51	0.22	0.11	0.13
403	0.79	0.089	1.7	0.15	0.27	0.13	0.081	0.12
407	5.1	1.2	6.5	0.73	0.33	0.15	0.077	0.098
409	0.76	0.099	1.9	0.17	0.33	0.16	0.089	0.1
411	0.2	0.029	0.62	0.056	0.3	0.14	0.086	0.12
414	1.5	0.64	1.2	0.082	0.13	0.062	0.045	0.051
415	2.6	0.76	6.1	0.52	0.33	0.16	0.082	0.1
419	0.059	0.023	0.17	0.014	0.4	0.19	0.17	0.15
421	3.9	0.77	5.4	0.67	0.27	0.14	0.061	0.084
423	0.36	0.1	0.48	0.026	0.49	0.25	0.081	0.12
425	0.44	0.08	0.99	0.057	0.32	0.16	0.071	0.093
427	0.24	0.062	0.57	0.039	0.36	0.17	0.093	0.099
429	1.3	0.35	2	0.19	0.36	0.2	0.09	0.12
431	0.82	0.27	0.53	0.029	0.11	0.058	0.016	0.066
432	0.067	0.027	0.18	0.0054	0.19	0.13	0.16	0.13
433	0.053	0.013	0.34	0.037	0.11	0.069	0.0092	0.1
436	0.52	0.13	0.57	0.1	0.41	0.22	0.092	0.12
437	13	2.2	17	1.9	0.34	0.18	0.088	0.11
438	2.4	0.55	2.5	0.45	0.24	0.13	0.06	0.091
443	2.7	0.29	2.5	0.64	0.22	0.11	0.047	0.085
444	6	0.8	7.4	0.57	0.23	0.11	0.058	0.08
451	4.7	0.41	4.5	0.78	0.32	0.15	0.088	0.1
452	0.12	0.042	0.22	0.038	0.33	0.19	0.13	0.15
454	0.094	0.058	0.18	0.012	0.33	0.23	0.2	0.18
455	0.18	0.088	0.2	0.012	0.61	0.36	0.24	0.2
456	0.0029	0.021	0.17	0.011	0.0023	0.00062	0.0018	0.016
459	0.082	0.025	0.31	0.019	0.29	0.15	0.14	0.13
461	0.25	0.1	0.3	0.019	0.25	0.16	0.11	0.12
462	0.8	0.073	1.1	0.099	0.29	0.14	0.075	0.086
465	0.11	0.076	0.2	0.021	0.34	0.21	0.11	0.12
466	5.7	0.75	4	0.86	0.13	0.079	0.015	0.056

Note. The rms error for each HARP's average gradient and unsigned flux relative to HMI during its lifetime for all SHARPs during the month of 2011 March are shown.

ORCID iDs

Andrés Muñoz-Jaramillo  <https://orcid.org/0000-0002-4716-0840>
 Anna Jungbluth  <https://orcid.org/0000-0002-9888-6262>
 Xavier Gitiaux  <https://orcid.org/0000-0002-6648-0225>
 Paul J. Wright  <https://orcid.org/0000-0001-9021-611X>
 Carl Shneider  <https://orcid.org/0000-0002-3689-6959>
 Shane A. Maloney  <https://orcid.org/0000-0002-4715-1805>
 Atılım Güneş Baydin  <https://orcid.org/0000-0001-9854-8100>
 Yarin Gal  <https://orcid.org/0000-0002-2733-2078>
 Michel Deudon  <https://orcid.org/0000-0000-0000-0000>
 Freddie Kalaitzis  <https://orcid.org/0000-0002-1471-646X>

References

- Asensio Ramos, A., Cheung, M. C. M., Chifu, I., & Gafeira, R. 2023, *LRSP*, 20, 4
- Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, *AJ*, 156, 123
- Bobra, M. G., Sun, X., Hoeksema, J. T., et al. 2014, *SoPh*, 289, 3549
- Borrero, J. M., & Ichimoto, K. 2011, *LRSP*, 8, 4
- Couvidat, S., Schou, J., Hoeksema, J. T., et al. 2016, *SoPh*, 291, 1887
- Dahl, R., Norouzi, M., & Shlens, J. 2017, arXiv:1702.00783
- Deudon, M., Kalaitzis, A., Goytom, I., et al. 2020, arXiv:2002.06460
- Díaz Baso, C. J., & Asensio Ramos, A. 2018, *A&A*, 614, A5
- Domingo, V., Fleck, B., & Poland, A. I. 1995, *SoPh*, 162, 1
- Forsyth, D. A., & Ponce, J. 2002, *Computer Vision: A Modern Approach* (Hoboken, NJ: Prentice Hall)
- Gitiaux, X., Maloney, S. A., Jungbluth, A., et al. 2019, arXiv:1911.01486
- Glogowski, K., Bobra, M. G., Choudhary, N., Amezcu, A. B., & Mumford, S. J. 2019, *JOSS*, 4, 1614
- Guo, J., Bai, X., Liu, H., et al. 2021, *A&A*, 646, A41
- Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, *Natur*, 585, 357
- Harvey, J. W., Hill, F., Kennedy, J. R., Leibacher, J. W., & Livingston, W. C. 1988, *AdSpR*, 8, 117
- Hathaway, D. H. 2010, *LRSP*, 7, 1
- Higgins, R. E. L., Fouhey, D. F., Antiochos, S. K., et al. 2022, *ApJS*, 259, 24
- Hoeksema, J. T., Liu, Y., Hayashi, K., et al. 2014, *SoPh*, 289, 3483
- Hunter, J. D. 2007, *CSE*, 9, 90
- Jungbluth, A., Gitiaux, X., Maloney, S. A., et al. 2019, arXiv:1911.01490
- Kingma, D. P., & Ba, J. 2014, arXiv:1412.6980
- Linker, J. A., Mikić, Z., Biesecker, D. A., et al. 1999, *JGR*, 104, 9809
- Liu, Y., Hoeksema, J. T., Scherrer, P. H., et al. 2012, *SoPh*, 279, 295
- Mackay, D. H., Green, L., & Van Ballegoijen, A. 2011, *ApJ*, 729, 97
- Muñoz-Jaramillo, A., Wright, P., Jungbluth, A., & Gitiaux, X., 2021a amunozj/magnetograph_2HMI_converter; v0.5, Zenodo, doi:10.5281/zenodo.5784205
- Muñoz-Jaramillo, A., Wright, P. J., Jungbluth, A., & Gitiaux, X. 2021b, Upscaled and calibrated GONG and MDI magnetograms via Deep Learning, v0.1, Zenodo, doi:10.5281/zenodo.5792172
- Paszke, A., Gross, S., Massa, F., et al. 2019, *Advances in Neural Information Processing Systems* 32, ed. H. Wallach et al. (NeurIPS), 8024, https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html
- Pesnell, W. D., Thompson, B. J., & Chamberlin, P. C. 2012, *SoPh*, 275, 3
- Pingle, K. K. 1969, in *Automatic Interpretation and Classification of Images*, ed. A. Grasselli (New York: Academic), 277
- Rahman, S., Moon, Y.-J., Park, E., et al. 2020, *ApJL*, 897, L32
- Raissi, M., Perdikaris, P., & Karniadakis, G. 2019, *JCoPh*, 378, 686
- Riley, P., Ben-Nun, M., Linker, J. A., et al. 2014, *SoPh*, 289, 769
- Scherrer, P. H., Bogart, R. S., Bush, R. I., et al. 1995, *SoPh*, 162, 129
- Scherrer, P. H., Schou, J., Bush, R. I., et al. 2012, *SoPh*, 275, 207
- Schou, J., Scherrer, P. H., Bush, R. I., et al. 2012, *SoPh*, 275, 229
- Shukla, A., Merugu, S., & Jain, K. 2020, in *A Technical Review on Image Super-resolution Techniques*, ed. V. K. Gunjan et al. (Singapore: Springer Singapore), 543
- The SunPy Community, Barnes, W. T., Bobra, M. G., et al. 2020, *ApJ*, 890, 68
- Tóth, G., Sokolov, I. V., Gombosi, T. I., et al. 2005, *JGRA*, 110, A12226
- van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., et al. 2014, *PeerJ*, 2, e453
- Virtanen, I., & Mursula, K. 2019, *A&A*, 626, A67
- Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, *NatMe*, 17, 261
- Wang, Z., Bovik, A., Sheikh, H., & Simoncelli, E. 2004, *ITIP*, 13, 600
- Wang, Z., Chen, J., & Hoi, S. C. H. 2020, arXiv:1902.06068
- Wang, Z., Li, H., Ouyang, W., & Wang, X. 2018, arXiv:1804.09398
- Yang, W., Zhang, X., Tian, Y., et al. 2019, *IEEE Trans. Multimed.*, 21, 3106