INFERRING SOLAR MAGNETIC STRUCTURE FROM EUV IMAGES USING A FULLY CONVOLUTIONAL NEURAL NETWORK

BY

Zena L. Stevenson, B.S.E.E.

A thesis submitted to the Graduate School

in partial fulfillment of the requirements

for the degree

Master of Science

Major: Electrical Engineering

NEW MEXICO STATE UNIVERSITY

LAS CRUCES, NEW MEXICO

July 2021

Zena Stevenson

Candidate

Electrical Engineering

Major

This Thesis is approved on behalf of the faculty of New Mexico State University, and it is acceptable in quality and form for publication:

Approved by the thesis Committee:

Dr. Laura E. Boucheron

Chairperson

Dr. Steven Sandoval

Committee Member

Dr. Nancy Chanover

Committee Member

DEDICATION

Dedicated to my family- without your encouragement I certainly would not have made it this far. Thank you for always being there to support me when things got hard. Special thanks to one very patient dog as well, whose services as a "rubber duck" and insistent reminders to take a break once in a while were vital to preserving my sanity throughout my graduate school career.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Laura E. Boucheron, for her dedication, advice, encouragement, and humor over the past few years that made this possible. Thank you for sharing your passion for this field with me, for all the times you reminded me that research is supposed to have its ups and downs, for the casually-mentioned solutions that ended up fixing huge problems that were driving me nuts, and for all the laughs had over squeaky ducks/pigs and carrot farms.

VITA

April 23, 1996	Born at Red River, New Mexico
2014 - 2018	B.S., New Mexico State University, Las Cruces, New Mexico
2016 - 2018	Research Assistant, NASA PDS Atmospheres Node, Department of Astronomy, New Mexico State University.
2018 - 2021	M.S., New Mexico State University, Las Cruces, New Mexico
2018 - 2021	Research Assistant, Klipsch School of Electrical and Computer Engineering, New Mexico State University.
2020 - 2021	Graduate R&D Intern, Sandia National Laboratory, Albuquerque, New Mexico

FIELD OF STUDY

Major Field: Electrical Engineering

Area of Specialty: Image Processing and Machine Learning

ABSTRACT

INFERRING SOLAR MAGNETIC STRUCTURE FROM EUV IMAGES USING A FULLY CONVOLUTIONAL NEURAL NETWORK

BY

ZENA L. STEVENSON, B.S.

Masters of Science

New Mexico State University

Las Cruces, New Mexico, 2021

Dr. Laura E. Boucheron, Chair

This work presents several approaches to training a fully convolutional neural network to predict an image of the solar magnetic field (a line-of-sight magnetogram) at a solar active region from an extreme ultraviolet 304 Å image of said active region. We examine the training dataset characteristics and network parameters needed to produce a model that can generalize the modality translation from extreme ultraviolet (EUV) images to magnetograms, and discuss several methods to address the challenges of predicting signed magnetograms specifically

(as opposed to predicting only the magnitude of magnetic activity). We conclude that the best approach for signed magnetograms is one that takes advantage of prior knowledge of how active region location affects the structure of the solar magnetic field, and discuss how this approach can be used in a multiple-model system to achieve performance on unseen active regions that is similar to performance on those seen in training. We also examine these models' ability to predict magnetic flux values accurately for a large range of flux values, and find that while there is a tendency to under-predict total unsigned flux and signed models predict positive flux more accurately than negative, there is generally good enough correlation between target and predicted flux levels that restricting target flux values to a small range is not necessary.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

This work describes the use of a fully convolutional neural network (FCNN), specifically a U-Net architecture [1], to predict an image of the magnetic field at a solar active region (AR) from an extreme ultraviolet image of the AR. We use 304 Å EUV images and line-of-sight (LOS) magnetograms from the National Aeronautics and Space Administration's (NASA) Solar Dynamics Observatory (SDO) [2] for this work.

We discuss expanding the use of this network architecture from small-scale tests, similar to those that it has proven to be successful at in the past [3], to the task of predicting full signed magnetograms for arbitrary ARs, which involves larger volumes of data and greater structural variation between images than previous applications.

In addition to exploring the limits of U-Net by extending it to this new dataset, we find suitable training dataset characteristics (including overall data volume, number of ARs data are drawn from, and location of those ARs on the Sun) to produce a model that can generalize the translation from EUV image to magnetogram. When generalization is achieved, these models perform on a similar level regardless of whether the AR in question has been seen before in training.

This work begins by predicting only the magnitude of magnetic activity, but is

extended to full signed magnetograms as well. It explores the challenges presented by switching to signed magnetograms and proposes several methods to address them, including the use of data augmentation and leveraging prior knowledge of the behavior of solar magnetic fields (and how this effects the structure of the magnetograms) to improve the quality of the predicted images. Working with data from only one hemisphere at a time is found to produce the best performance among models predicting signed magnetograms.

The choice of 304 Å as the EUV data modality was motivated by the possibility to extend this image translation to similar data from other sources besides SDO. Current methods for estimating the magnetic structure of ARs on the far side of the Sun (not visible from earth) rely chiefly on helioseismology instead of direct imaging (e.g., [4, 5, 6]), and these estimates tend to be low resolution. If 304 Å EUV images can be used to predict the magnetic structure instead, higher resolution estimates of far-side magnetic activity could be acquired, as there are other missions that have 304 Å data available for the far side of the Sun as well (for instance, NASA's Solar Terrestrial Relations Observatory (STEREO) [7]).

Past work has been done [8] using deep learning and 304Å images to predict far-side magnetograms, and does produce higher resolution images compared to methods based in helioseismology. However, these predictions are still significantly lower resolution than true magnetograms. This approach also clipped the range of magnetic flux values significantly and thus the predicted flux levels are often very

different than those in the original magnetograms- we allow for a much larger range of possible flux values, and evaluate how well predicted magnetic flux matches target flux for this expanded range. We find that while our methods do tend to under-predict flux levels to various degrees and signed models show better prediction of positive flux than negative, this is not so severe that the flux values in target images should need to be clipped to the same degree as seen in [8] in future work.

In Chapter 2 we discuss background information and previous methods that have used a similar network architecture for other scientific applications, as well as other work done related to image translation for SDO data. In Chapter 3 we discuss the choice of data for this work, preprocessing methods used, the data selection process, the specific characteristics and training parameters of each experiment, and metrics used to evaluate performance. We present and discuss the results of these experiments in Chapter 4. Finally, we conclude and discuss future work in Chapter 5.

2 RELATED WORK

Here we will outline relevant background information related to fully convolutional neural networks, the development of and previous applications for the specific network architecture we utilize for this project, and related works involving using deep learning models for image translation on SDO data.

2.1 Fully Convolutional Neural Networks

Convolutional neural networks (CNNs) [9] are a widely used type of deep learning model with a variety of applications in image processing, including image classification [10], segmentation [11], and computer vision [12]. They differ from fully connected neural networks in that, instead of learning separate weights for each hidden node, a CNN learns filter kernels that are convolved with the input tensor using a sliding window.

A typical CNN used for image classification tasks consists of a predetermined number of convolutional layers (the number varying depending on the desired complexity of the model), each followed by a nonlinear activation function and a pooling/downsampling operation to reduce the resolution of the output. These layers are followed by one or more fully connected layers at the output of the network, which provide a single predicted label for the image.

The Fully Convolutional Neural Network (FCNN) [11] removes the final fully

connected layers found in a conventional CNN classifier, and outputs not a single label (as in image classification) but a 2D array (image) of labels. This was originally formulated as a means to extend image classification to image segmentation, and fully convolutional networks have been shown to match or outperform conventional CNNs in semantic segmentation tasks [11]. Subsequent work has explored requesting a more complex image than a segmentation out of an FCNN; for example, using an FCNN to perform image translation [13]. As with many deep learning methods, the amount of annotated training data required to successfully train FCNNs can be a significant obstacle–the curation of appropriately sized training datasets is often difficult and expensive, and large amounts of data may not be available at all for certain applications.

2.2 U-Net

U-Net [1] is an FCNN architecture which combines a contracting path that captures the context of features in the image (as a traditional classifier would) with an expanding path that preserves the localization of these features.

A diagram of the U-Net network architecture is shown in Figure 1. This network consists of two symmetrical branches: a contracting branch on the left, and an expanding branch on the right. Essentially, the contracting branch is learning to extract the features that describe the signal (input) image, and the expanding branch is learning how to reassemble those features into a target (output) image.

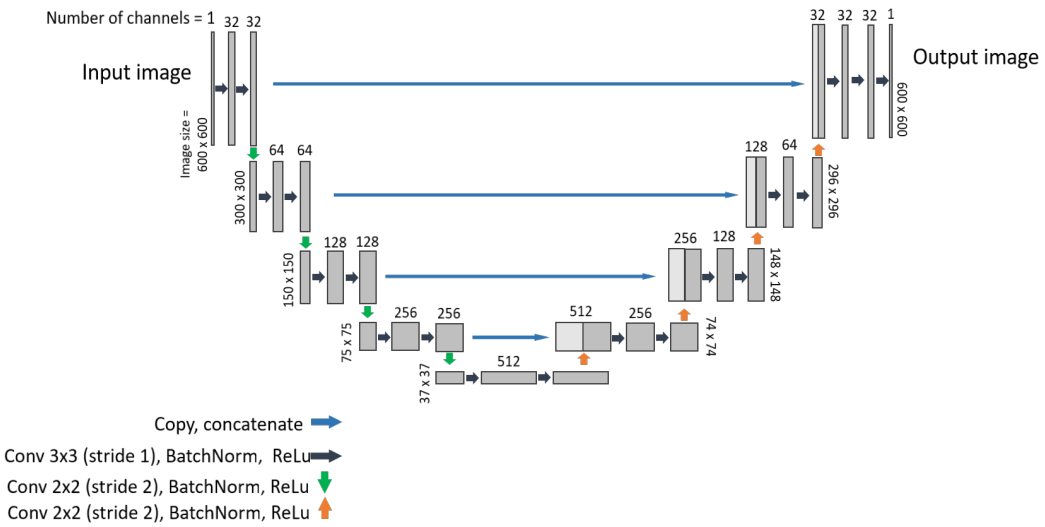


Figure 1: U-Net network architecture diagram

The number of levels in each branch (excluding the input/output level) is referred to as the depth of the network; we used a depth of 4 for this work.

Each gray block in Figure 1 represents the operation of a convolution with a set of filter kernels of size 3×3 and stride of 1, followed by a BatchNorm (which performs layer normalization separately for each training batch according to the statistics of that batch [14]) and then a ReLu operation. Each level repeats those operations twice. In the contracting branch, the feature map resulting from these operations is then copied across to the expanding branch, as well as down-sampled (via convolution with kernel size $= 2 \times 2$, stride $= 2$; this is followed by BatchNorm and ReLu operations as well) and passed to the next layer down in the contracting branch. A level of the expanding branch contains the same operations [conv(3×3, stride=1), BatchNorm, ReLu], but following this the feature map is up-sampled

via a transposed convolution (also known as a fractionally-strided convolution, where the output tensor has increased in size versus the input. The stride of these operations is still listed as = 2 in Figure 1, to match the convention used in Pytorch documentation). This is again followed by a BatchNorm and ReLu, then passed to the next layer up and appended to a copy of the feature map from the corresponding level in the contracting branch. Up- and down-sampling via convolution operations with learnable parameters is designed to let the network learn how to effectively preserve image features when up- and down-sampling the feature maps.

The practice of copying feature maps from the contracting branch across to the expanding branch is designed to introduce global context to the features that the network has learned, and prevent a loss of localization accuracy due to the reduced resolution of the feature maps in the expanding branch. Some loss of detail in predicted images compared to target images is expected regardless due to repeated down-sampling, however.

2.3 Applications in Biomedical Image Processing

U-Net was developed for applications in cellular microscopy, specifically the segmentation of cellular structures in situations where comparatively few training images were available [1]. The work done by [1] used U-Net, along with significant data augmentation (namely elastic deformations), to outperform prior

methods on International Symposium on Biomedical Imaging (ISBI) cell tracking and segmentation challenges.

A similar approach was later proposed as a means to infer fluorescence images that highlight subcellular structures from bright-field images of cells [3]. Fluorescence microscopy is used to image subcellular structures in living cells, and involves the use of fluorescent dyes and proteins to tag specific structures (e.g., DNA, cell walls). These methods are often time-consuming, expensive, damaging to the samples, and typically cannot highlight more than one structure at a time. The work done by [3] involved using U-Net to predict these fluorescence images from a bright-field microscopy image; bright-field images are collected via a much simpler sample preparation procedure than fluorescence images and at lower cost. The work in [3] was successful in creating predicted fluorescence images from both static bright-field images and from time series of multiple bright-field images of the same group of cells. This work included training several models, each to highlight different subcellular structures in the target fluorescence images, in order to combine the outputs from these models into composite images where several structures are highlighted at once.

In the same way that it is difficult for the human eye to look at a bright-field microscopy image and pick out the structures highlighted by fluorescent tagging, it is expected that there are similar structures in EUV images of solar ARs that are indicative of the underlying magnetic field structure. It has been demonstrated

that there is a correlation between the EUV intensity and the underlying magnetic field strength [15]. The accuracy and intricacy of the predicted fluorescence images in [3] indicate the potential for these FCNNs to learn to translate between EUV and magnetogram images in a similar manner, that goes beyond simple mapping of EUV intensity to magnetic field strength.

The time series experiments conducted by [3] involved using several images from a time series of the same group of cells as training data, but reserving an image from the center of the time series for testing. Other than this method of choosing test data, these time series experiments are conceptually similar to what we would like to do with solar AR images, as we have a time series of images available for each AR as well. Selecting training data based on position in a time series is not ideal for our purposes, as we would like to be able to predict magnetic activity for an AR without the network having any prior knowledge of that particular AR or its evolution over time. We also expect the transfer between solar imaging modalities to be more difficult for the network to generalize than the transfer from bright-field to fluorescence microscopy images; the shape and structure of different active regions is highly variable compared to cellular structures, which have variation in their visual characteristics as well but will generally show similar shapes and structures between samples.

2.4 Image Translation for SDO Data

Several previous works have used deep learning for image translation between modalities of SDO data.

The work of [16] provides a precedent for a CNN performing image translation between HMI and AIA data, although the translation in this work is the reverse of what we intend to do and uses slightly different data modalities. The authors in [16] predict multiple AIA wavelengths from vector magnetograms, as opposed to predicting line-of-sight magnetograms from a single AIA wavelength, and are using images of the full solar disk (sized 256×256 pixels) as opposed to images focused on single ARs (which is what we focus on in this thesis). The work in [16] uses a different CNN architecture as well (an initial ResNet feature extraction followed by a variable number of convolutional layers, then a bilinear upsampling to produce a predicted image). The quality of the predictions in [16] varies with which wavelength of AIA image is being predicted–their lowest error when predicting 304 Å images is below the average across all wavelengths for their overall best-performing model, and is the sixth best performing wavelength (out of nine) for that model.

Similarly, the work of [17] uses LOS HMI magnetograms to predict images for nine AIA modalities (including 304 Å). They utilize a network architecture modeled after an encoder-decoder network, which has two separate branches that

down-sample and then up-sample feature maps respectively, similar to U-Net. However, the network in [17] lacks the practice of copying feature maps between the two branches. The authors in [17] compare the performance of this encoder-decoder-like network on its own to performance when this network was used as the generator network in a generative adversarial network (GAN) system [18]; where a second network is added as a discriminator and trained to tell the difference between the generator's predictions and real ground truth images. Similar to [16], [17] saw average performance for 304 Å predictions compared to other wavelengths. The authors in [17] note that the second network configuration (the GAN) provided better performance; we discuss this as an avenue for future work in Chapter 5. The work in [17] uses both full-disk images of the Sun (sized 1024×1024 pixels) as well as images focused on single ARs. They note that their models struggle to predict local structure compared to global structures, especially for those wavelengths not imaging the photosphere (304 Å, for example).

The work of [19] uses GANs to predict AIA images from LOS HMI as well, and sees a significant improvements over [17] by using Pix2Pix and Pix2PixHD GAN architectures to predict multiple AIA wavelengths. In [19], 1024×1024 pixel full-disk images are used exclusively.

These works indicate that while 304 Å may not have the closest correlation to HMI magnetograms out of the AIA wavelengths, there is clear evidence of a learnable relationship between the two modalities.

Predicting HMI magnetograms from AIA images (instead of the reverse as [16], [17], and [19] did) presents an additional challenge. All of these works show evidence of high magnetic flux in magnetograms corresponding to bright areas in AIA images, and it is reasonable to expect that this correlation will hold in reverse to some degree. However, HMI images also show the sign of the magnetic flux, and it is likely that there is no clear way to tell if magnetic activity is positive or negative based on AIA signal magnitude alone. We address this via purposeful choice of the active regions used to source training data when making predictions of signed magnetograms, as discussed in detail in Chapter 3.

The work of [8] involves predicting LOS magnetograms from 304 Å EUV images, with the goal of creating far-side magnetograms from STEREO 304 Å data. They used a GAN with a U-Net architecture as the generator. The work in [8] uses both 1024×1024 pixel full-disk images, and 128×128 pixel patches focused on individual active regions. We produce higher resolution predicted magnetograms by using significantly larger images (600×600 pixel) of single active regions. The work of [8] also clipped target magnetograms to the range [-100G, 100G] prior to being used for training. This is a large degree of clipping for these magnetograms– the maximum flux values seen in the magnetograms from our datasets are often well over 1000G. We allow for a larger range of magnetic flux values, clipping to [-2550G, 2550G] instead, in order to more effectively evaluate how well the flux in predicted images corresponds to the flux levels in the original magnetograms.

3 DATA AND EXPERIMENTAL SETUP

This chapter presents an overview of the data used to train the network to predict line-of-sight magnetograms from EUV images, as well as outlining the specific datasets and training parameters used for each experiment. It includes details of the SDO data chosen for this project, the preprocessing methods applied to the data, and the process of building training and testing datasets to fit the constraints required by each experiment. We also describe each experiment (differentiated by data volume and preprocessing methods used) in detail, describe the training process and choice of network parameters, and review the metrics that will be used to asses model performance.

3.1 Data

This project uses data from NASA's Solar Dynamics Observatory (SDO). SDO has been observing the Sun since May 2010 with the goal of studying the solar atmosphere in several wavelengths and modalities simultaneously to improve our knowledge of space weather and Earth-Sun interactions [2]. It carries three instruments: the Atmospheric Imaging Assembly (AIA) [20], the Helioseismic and Magnetic Imager (HMI) [21], and the Extreme ultraviolet Variability Experiment (EVE) [22]. For this project, we are using data from AIA and HMI.

Data products available from HMI include dopplergrams (maps of solar sur-

face velocity), continuum filtergrams (broad-wavelength photographs of the solar photosphere), and both vector and line-of-sight (LOS) magnetograms (maps of the photospheric magnetic field) [6]. A dataset consisting of LOS magnetogram images focused on 1655 NOAA active regions that appeared on the Sun from 01 May 2010 through 31 December 2018 has previously been developed by [23]. This dataset consists of 1,372,004 images of 600×600 pixel magnetograms centered on NOAA ARs. In this work, we use these HMI magnetograms as the target image (desired output) for the FCNN.

For signal images (input to the FCNN), we collected AIA 304 Å images corresponding to each of these HMI magnetograms. We thus essentially double the size of the dataset reported in [23]. AIA simultaneously images the Sun in 10 different wavelengths, ranging from 94 to 4500 Å [20]. 304 Å is emitted by Helium-II at roughly 50,000 K, and is emitted from the transition region and chromosphere of the solar atmosphere, just above the photosphere. While there are several AIA modalities that do image the photosphere itself, 304 Å was chosen instead because if 304 Å can be used to predict magnetic activity, we could extend our approach to data from other spacecraft that do not image the photosphere. Specifically, NASA's Solar Terrestrial Relations Observatory (STEREO) [7] would be an excellent candidate for this because it has 304 Å images available for the entire Sun, while SDO only sees the near side of the Sun facing the Earth. The inference of magnetic field from STEREO images is particularly intriguing since most methods

that infer far-side AR structure rely on other methods such as helioseismology, e.g., [4, 5, 6], which result in low spatial resolution estimates.

As discussed in Chapter 2, there has been other work aiming to use deep learning to predict far-side magnetograms from STEREO data [8], though these methods also have their limitations related to image resolution and available flux range.

The LOS HMI images used here have a minimum cadence (how often an image is taken) of 720 s, while AIA images have a minimum cadence of 12 s. The 304 Å images were downloaded at a reduced cadence of 720 s to align with corresponding HMI images. As such, the time interval separating corresponding AIA and HMI images is 720 s or less, but can vary slightly due to jitter in image acquisition times.

3.1.1 Preprocessing

The data went through several preprocessing steps, some of which differed according to the needs of each experiment, prior to being introduced to the network for training and testing. This preprocessing is designed to compress the data and reduce storage space and memory requirements, while introducing minimal quantization and conversion error.

Because storage and transfer speed can be significant obstacles when working with large datasets, both AIA and HMI images were scaled to the `uint8` range

of $[0, 255]$ and additionally converted from their native `.fits` format into `.tiff` (which is the format the network expects as input, as well as one that generally requires less storage space than `.fits`).

HMI data were first clipped to the range [-2550, 2550] G prior to this scaling. Error introduced by this clipping is expected to be negligible, as few HMI images have intensities outside of this range–only 0.005% of the images were affected by the clipping operation. A linear intensity scaling to [0, 255] results in a quantization step size of

$$Q = \frac{2550 - (-2550)}{255} = 20 \text{ G} \tag{1}$$

and a quantization noise in the range $-10 \leq e \leq 10$ G with standard deviation $\sigma_e = 5.8$ G for signed HMI data [24]. The HMI instrument has a noise with standard deviation $\sigma = 6.3$ G for 720 s LOS magnetograms [25], so any error from the conversion to `uint8` is on the same order as the inherent noise of the instrument and thus insignificant for our purposes.

AIA data were also scaled to the `uint8` range of [0, 255]. No clipping was performed beforehand, so quantization noise will vary slightly depending on the range of values in the original image. The average maximum value in the AIA images was found to be 2205 and the average minimum value was 0, giving an average quantization step size when scaled to [0, 255] of

$$Q = \frac{2205 - (0)}{255} = 8.647 \tag{2}$$

and a quantization noise in the range $-4.324 \leq e \leq 4.324$ with standard deviation $\sigma_e = 2.5$.

Not performing any clipping and scaling according to the range of each image instead of a uniform maximum value raises concerns about losing information related to the relative intensity between images– i. e., images with very different maximum values in the original `.fits` files will end up scaled to a similar range in the conversion to `uint8` values. Due to this, two experiments using AIA data that had been uniformly clipped (to [1, 2550]) before scaling were conducted, but these experiments had poorer performance than their directly scaled counterparts. A brief discussion of these results and possible reasons behind the drop in quality are included in Appendix I.

AIA images are photon-noise limited (except at very faint signal levels), and this noise scales as $\sqrt{DN}$ where DN is the digital number (photon counts converted from electron charge into a pixel value) [26]. Therefore we can assume that, except at relatively low pixel values, the quantization noise introduced by scaling and converting the images is less than the photon noise already present.

Depending on which experiment the data is being prepared for, AIA data may be converted to a log10 scale prior to scaling to [0, 255] and converting to `.tiff`. The log10 scaling of AIA data is often done for the ease of a human viewer, since

the large range of values present can make it difficult see details in the image. We use both methods of scaling in order to establish whether or not the FCNN's ability to extract features from the AIA images that are relevant to the translation will benefit from log10 scaling in a similar manner.

Similarly, several experiments use an absolute value version of HMI data, to show only the magnitude of the magnetic activity. This is also done prior to scaling and converting to `.tiff`. Predicting only the magnitude of magnetic activity is assumed to be an easier problem for the network to solve, because while there is usually visible correlation between bright spots in an AIA image and high magnetic activity in a magnetogram, there is no obvious way to tell the sign of that magnetic activity from the AIA image alone. For this reason, earlier experiments use only absolute value HMI data. For later experiments with larger volumes of data, different combinations of log10- or linearly-scaled AIA and signed or unsigned magnetograms were introduced in order to explore any effect these different preprocessing methods had on prediction accuracy; further details are provided in Section 3.2. As this operation is done prior to converting to `uint8` values, these absolute value versions of the HMI data now have a quantization step size of

$$Q = \frac{2550 - (0)}{255} = 10 \text{ G} \tag{3}$$

with noise in the range $-5 \leq e \leq 5$ and standard deviation $\sigma_e = 2.9$ G.

An additional normalization that subtracts the mean of the image and divides

by its standard deviation is done immediately prior to an image being used as input to the network, in order to help with stability and convergence of the FCNN during the training process. This is reversed prior to any evaluation of metrics or other comparison between images.

3.1.2 Effect of AR Location

Several of our experiments were conducted using signed magnetograms as target images, and we would like the eventual final model to perform well on signed images specifically. Using signed HMI data introduces a slight complication to the data selection process–because the ARs used for training are selected randomly, it is possible to unintentionally select more ARs from one hemisphere (north or south) than the other. This matters because which hemisphere an AR forms in determines whether negative flux "leads" positive or vice versa, i.e., Hale's law [27]. Thus, HMI images from one hemisphere will tend to have negative flux on one side of the image and positive on the other. This effect is not perfectly uniform for all active regions from a given hemisphere, but is reliable enough that it had a noticeable effect on the predictions produced by the network. It was discovered after the first round of experiments using signed magnetograms that a majority of the training data were from the southern hemisphere and tended to have most of the positive flux on the right side of the image and most of the negative flux on the left, and so the network would apply this convention to nearly all predicted

images regardless of the orientation present in the target image. We addressed this imbalance by selecting signed training data for subsequent experiments in one of two ways.

Data Augmentation by Rotation: One method used to create a balanced training set was to introduce a random rotation (by either 0 or 180 degrees) to each training image pair for any experiments involving signed HMI images from both hemispheres. This data augmentation simulates the characteristics of a dataset drawn from both hemispheres without needing to repeat the dataset building process from scratch.

Single Hemisphere Dataset: For the second method, we instead choose all training data from one hemisphere (the northern hemisphere was chosen arbitrarily in this work). We found that this helps to remove ambiguity when predicting the sign of magnetic activity and leads to improved prediction quality overall. Test data for these experiments were selected with respect to hemisphere as well, discussed further in Section 3.1.4.

Both methods for balancing signed HMI training data (a random rotation to replicate characteristics of an even North-South split among the training data, and training data from only the Northern hemisphere) are used in experiments with both linearly- and log10-scaled AIA, as detailed in Section 3.2.

3.1.3 Building Training and Testing Datasets

For experiments that use training data from more than a single active region, the process of selecting and preprocessing image pairs to build training and testing datasets was as follows.

1. Select AR number
    (a) Check hemisphere (if applicable)
2. Select desired # of images
    (a) Verify that enough images are available
    (b) Randomly select only that many images
3. Discard invalid image pairs
    (a) Discard if timestamps >720s apart
    (b) Discard if not-a-number values (NaNs) are present
4. Clip (if applicable), re-scale, and convert to `.tiff`
    (a) Create both log10 and linear AIA, abs and signed HMI copies of images
    (b) Scale to `uint8` range
    (c) Save as `.tiff`
5. Set aside test data

First, an AR number is randomly selected from the range [1076, 2693], which are all of the NOAA ARs from which we currently have data available. If the experiment requires data from a certain hemisphere, we check which hemisphere the randomly selected AR is from using the NOAA Space Weather Prediction Center Solar Region Summary (`ftp://ftp.swpc.noaa.gov/pub/warehouse/`) file for the AR. Solar Region Summary (SRS) files are organized by date, which we can match up with the timestamp of the first image of the AR. SRS files contain a list of active region numbers that appear on that date and list general information (including coordinates) for each AR. We can then either skip the AR and randomly select another if it is from the wrong hemisphere, or continue if it is from the correct hemisphere.

Second, because the number of images available can vary significantly between ARs, we must verify that there are enough images present for the AR we have selected. For balanced representation from all ARs, the exact number of images needed per AR is N_i/N_{AR} where N_i is the total number of images desired in the dataset and N_{AR} is the number of ARs desired in the dataset. A buffer of an additional 5 images was included because subsequent steps may remove some image pairs from consideration. For example, for the 52-AR datasets where $\sim$5000 training image pairs were desired, any ARs that had less than 115 image pairs available were excluded. This should give us at least $\sim$ 100 training and $\sim$ 10 testing images from each AR, plus a small buffer. As some ARs can have many

more images than this, we then randomly select only enough images to satisfy the N_i/N_{AR} constraint.

Third, we check the timestamps of the image pairs. If the AIA and HMI images were taken more than 720 seconds apart, we remove that image pair from consideration. We also remove any images pairs that contain not-a-number (NaN) values.

Fourth, we create multiple versions of the images (both a linear- and log10-scaled version of the AIA image and both an absolute value and signed version of the HMI image), scale to `uint8` range of [0, 255], and save the images as `.tiff`. Saving multiple versions of the same images means that several of the experiments use the same image pairs, but with different combinations of preprocessing depending on the experiment. This allows for some sharing of images between experiments, which helped to decrease storage space requirements.

Fifth and finally, we set aside test data. As the process of selecting test data became more involved than simply partitioning the images processed by the above steps into a train group and a test group, this is discussed in further detail in Section 3.1.4.

These five steps are repeated until we have at least the number of ARs desired N_{AR} and at least the total number of training images desired N_i.

Table 1 summarizes the experimental datasets defined using the process outlined above. Note that this process can result in slight differences in the number

of images present from each AR during training, but the differences are small enough to not have any significant effect on the balance of the dataset.

3.1.4 Test Data

We define four processes for setting aside test data. In the first definition of test data, "CenterAR," applicable only for the 1-AR experiment, we set aside the central 10% of the image pairs (with respect to time) to be consistent with tests reported in the literature for the FCNN network we use here [3]. In the second definition of test data, "SameAR," we randomly selected (without respect to time) 10% of the image pairs for each AR to reserve for testing, while the rest are used for training. In the third definition of test data, "DiffAR," we sampled data from ARs not used in the training data. Using a similar process to the steps outlined above, we select a random AR number, verify that it is from the correct hemisphere (if applicable), make sure that the AR was not already included in the training dataset, and then process only as many images as were reserved for the corresponding SameAR test set. In the fourth definition of test data, "SouthAR" (applicable only for the signed magnetogram experiments using training data exclusively from the Northern hemisphere), we sample data from Southern hemisphere ARs, rotate those images 180 degrees, and process only as many images as were reserved for the corresponding SameAR test set.

Separating test data and evaluating performance separately in this way allows

us to better probe the generalization abilities of the trained network- examining any performance gaps between SameAR and DiffAR (or SouthAR) gives us a simple way to tell how well a model has generalized the modality transfer from AIA to HMI.

Including a SouthAR test set as well for models trained on only the Northern hemisphere allows us to more easily compare the performance of the single-hemisphere models to the models trained on data from both hemispheres. We used this approach to determine whether a single model trained using either approach (both hemispheres or only one), or two models (one trained on each hemisphere) provides the best predictions on arbitrary ARs.

Table 1 lists the number of test images present for each test dataset and each experiment.

3.2 Experimental Setup

This section covers details of each experiment, with further motivation for the data volume used in training and testing and the chosen preprocessing methods (log10 versus linear AIA, signed versus absolute value HMI). All experiments conducted, the volume of data used, and preprocessing methods used are summarized in Table 1.

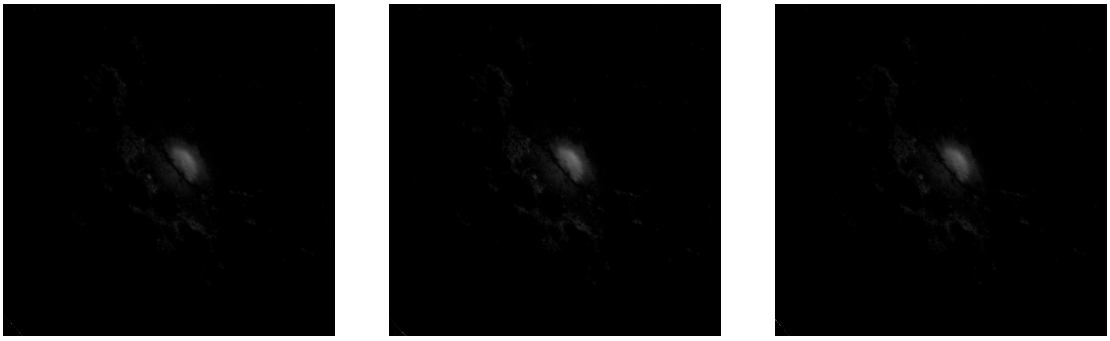
Table 1: Summary of datasets, including number of ARs used in training, type of preprocessing, and total number of images for training and test. The AIA data used as the signal has intensities either log10 scaled prior to scaling to $[0, 255]$ or a direct linear scaling to $[0, 255]$. The HMI data used as the target has the signed fluxes directly linearly scaled to $[0, 255]$ or the absolute flux values scaled to $[0, 255]$ (i.e., unsigned). Signed HMI data were balanced by random rotation by 0 or 180 degrees, or by choosing only northern hemisphere data.

#ARs	**Signal**	**Target**	**#Image Pairs**				
			Train	Center	Same	Diff	South
1	log10 AIA	unsigned (abs) HMI	30	3		6	
10	log10 AIA	unsigned (abs) HMI	5014		509	508	
52	log10 AIA	unsigned (abs) HMI	5042		521	508	
52	linear AIA	unsigned (abs) HMI	5042		521	508	
52	log10 AIA	signed HMI, rotated	5042		521	508	
52	log10 AIA	signed HMI, N only	5039		507	512	519
52	linear AIA	signed HMI, rotated	5042		521	508	
52	linear AIA	signed HMI, N only	5039		507	512	519

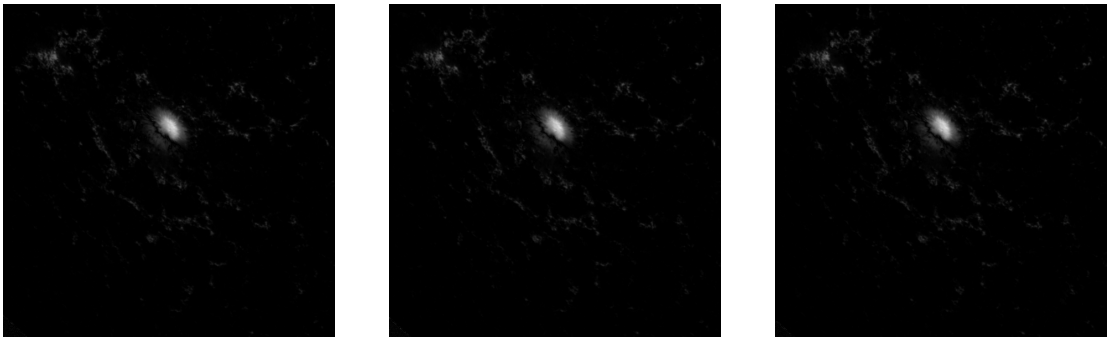
3.2.1 1-AR Experiments

We began with a small-scale proof-of-concept experiment using only one AR. This experiment was conducted both to establish whether or not the modality transfer task was successful on a new dataset at a similar volume to [3] and to provide a performance baseline for later experiments.

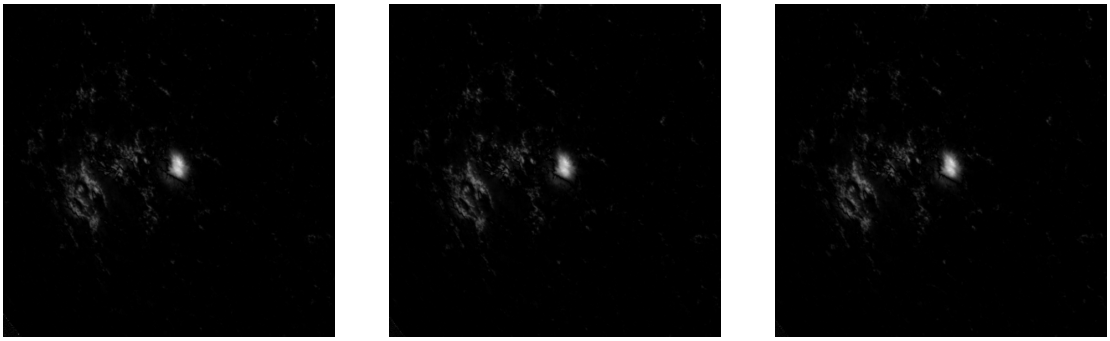
Small-Scale Experiment 1, CenterAR: Similar to the experimental setup for each of the models trained on individual cellular structures in [3], we began by training the network on a small dataset of image pairs from the same AR. Of 33 image pairs, 30 pairs were used for training, and the central three pairs (with respect to time–see Figure 2) were reserved for testing (CenterAR). This was done



(a) AR 1131, $t-2$ (b) AR 1131, t (c) AR 1131, $t+2$



(d) AR 1140, $t-2$ (e) AR 1140, t (f) AR 1140, $t+2$



(g) AR 1147, $t-2$ (h) AR 1147, t (i) AR 1147, $t+2$

Figure 2: Example small-scale experiment data where each row is a different AR and the columns are separated by two time steps. These data are the absolute value magnetograms used as target images. In Small-Scale Experiment 1 (CenterAR), for each AR (row), the first and last columns are example training data and the middle column of that row is example test data. In Small-Scale Experiment 2 (CenterAR), for each AR (row), the first and last columns are example training data and all three middle column images are test data.

independently for three ARs (each row of Figure 2).

Relative to Small-Scale Experiment 1, for each AR, example test data are illustrated in the center column of Figure 7. Note that the structure of these images is very similar to those images close in time which constitute example training data (first and last columns of the corresponding row), and thus it is expected that the network will be able to learn that specific structure easily. Three models were trained independently, each on a different AR. Each model was allowed 5×10^4 training iterations (to match [3]). The patch size was 128×128 pixels– details of the random patch selection process used during training are covered in Section 3.3.

Small-Scale Experiment 2, DiffAR: The small-scale models were also tested on image pairs from ARs not used in training. Our eventual goal is to train the network in such a way that it can make accurate magnetogram predictions for any AR–as ARs tend to vary significantly in structure, size, and location, it was expected that training on a single AR would not be sufficient to learn the underlying relationship between the two image modalities.

Relative to Small-Scale Experiment 2, training data again consists of a single AR (one of the rows in Figure 7), but now the test data are the central image pairs for the other two ARs (the middle column of Figure 7). In other words, the DiffAR test data for the model trained on AR 1131 are the same as the CenterAR test

data for the other ARs (1140 and 1147). Note that the structure of image pairs from different ARs is generally not similar, and thus the network was expected to struggle to predict this new structure.

3.2.2 10-AR Experiments

To provide a more robust training dataset, the number of ARs was increased to 10, randomly selected from the full dataset of 1665 ARs. From those 10 ARs, ~ 5000 image pairs were used to train the network, and ~ 500 pairs were reserved for testing (SameAR). In the SameAR test set, training and testing pairs were chosen randomly, without respect to time. Additionally, ~ 500 additional image pairs from 50 ARs not used in training were added as a separate test set as well (DiffAR). The 10-AR experiments used only absolute value HMI data and log10-scaled AIA data, to allow for easy comparison with the 1-AR experiments.

At this scale, we began evaluating several training parameters to determine which would give the best performance for a training dataset of this size. The details of this process are discussed in Section 3.3.1. The final values used for this experiment (and the subsequent 52-AR experiments) were 2×10^5 iterations and a patch size of 256×256 pixels.

3.2.3 52-AR Experiments

To increase the variety of training data without an increase in data volume (and the increase in training time that would come with it), the number of active regions was increased to 52, while keeping the total number of image pairs at ~ 5000. As before, ~ 50 image pairs were reserved for the SameAR test. The DiffAR testing pairs were again chosen randomly, and consisted of ~ 500 additional test pairs from 50 active regions not used in training.

Iterations, patch size, training time, and other network parameters were unchanged from the final values chosen for the 10-AR experiment. At this data volume, we began including different combinations of preprocessing methods for the training data as well (both log10 and linear AIA data, both absolute value and signed HMI data), as listed in Table 1. The introduction of signed HMI data at this stage led to separate experiments for the two methods of balancing the data with respect to AR location, discussed in Section 3.1.2.

3.3 Hardware and Training Parameters

The approximate training times, number of training iterations, and hardware used for each scale of experiment (separated by number of ARs used in training) are listed in Table 2.

All models used the Adam optimizer [28] with $\beta = [0.5, 0.999]$ and an initial learning rate of 0.001. The batch size for all experiments was 30 images, and

Table 2: Summary of training time, hardware, and number of iterations for each class of experiment (separated by the number of ARs used in training).

# ARs	**Time (hrs)**	**Computer**	**Iters**	**Patch size**
1	4	Quaffle	5E4	128×128
10	12	Discovery	2E5	256×256
52	12	Discovery	2E5	256×256

pixel-wise mean-squared error (squared L2 norm) was used as the loss function for training. All of this was implemented and run via Pytorch.

The majority of training and testing was done using NMSU's Discovery high-performance computing cluster. We worked on a single GPU node using a Nvidia Tesla P100 GPU, with 3584 CUDA cores and 16 GB GPU memory. As the small scale of the 1-AR experiments did not require use of the high-performance computing cluster to reduce training time, these were instead trained on CPU (listed as 'Quaffle' in Table 2), using an AMD FX 8120 8-core processor with a base clock speed of 3.1 GHz and 16 GB RAM.

Training time is given as a number of iterations that the network trained for rather than a number of epochs. This is the simpler way to refer to this particular training process because this implementation of U-Net does not use the entire image in each training iteration, meaning that going through a number of iterations equal to the number of training pairs does not necessarily mean an epoch has technically been completed. What the network does is select random patches of the training data and conduct a single training iteration on those patches, then

move on to another random selection of patches from the next batch of images for the next iteration. We often set the number of iterations significantly higher than the number of training pairs to make it likely that the entire training dataset will be used. This practice allows us to keep the images at their original resolution (600×600 pixels) while only using the amount of memory required to handle each patch.

3.3.1 Training Parameter Selection

Several early versions of the 10-AR experiment were conducted in order to find the best training parameters for the rest of the larger scale experiments.

The number of iterations allowed for training the 1-AR model was selected to match [3]. For the larger experiments, trial and error led to the choice of 2E5 iterations as a suitable one. Figure 3 shows training loss for an early 10AR model that was allowed to train for 5E5 iterations–it shows that the loss had stopped improving by 2E5 iterations and there was no benefit to continuing to train longer. Other parameters were altered after this longer experiment as well, which improved the stability of the final loss significantly compared to what is shown in Figure 3.

Various depths for the network (depth referring to the number of levels in each of U-Net's branches; see Section 2.2) were tested early in the process as well, because altering depth is a simple way to control the number of trainable

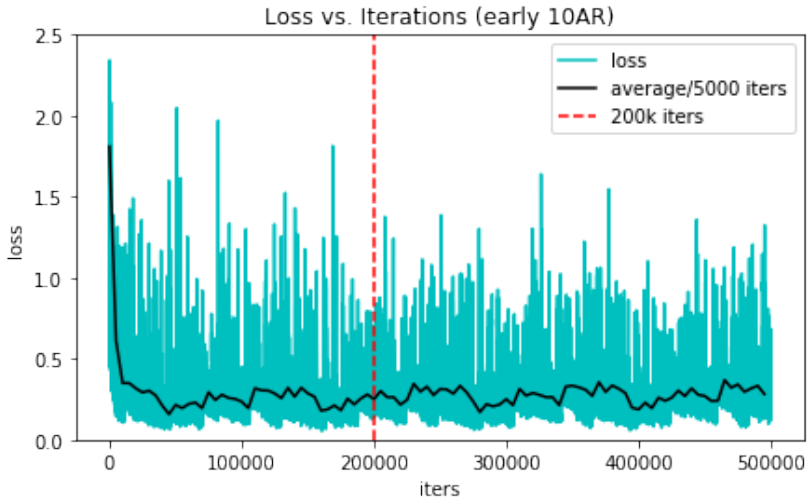
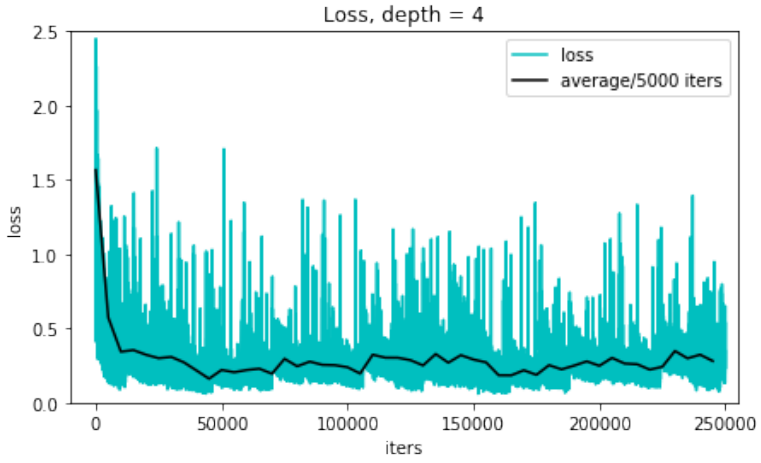


Figure 3: Loss vs. iterations for an early version of the 10AR experiment, which was allowed to run for 5E5 iterations.

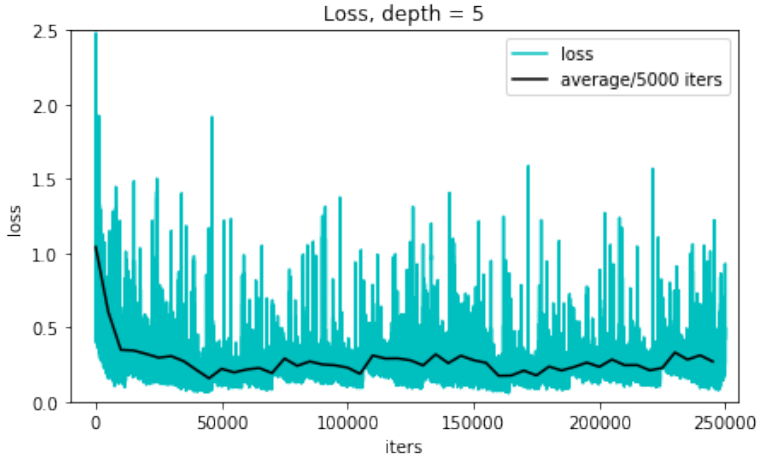
parameters in the network. Figures 4(a), (b), and (c) show training loss vs. iterations for networks with depths of 4, 5, and 6 respectively. Increasing the depth past 4 (the value used in [3]) did not yield any noticeable performance change and increased training time significantly, so network depth was left at 4 for all experiments in this work.

Multiple patch sizes were tested as well, to see if the value of 128×128 used in [3] would be suitable for the magnetogram prediction task; Figure 5 shows the results of this. We found that increasing the patch size to 256×256 pixels results in lower training loss that decreases faster and is more stable than 128×128. Increasing further to 512×512 led to the training process becoming overly memory intensive, so 256×256 was chosen as the final value.

Figure 5(b) also suggests that the final loss could have been achieved in fewer



(a)



(b)

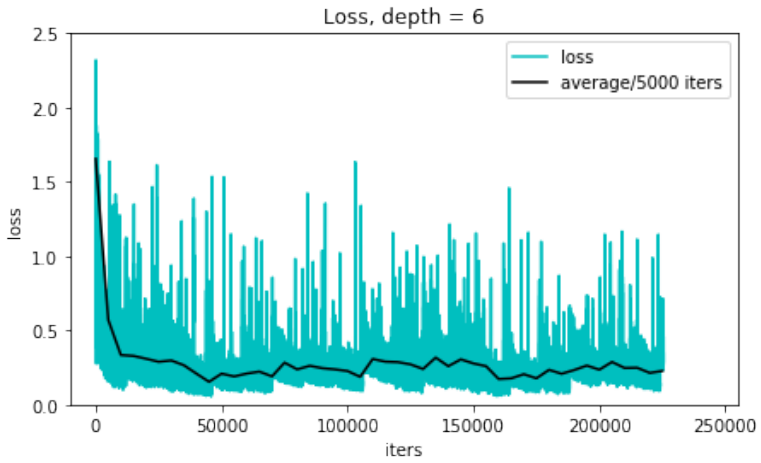
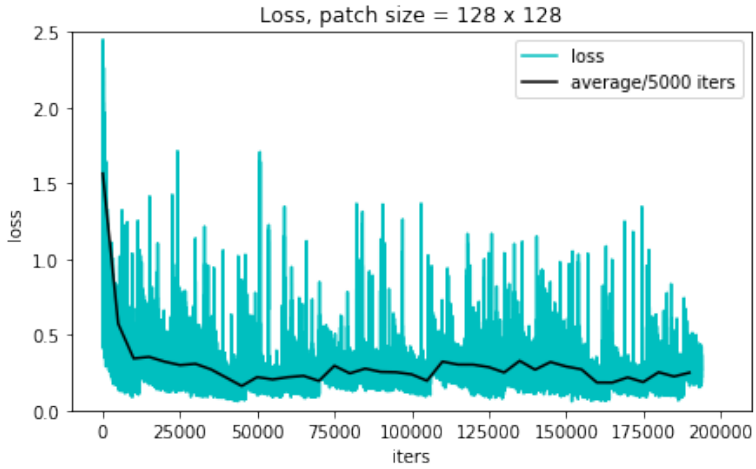
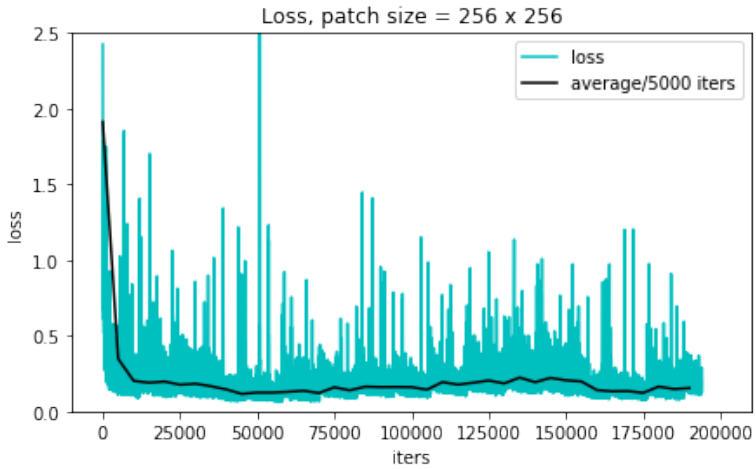


Figure 4: Training loss vs. iterations for various network depths



(a)



(b)

Figure 5: Training loss vs. iterations for various patch sizes.

than 2E5 training iterations. However, since the number of iterations where the loss spikes to significantly higher than the average is still decreasing after 1.75E5, the final value was left at 2E5.

3.4 Metrics

In this section we discuss the metrics used to assess the performance of the FCNN.

3.4.1 Normalized Root-Mean Squared Error (NRMSE)

A normalized, pixel-wise root-mean-squared error (NRMSE) was used to evaluate network performance during testing. This error is normalized using the intensity range of the ground truth (target) image and is implemented in python via `skimage.measure.compare_nrmse()`. The NRMSE is formally defined as

$$NRMSE = \frac{1}{R_T}\sqrt{\frac{1}{N}\sum_{n=1}^{N}(I_T - I_P)^2}, \tag{4}$$

where I_T is the target image, I_P is the predicted image, N is the total number of pixels in I_T and I_P, and R_T is the maximum value in the target image.

3.4.2 Structural Similarity Index Measure (SSIM)

We also used the mean structural similarity index measure (SSIM) [29] between the target and predicted images as an additional method of evaluating performance. We are interested in the structural quality of predicted images in addition to a simpler pixel-wise error measure because it helps give us a more complete picture of prediction quality. SSIM is considered a closer metric to how the human eye would evaluate the similarity of the two images, and is less sensitive than MSE to small mistakes in the location of predicted activity. SSIM was implemented in

python via `skimage.measure.compare_ssim()` and is defined as

$$SSIM = \frac{(2\mu_T\mu_P + C_1)(2\sigma_{TP} + C_2)}{(\mu_T^2 + \mu_P^2 + C_1)(\sigma_T^2 + \sigma_P^2 + C_2)}, \tag{5}$$

where μ_T and μ_P are the means of the target and prediction images, respectively; σ_T and σ_P are the standard deviations of the target and prediction images, respectively; σ_{TP} is the covariance between target and prediction image; and C_1 and C_2 are constants included to avoid instability in the denominator. These constants are defined as $C_1 = (K_1 L)^2$ and $C_2 = (K_2 L)^2$, where L is the dynamic range of the images, and we select $K_1 = 0.01$ and $K_2 = 0.03$ to match the implementation in [29].

SSIM is generally considered more useful when applied locally rather than over the entire image [29]. To arrive at the single mean SSIM values that we report, Equation (5) is applied to sections of the image via a sliding window and the mean over these sections is used as the final SSIM value. Here we used the default parameters from `skimage.measure.compare_ssim()` of a uniform window with a side length of 7, allowing for 3 pixels of overlap between sections.

3.4.3 Histograms of Predicted versus Actual Magnetic Field Strength

In addition, we examined 2D histograms (visualized as heatmaps) comparing magnetic field strength in the target images versus magnetic field strength in the prediction images (both images are first re-scaled so that the target image occupies

the same range as the original `.fits` image, so that the values correspond to the original flux levels as well as they can after being quantized). These histograms show the range of target field strength on the x-axis and predicted strength on the y-axis, and count how many pixels in the test dataset fall into each (x,y) bin. Visualizing this 2D histogram by mapping counts to colors creates a heatmap.

The characteristics of these heatmap images can be used to explore any trends that may appear in the predicted strength of magnetic activity. For example, we can examine if the network has a tendency to over- or under-predict magnetic field strength in general or performs noticeably better or worse for certain levels of field strength. This provided insight into whether or not flux strength can be predicted accurately with these methods when target magnetograms are allowed to have a larger dynamic range than that used by [8]. As this method relies heavily on pixel-to-pixel accuracy between target and predicted magnetograms, it can be expected to break down when the location of predicted magnetic activity is inaccurate.

3.4.4 Comparison of Extracted Flux Features

As the heatmaps discussed above are extremely sensitive to any errors in the location of predicted magnetic activity, using them to analyze the accuracy of predicted flux levels became difficult for experiments where pixel-wise accuracy was lower. In order to still have a way to compare flux levels between target and

predicted magnetograms, we extract and compare four image features described by [30] related to the magnetic flux. These features are the total signed magnetic flux B_{total} (sum of all pixels in the image), the total unsigned magnetic flux B_{abs} (sum of the absolute value of the image), the total positive flux B_{pos} (sum of all pixels with positive values), and the total negative flux B_{neg} (sum of all pixels with negative values).

Only B_{abs} is meaningful for absolute value magnetograms, while we examined all four for signed magnetograms. These features, while simple, are useful information to have about a magnetogram–they can and have been used as part of automated solar flare prediction [30], and revealed interesting trends about how the methods presented here tend to predict magnetic flux.

4 RESULTS

This chapter discusses the results of the aforementioned experiments. We will give a broad overview of how the results between experiments compare, followed by a more in-depth discussion of the results of individual experiments. This includes discussion of the quantitative metrics NRMSE and SSIM, the heatmaps and flux feature comparisons described in Section 3.4, as well as more qualitative analysis of example prediction images. These sets of example prediction images for other experiments have been chosen to show a range of image qualities and AR structures (unless otherwise noted), and values in these images are allowed to occupy the full display range of [0, 255] to provide enough contrast to show structure clearly. Because of this, discussion of pixel value accuracy, relative brightness, etc. between images will rely on the quantitative metrics and flux comparison methods (heatmaps and extracted flux features) instead.

4.1 Overview of Experimental Results

We examine both the pixel-wise normalized root mean squared error (NRMSE) and the structural similarity index measure (SSIM) for each experiment and present separate results for the test datasets defined in Section 3.1.3. These results are summarized in Tables 3 and 4. As we are also interested in any performance gaps between SameAR and DiffAR (and SouthAR, if applicable) performance,

Table 5 summarizes these gaps. CenterAR results for the 1-AR test have been included under the SameAR columns for space.

The first notable characteristic of these results is that, as expected, the performance gaps between SameAR and DiffAR does decrease as the number of active regions included for training increases, for both NRMSE and SSIM. We see this in the first three rows Table 5, comparing the 1-AR benchmark experiment to the results of the 10-AR experiment and the 52-AR experiment that used the same preprocessing methods (log10 AIA data and absolute value HMI data). There is a large gap between the results of the SameAR and DiffAR tests in the 1-AR results, but this gap decreases significantly for 10 ARs, and we see it nearly disappear for 52 ARs. This indicates that $\sim$52 active regions and $\sim$5000 training image pairs is a suitable lower bound for data volume to generalize the modality transfer and avoid overfitting.

For the 52-AR experiments that used absolute value HMI data, linearly scaled AIA data provides a slightly higher SSIM on both tests. However, the improvement is small and there is next to no difference in NRMSE between the two tests, so it is unlikely that we can claim any significant performance advantage to using linear versus log10 AIA data with unsigned HMI data.

When signed HMI data is introduced, we see a significant improvement in NRMSE overall compared to the experiments using absolute value HMI data. However, we also see a large decrease in SSIM values for signed HMI experiments

Table 3: Summary of NRMSE for all experiments

	NRMSE		
Experiment	**SameAR**	**DiffAR**	**SouthAR**
1-AR	0.0063	0.0465	
10-AR	0.0484	0.0521	
52-AR, abs HMI			
log10 AIA	0.0504	0.0512	
linear AIA	0.0503	0.0512	
52 AR, signed HMI			
Unbalanced, log10 AIA	0.0333	0.0326	
Unbalanced, linear AIA	0.0341	0.0331	
Rotated, log10 AIA	0.0333	0.0330	
Rotated, linear AIA	0.0340	0.0336	
N hemisphere, log10 AIA	0.0294	0.0295	0.0317
N hemisphere, linear AIA	0.0287	0.0297	0.0315

Table 4: Summary of SSIM for all experiments

	SSIM		
Experiment	**SameAR**	**DiffAR**	**SouthAR**
1-AR	0.7407	0.3985	
10-AR	0.5228	0.4961	
52-AR, abs HMI			
log10 AIA	0.4987	0.5024	
linear AIA	0.5029	0.5103	
52 AR, signed HMI			
Unbalanced, log10 AIA	0.0137	0.0135	
Unbalanced, linear AIA	0.0136	0.0133	
Rotated, log10 AIA	0.0132	0.0139	
Rotated, linear AIA	0.0132	0.0140	
N hemisphere, log10 AIA	0.0141	0.0135	0.0132
N hemisphere, linear AIA	0.0147	0.0150	0.0136

Table 5: Differences between SameAR and DiffAR (and SouthAR) metrics. Δ_{diff} = |SameAR - DiffAR|, Δ_{south} = |SameAR - SouthAR|.

	NRMSE		**SSIM**	
Experiment	Δ_{diff}	Δ_{south}	Δ_{diff}	Δ_{south}
1-AR	0.0402		0.3342	
10-AR	0.0037		0.0267	
52-AR, abs HMI				
log10 AIA	0.0008		0.0037	
linear AIA	0.0009		0.0074	
52 AR, signed HMI				
Unbalanced, log10 AIA	0.0007		0.0002	
Unbalanced, linear AIA	0.0010		0.0003	
Rotated, log10 AIA	0.0003		0.0007	
Rotated, linear AIA	0.0004		0.0008	
N hemisphere, log10 AIA	0.0001	0.0023	0.0006	0.0003
N hemisphere, linear AIA	0.0010	0.0028	0.0003	0.0011

compared to unsigned experiments. Signed experiments did generally produce predictions that appear structurally less accurate than the unsigned experiments, but the size of this decrease may also be a product of the specific types of errors the network is making in its predictions for signed magnetograms. A common issue among these models is a tendency to reverse the polarity of predicted magnetic activity–the degree to which they do this varies between models, and is discussed in detail in subsequent sections. Any areas in the images where this occurs will have comparatively very low (often < 0) local covariance (σ_{TP} in Equation 5) between the two images, which will drag the overall average SSIM down significantly.

The initial unbalanced experiments and the later two experiments balanced by rotation of the training data have very similar NRMSE values for both tests,

with unbalanced performing slightly better in DiffAR especially. SSIM behaves much the same between these experiments. For both unbalanced and rotated experiments, there are only very small differences between the performance of experiments using log10 AIA data versus linear.

Both of the single-hemisphere experiments generally outperform the unbalanced and rotated experiments in NRMSE. SSIM is a closer race– only the single hemisphere experiment with linear AIA as signal data has a higher SSIM for DiffAR than the other signed experiments. We also note that SouthAR performance for both models is generally not as good as their DiffAR performance, and Δ_{south} = |SameAR - SouthAR| in Table 5 is generally larger than Δ_{diff} = |SameAR - DiffAR|. The exception is the log10 AIA DiffAR test's SSIM, which was unexpectedly low.

This indicates that models trained on a single hemisphere will give the best performance for signed magnetograms, but that one model trained in this way does not perform as well on test data from the other hemisphere even if the data has been rotated.

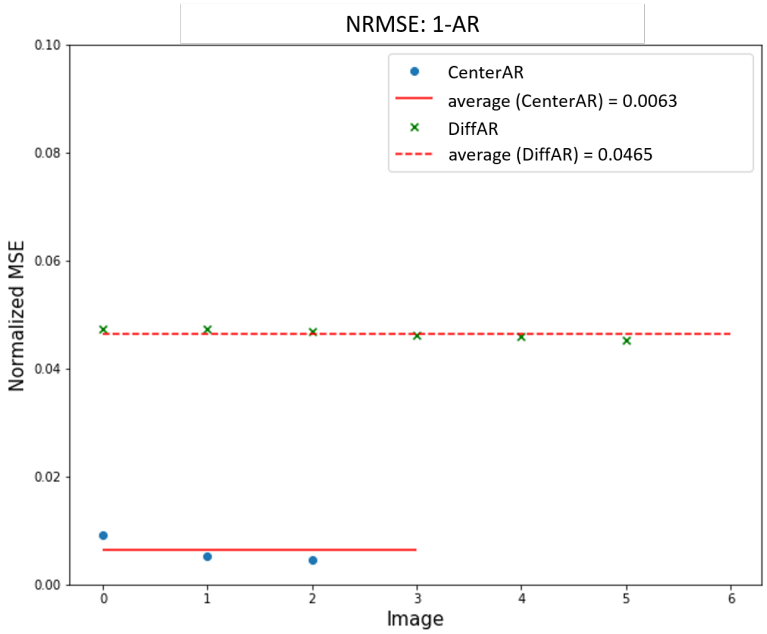
Between the two single-hemisphere experiments, linear AIA provided generally better performance than log10 across both metrics and all tests. However, this margin is fairly small.

4.2 1-AR Experiments

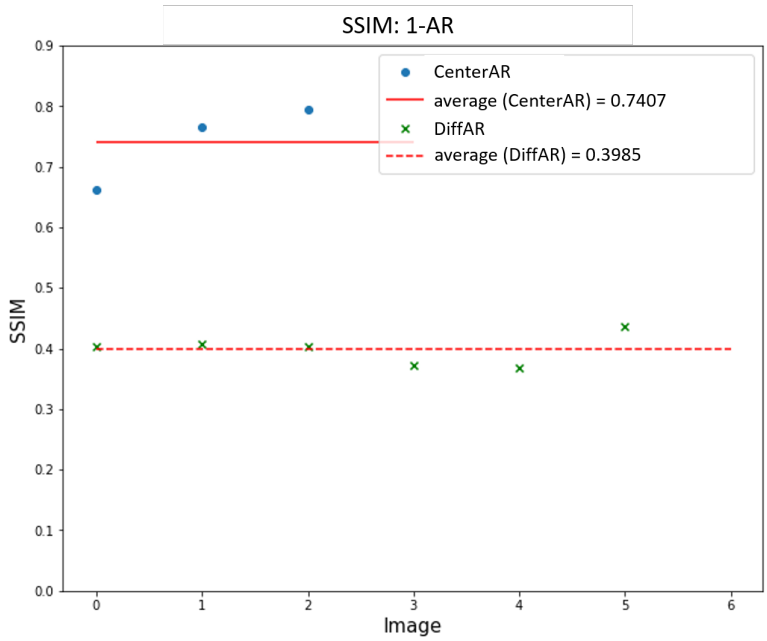
For the model trained on AR1 (NOAA AR 1131), the NRMSE and SSIM are shown in Figure 6. For figures showing NRMSE and SSIM for a given experiment, we produce scatter plots that show the metrics for each image in the test datasets (blue dots for SameAR, green x's for DiffAR), as well as lines showing the overall average values for each of the test datasets. We are looking at not only the average values themselves, but at the size of the gap between the average values for SameAR and DiffAR–the closer the lines representing these averages are, the better the model is generalizing and the less it is overfitting to the training data. The exact sizes of these gaps are listed in Table 5.

We see significantly lower error and higher similarity for the three images from the same AR as in training (NRMSE of 0.0063 and SSIM of 0.7407), while the images from other ARs elicited much worse performance (NRMSE of 0.0465 and SSIM of 0.3985).

Comparisons between the target and predicted magnetograms for small-scale experiment 1 (using CenterAR test images, chosen from the same AR used in training and central with respect to time) are shown in Figure 7. Each predicted image was constructed by the network that was trained on that same AR, and the test images are at the center time as in Figure 2. In Figure 7 we see that, as expected, the networks perform very well on images that are extremely similar to



(a) NRMSE for small-scale experiment

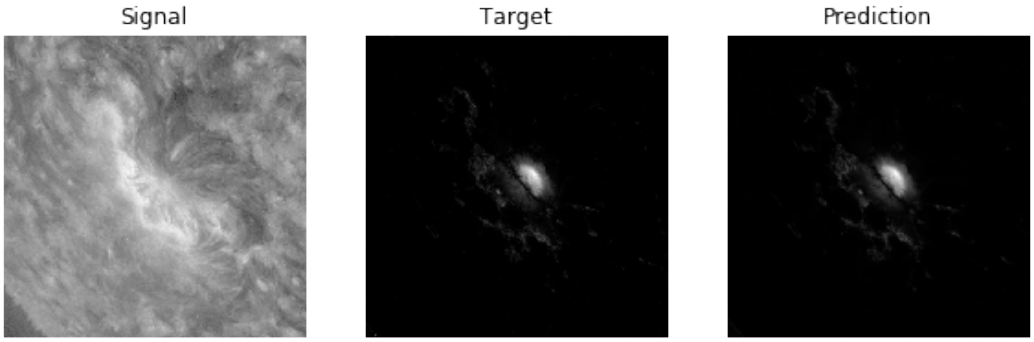


(b) SSIM for small-scale experiment

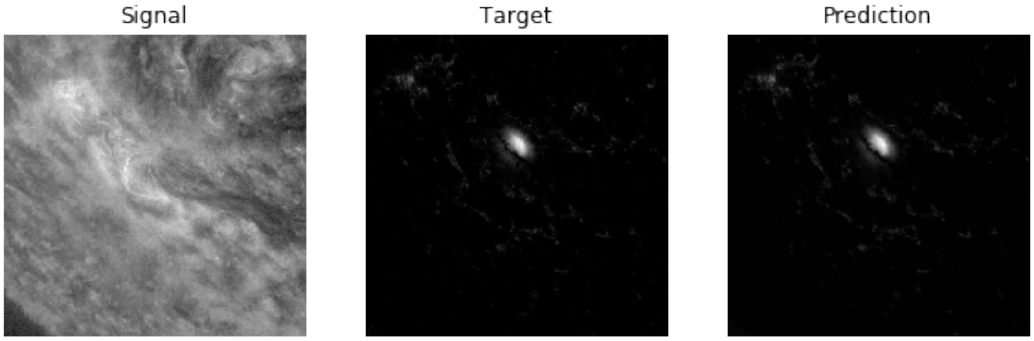
Figure 6: NRMSE and SSIM results for model trained on 30 images from NOAA AR 1131 and tested on 3 images from the same AR and 6 images from two different ARs.

the training data.

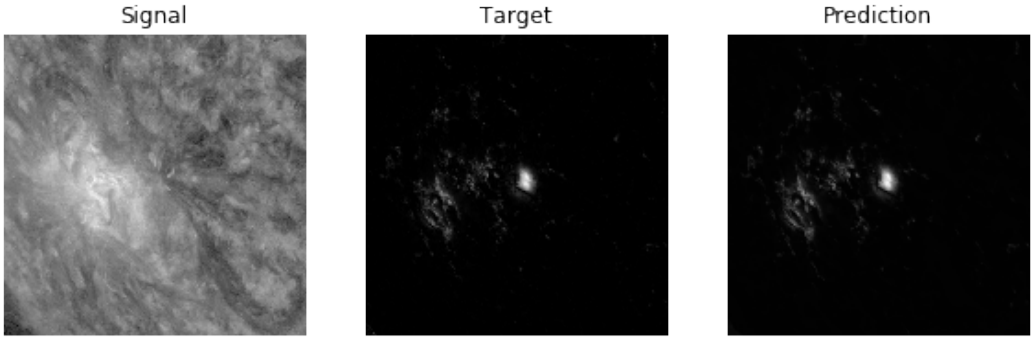
Each of the three models (trained on AR1, AR2 (NOAA AR 1140), or AR3 (NOAA AR 1147)) was also tested on images from the other two active regions. A comparison between target magnetograms from AR2 and AR3 and predicted magnetograms from the model trained on AR1 are shown in Figure 8. We see that the network performs very poorly on images from AR2 and AR3. Even though both AR2and AR3 have very similar structure to AR1, the network struggles to correctly predict this. This model has overfitted to to AR1 and has been unable to generalize the modality transfer from so little training data, as expected.



(a) AR1

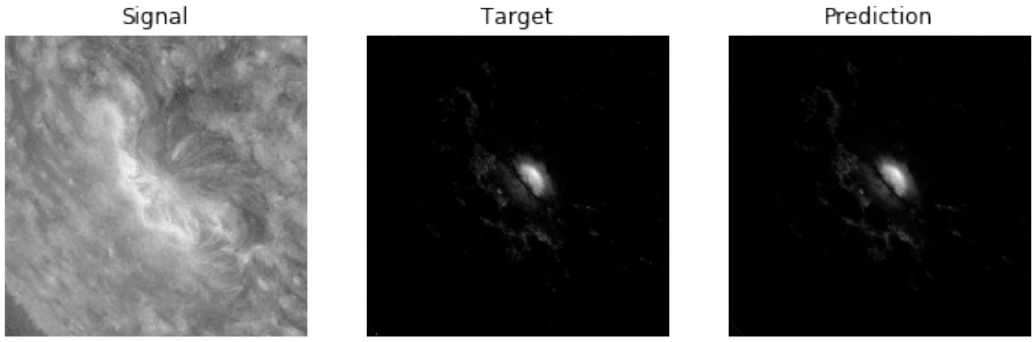


(b) AR2

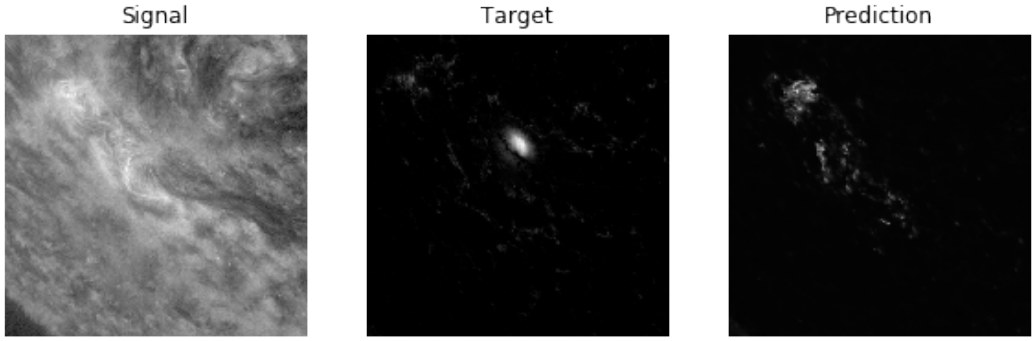


(c) AR3

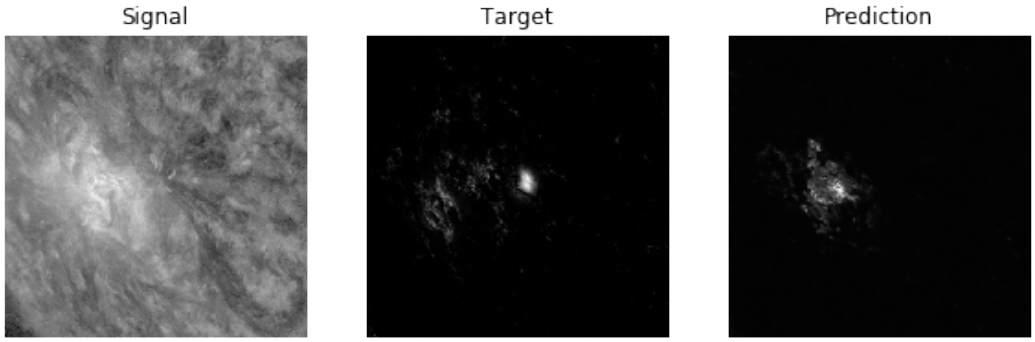
Figure 7: Predicted unsigned magnetograms for small-scale experiments for CenterAR. The prediction in each row was made by the model trained on the same AR that the test image is sourced from.



(a) Target image from AR1



(b) Target image from AR2



(c) Target image from AR3

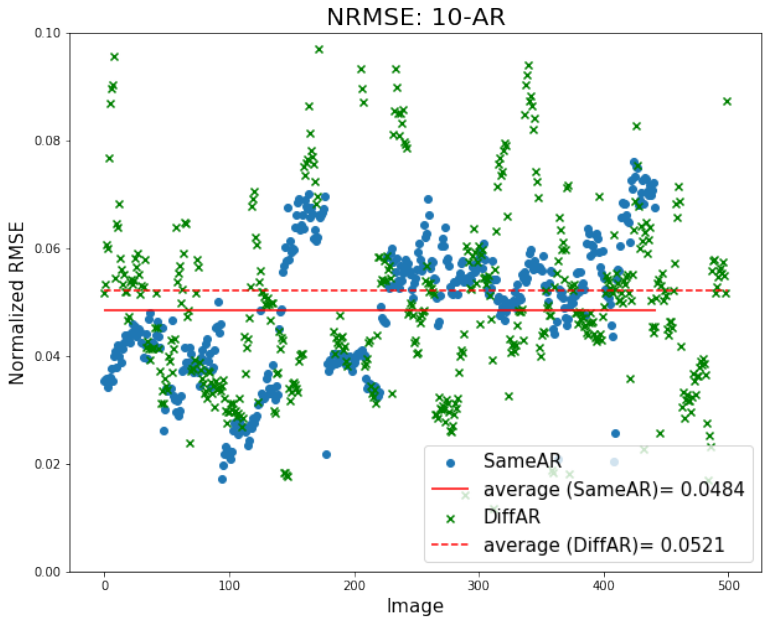
Figure 8: Predictions from the model trained only on AR1, with test images from multiple ARs.

4.3 10-AR Experiments

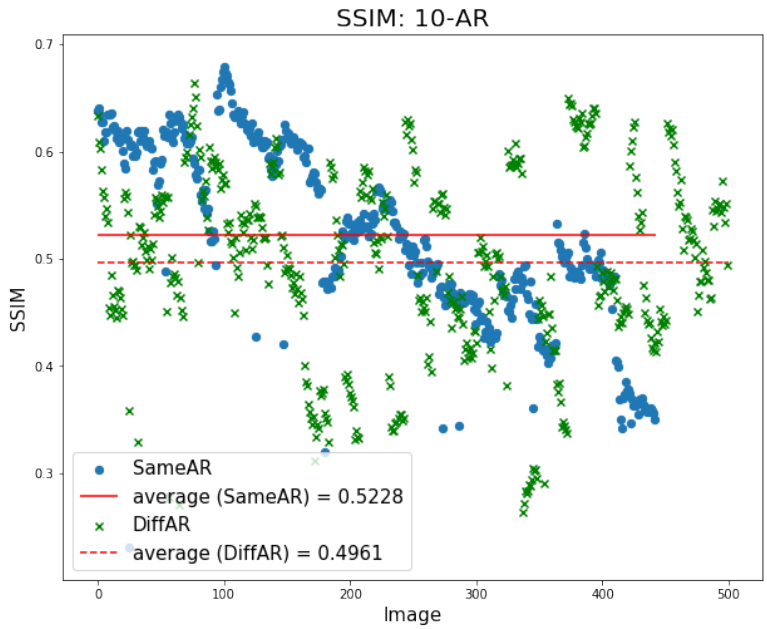
Figure 9 shows error and similarity for the model trained on ~ 5000 images from 10 active regions. We see average NRMSEs for SameAR and DiffAR of 0.0484 and 0.0521, respectively. Compared to the small-scale experiment NRMSE (0.0063 and 0.0465 for CenterAR and DiffAR, respectively), we see significantly higher error for SameAR. However, as the exceptionally low NRMSE for small-scale CenterAR results was due to overfitting, a higher error with more ARs is expected. We see higher error than in the small-scale test for DiffAR as well, although the increase compared to the small-scale experiments for DiffAR is smaller than for SameAR.

More notable is the fact that the gap between SameAR and DiffAR NRMSE has decreased significantly in the 10-AR experiment compared to 1-AR. This is a good sign that indicates we are not overfitting so severely to the ARs used in training, and are obtaining a model that is closer to generalizing the modality transfer. We do still see a gap between SameAR and DiffAR results–ideally we would like to get that gap as small as possible in order to demonstrate a model that can generalize.

We conclude similarly for the structural similarity– SSIM for SameAR has worsened (down to 0.5228 from 0.7407 in the small-scale SameAR). For DiffAR, however, we see an improvement over the small-scale DiffAR results (up to 0.4961 from 0.3985). The gap between the SSIM for the two test datasets has decreased



(a) NRMSE



(b) SSIM

Figure 9: NRMSE and SSIM for 10-AR experiment

as well.

Comparisons between target and predicted images from the 10-AR experiment are shown in Figure 10 for SameAR and Figure 11 for DiffAR. The SameAR results in Figure 10 confirm that the overall prediction quality is not as good as the results for the small-scale SameAR results (Figure 7), but the global structure of the active region is being predicted fairly well. The network seems to do well with bright areas of high flux, but struggles with small details and fine structures, giving the predictions in Figure 10 a slightly blurry appearance compared to the target images.

The DiffAR results in Figure 11 are a clear improvement over the 1-AR DiffAR results (Figure 8). The network is picking up the basic shape and structure of the magnetic activity, although there is a decrease in quality from the SameAR examples. This supports what we saw looking at the quantitative metrics–although we have improved performance on new ARs by increasing the training data volume, more is needed to help close the performance gap between SameAR and DiffAR predictions.

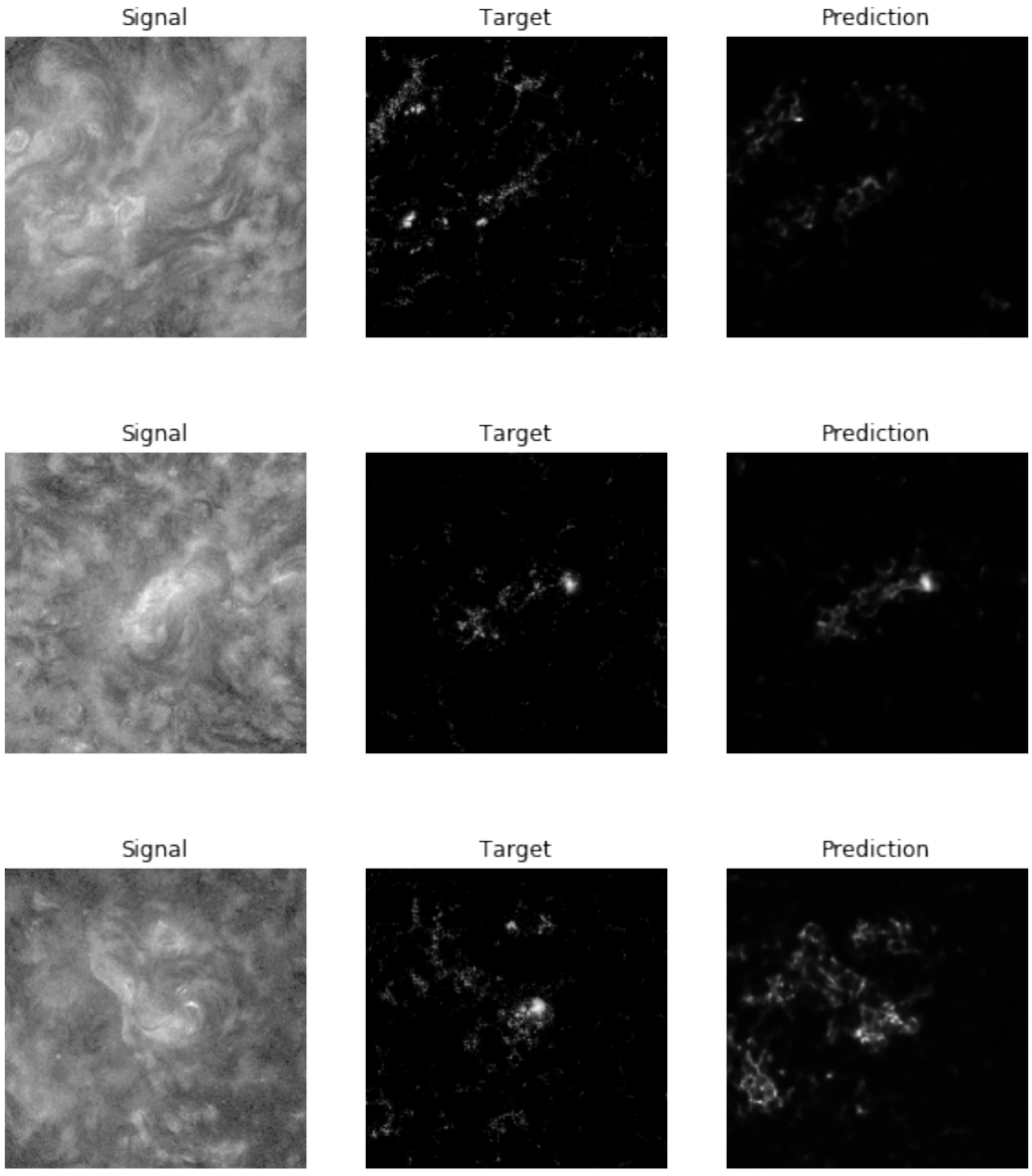


Figure 10: Predicted unsigned magnetograms for 10-AR experiment for SameAR

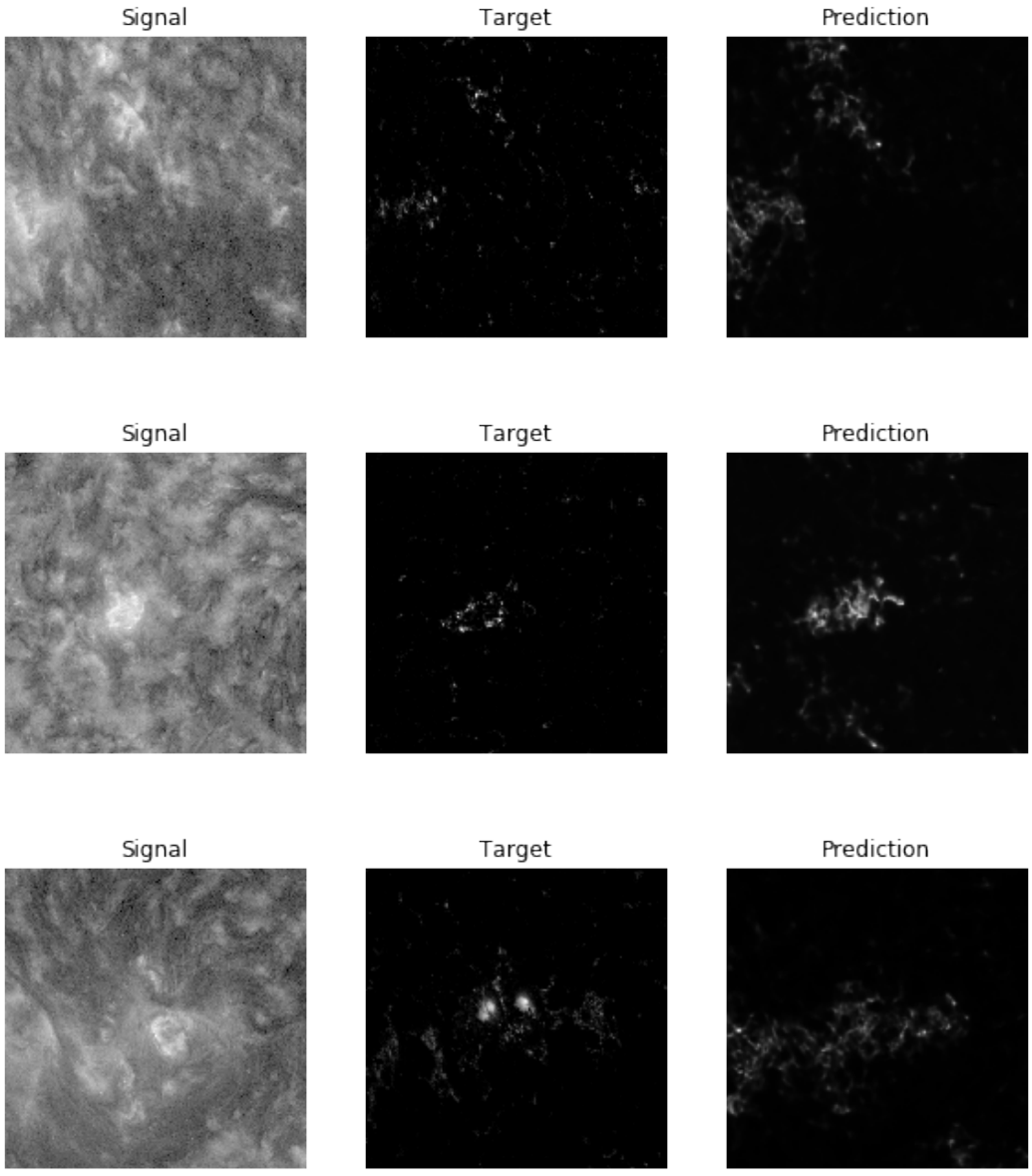
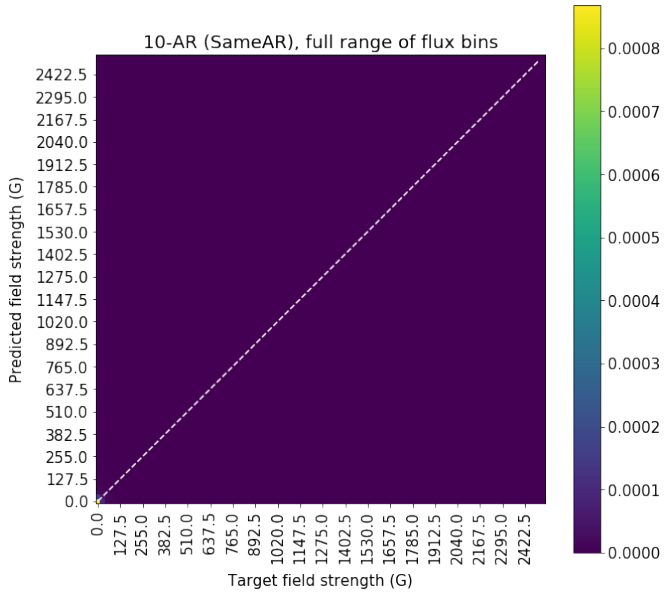


Figure 11: Predicted unsigned magnetograms for 10-AR experiment for DiffAR

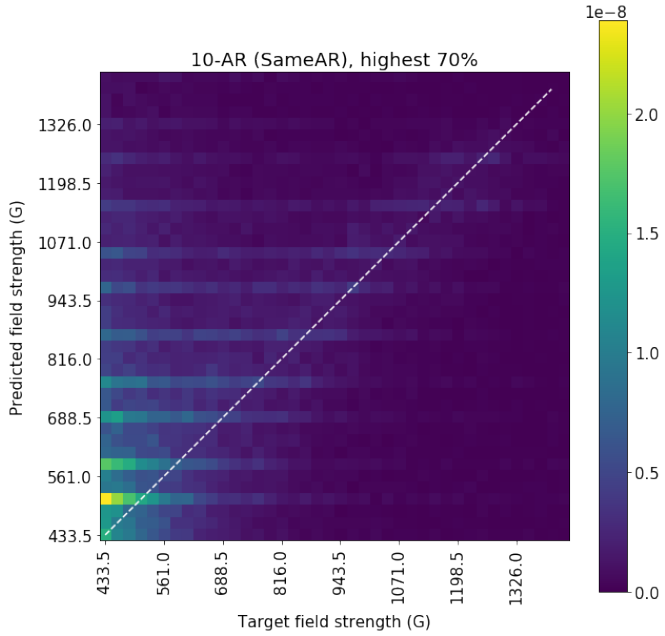
Along with measures of error and similarity, it is also of interest to evaluate how well the network can predict the strength of an active region's magnetic field via a 2D histogram or heatmap of predicted magnetic field strength versus target magnetic field strength, as described in Section 3.4. We include a dashed reference line with a slope of 1 on each of these heatmap images–for a very high accuracy prediction where predicted field strength is equal to target field strength everywhere, the bright areas of the heatmap would fall only along this line.

Figure 12(a) shows the target versus predicted flux heatmap for the 10AR SameAR test, covering the full range of flux bins ([0G, 2550G]). We cannot make out any meaningful features in Figure 12(a)–as a majority of the pixels in the magnetograms are usually background quiet sun with values close to 0 G, the high density of counts in bins corresponding to zero flux tends to overwhelm any other characteristics in the heatmap. For this reason, we will look at a subset of the flux bins instead of the full range. Figure 12(b) is the same heatmap, but only bins covering the highest 70% of flux values present in this test dataset are included. By excluding the bins corresponding to the lowest flux values, we can now see much more clearly that there are interesting structures present in the heatmap.

We display heatmaps showing the highest 70% and highest 50% of flux values for the experiments predicting unsigned magnetograms, which are the experiments where localization was good enough that the heatmaps could be interpreted mean-



(a) SameAR full heatmap



(b) SameAR heatmap, highest 70% of flux bins

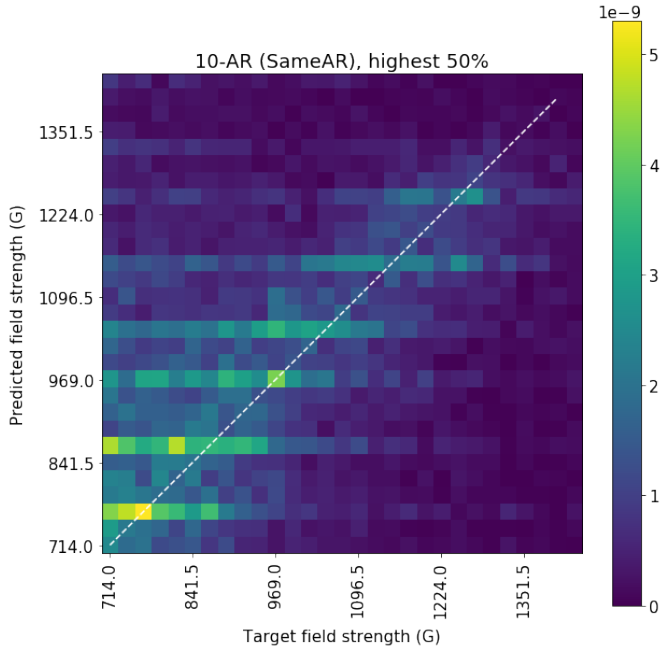
Figure 12: Heatmaps for 10AR SameAR test, showing why we limit these heatmaps to show only a subset of the full range of flux bins.

ingfully.

The color values in the heatmaps represent the probability density at that bin (number of counts in the bin / total counts / bin area). In Figure 12(b) we can see scattered bright areas towards the left side of the heatmap along the y-axis, caused by the network predicting higher flux values where there is low flux in the target image– this corresponds to the blurred appearance of the example images in Figure 10. Figure 12(b) also shows brighter areas following the dashed reference line. This is a good indication that, when the network does place areas of higher flux at the correct locations, the predicted flux value is accurate. This supports the idea that our methods have the potential to predict a wider range of flux values than those used in [8] with accuracy, though the fact that we must rely on detail images to see this indicates a need for improvement in the pixel-wise accuracy of the predictions.

Figure 13 shows a detail view of the highest 50% of flux values present in the SameAR test dataset. We can now see bright areas following the dashed line very clearly, which indicates that predicted flux corresponds well with target flux at these higher flux values.

Figure 14 shows heatmaps for the 10-AR DiffAR experiment, with the highest 70% of flux bins shown in Figure 14(a) and the highest 50% shown in Figure 14(b). This is an example of a test where localization of predicted activity is poor enough that we cannot gather meaningful information about the relationship between target and predicted flux values from these heatmaps–the only clear structure in

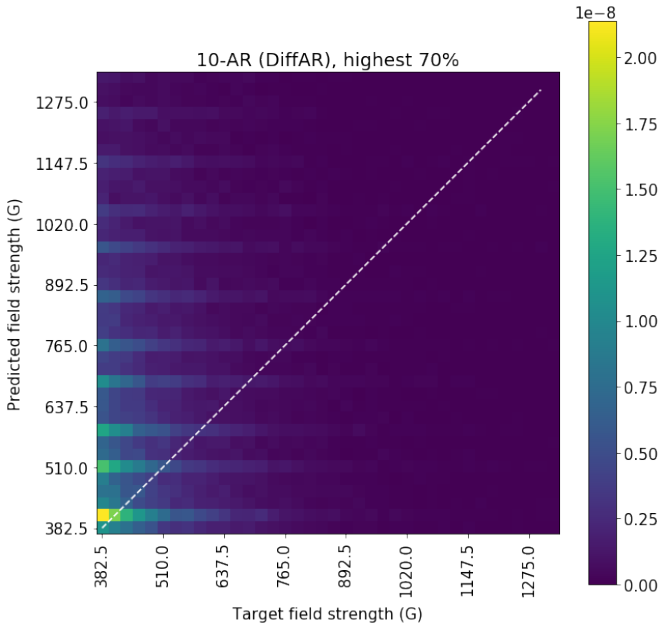


(a) SameAR heatmap, highest 50% of flux bins

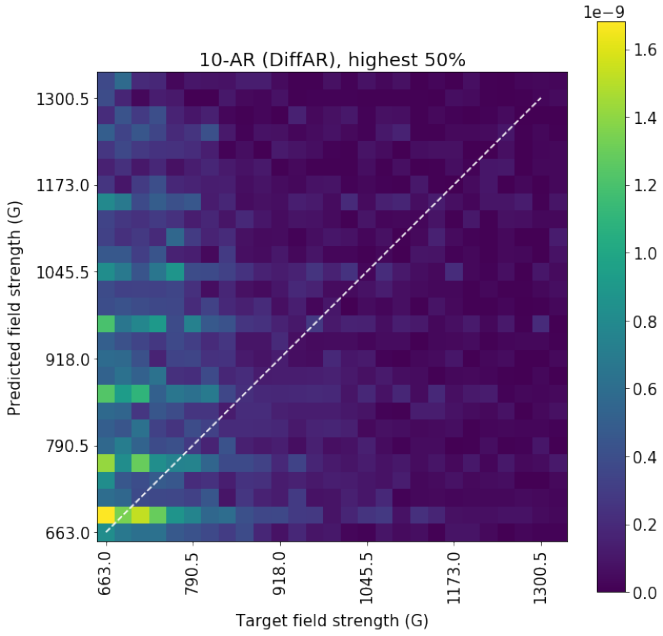
Figure 13: Detail heatmap for 10AR SameAR test

both heatmaps is bright areas along the y-axis, indicating the network predicting bright areas where there should be dark areas. This corresponds with the generally blurrier and less structurally accurate example predictions seen in Figure 11, as well as the performance decrease in DiffAR metrics vs. SameAR as seen in Figure 9.

Figure 15 shows target versus predicted total unsigned flux, B_{abs}, for the 10AR experiment. Each point represents an image in the relevant test dataset, and a dashed line with a slope of 1 is included for reference as with the heatmaps– the closer the points in the scatter plot are to following the line, the better the correspondence between target and predicted B_{abs}. If most of the points are

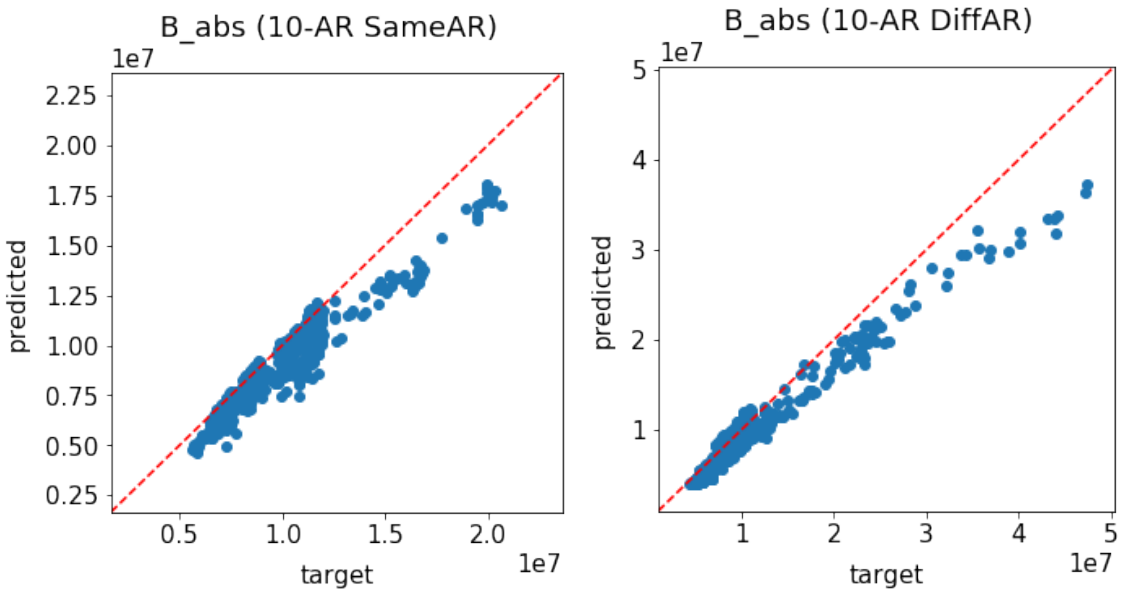


(a) DiffAR heatmap, highest 70% of flux bins only



(b) DiffAR heatmap, highest 50% of flux bins only

Figure 14: Heatmaps for 10AR DiffAR test



(a) (b)

Figure 15: Target vs. predicted total absolute flux (10-AR experiment)

placed under the reference line, the model is under-predicting the total flux; if points are mostly above the line it is over-predicting the total flux. For both SameAR (Figure 15(a)) and DiffAR (Figure 15(b)), total absolute flux is being predicted fairly well, although there is a slight tendency to under-predict which worsens for images with large B_{abs}. As the majority of the images have lower B_{abs} values, it's likely this under-prediction did not contribute enough to the counts in the heatmap for this tendency to be visible in Figures 12 and 13.

4.4 52-AR Experiments, Unsigned Magnetograms

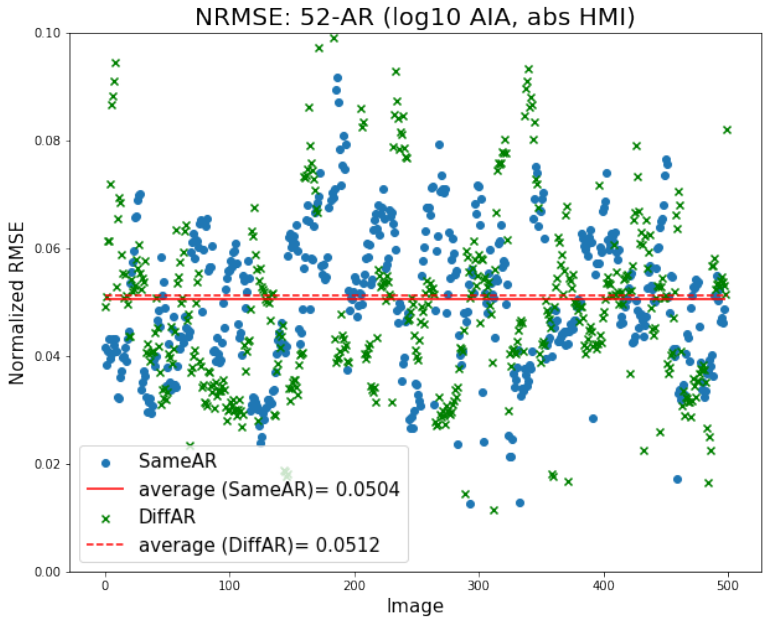
This section details the results of the two experiments using data drawn from 52 active regions and using unsigned HMI data as target images. The first used log10 scaled AIA data as signal, as the previous experiments did, and the other used linearly scaled AIA data as signal.

4.4.1 Log10 AIA

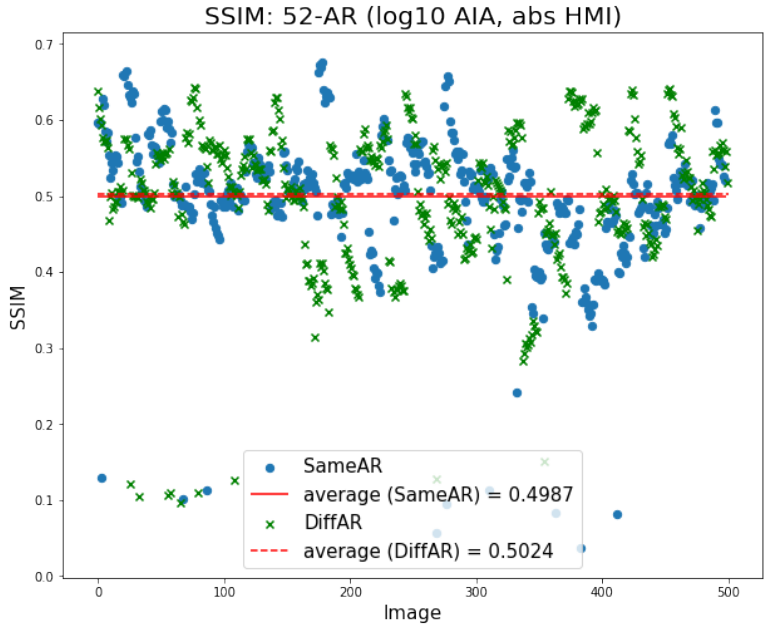
Figure 16 shows the NRMSE and SSIM for predictions from the model trained on data from 52 active regions, with log10 scaled AIA data as signal and absolute value HMI as target. NRMSE for SameAR is higher than we saw with 10 ARs (up to 0.0504 compared to 0.0484), but has decreased for DiffAR (down to 0.0512 compared to 0.0521). The gap between NRMSE for SameAR and DiffAR in the 52-AR experiment has decreased significantly, although the NRMSE for DiffAR is still slightly higher than SameAR.

We see similar trends in SSIM–compared to the 10-AR experiment, SSIM has decreased for SameAR (down to 0.4987 from 0.5228), but increased for DiffAR (0.5024 versus 0.4961). Comparing between the two test datasets for the 52-AR log10/abs experiment, SSIM for DiffAR is slightly higher than for SameAR, but the difference between them is small.

The lack of a significant performance gap between SameAR and DiffAR indicates that this model is not overfit to the training data and the network has



(a) NRMSE



(b) SSIM

Figure 16: NRMSE and SSIM for 52-AR abs/log10 experiment.

generalized the modality transfer. The average NRMSE is close to what we saw with the small-scale (1-AR) DiffAR experiments, indicating that localization of magnetic activity still is not perfect on a pixel-wise level. The SSIM, however, has improved significantly, indicating prediction of the overall structure of the predicted magnetic field is more accurate even if pixel-to-pixel accuracy is not. While there is still room for improvement in overall prediction quality, it appears that this 52-AR experiment provides a suitable data volume, both number of images and number of ARs, to avoid the overfitting problems seen in the smaller experiments.

Comparisons between target and predicted images from the model trained on $\sim$ 5000 images from 52 active regions are shown in Figure 17 for SameAR and Figure 18 for DiffAR. Qualitatively, these are on par with those from the 10-AR SameAR results in Figure 10. The predictions in Figure 17 do appear blurrier and overall less accurate than the 1-AR results in Figure 7–this is not surprising as a model trained on a larger number of ARs will likely not predict as accurately as one that is overfit to a single AR and with test images chosen to boost accuracy. Similar to the blurriness noted in the discussion of the 10-AR results, here we see that many of the small details and fainter structures in the target magnetograms are not present in the prediction images, while brighter and stronger areas of activity are predicted more accurately.

Comparing Figure 18 qualitatively with Figure 17, predictions for DiffAR im-

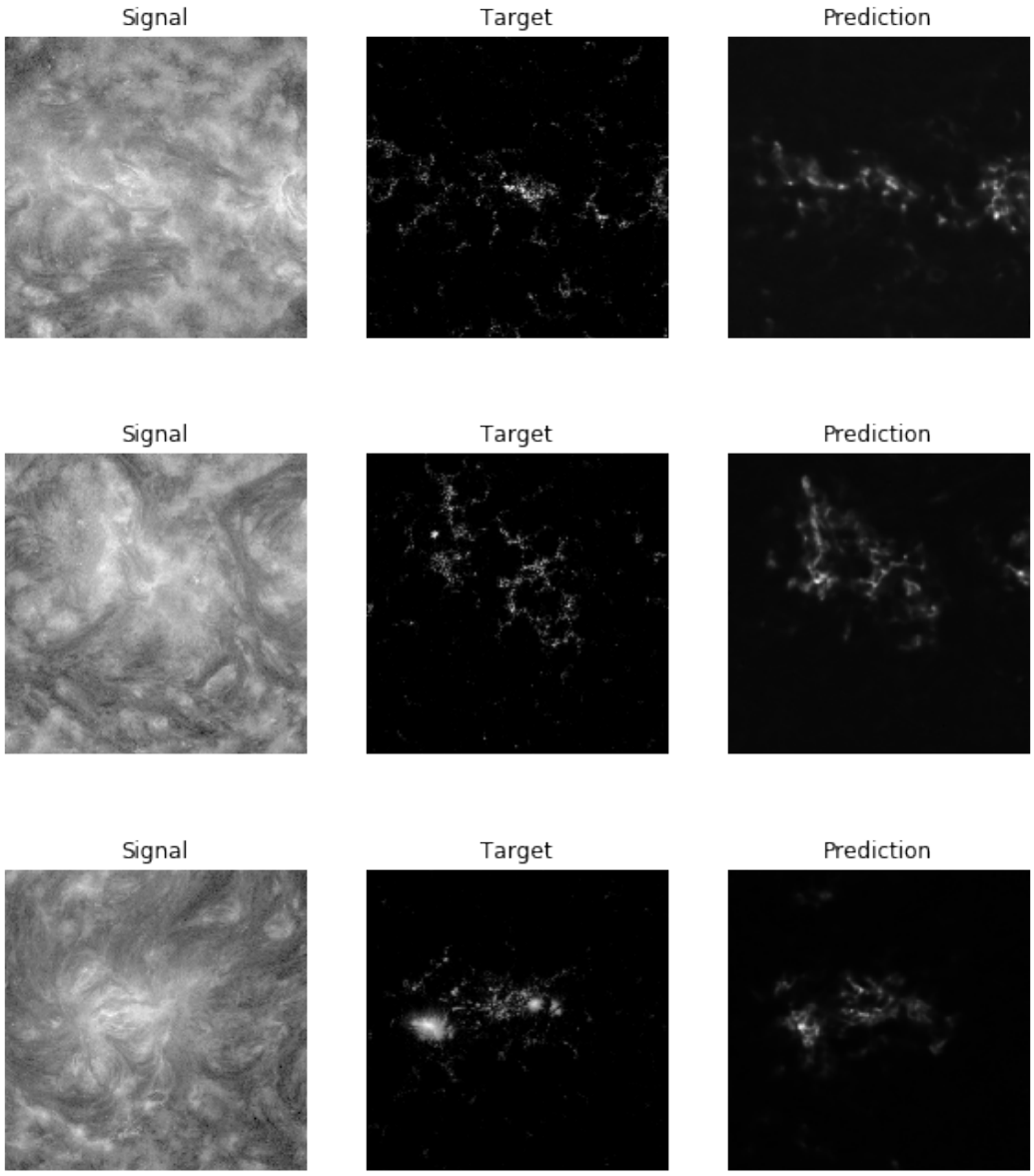


Figure 17: Predicted unsigned magnetograms for the log10/abs 52-AR experiment for SameAR (test data from the same ARs as training).

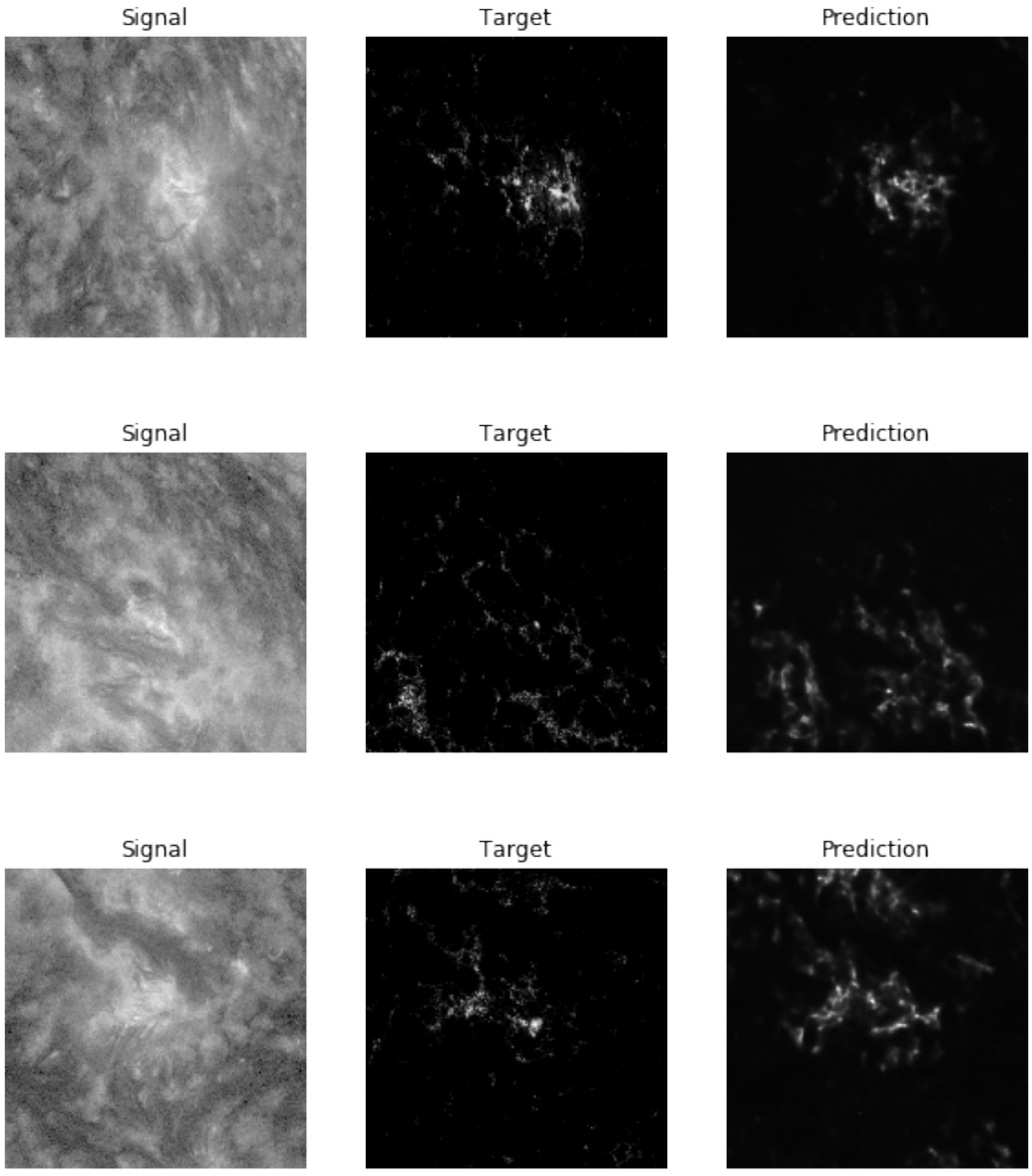
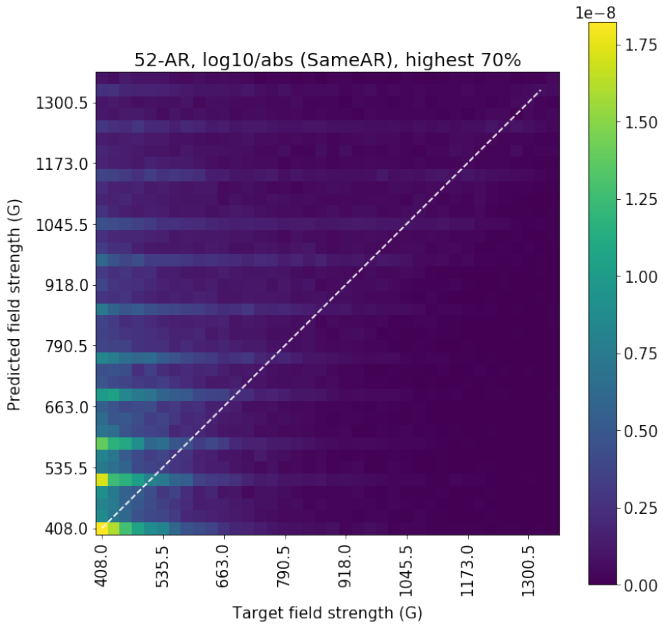


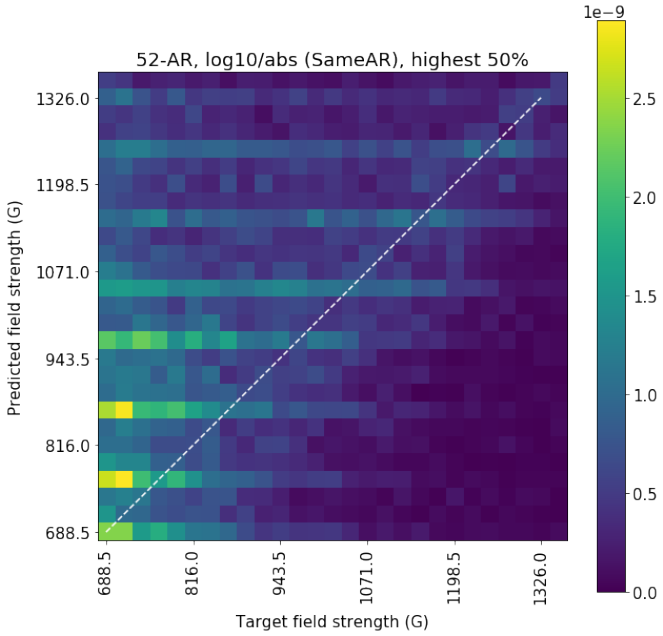
Figure 18: Predicted unsigned magnetograms for the log10/abs 52-AR experiment for DiffAR (test data from new ARs).

ages look very similar to those from SameAR images. It would be difficult to pick out which predictions were for which test if one did not know beforehand which dataset the target images came from. Compared to the 10-AR DiffAR results in Figure 11, the predictions shown in Figure 18 have improved–the network is capturing the overall shape of the magnetic activity and placing stronger flux areas accurately just like with SameAR images, which the 10-AR model struggled to do when making DiffAR predictions. We again see a mix of quality in these predictions and some ARs look better than others, but they are an improvement over the 10-AR predictions in Figure 11 and dramatically better than the small-scale DiffAR predictions in Figure 8.

Figure 19 shows heatmaps for the SameAR test. The structure present in these maps is less clear than those seen in the 10-AR SameAR heatmaps. We still see activity along the y-axis due to blurring and activity along the dashed line due to good predicted flux strength, but these groups of activity have bled together and are less distinct than they were for the 10-AR test. In Figure 19(b) especially, this leads to the entire upper left half of the image showing notable activity, while the lower right shows very little activity. If there was little correlation between target and predicted flux levels, or if localization accuracy was much worse, we would expect very little activity along the dashed line at all (similar to the 10-AR DiffAR heatmap in Figure 14(b)). Although localization accuracy is low enough that it is difficult to make more concrete conclusions, the fact that there is still



(a) SameAR heatmap, highest 70% of flux bins only



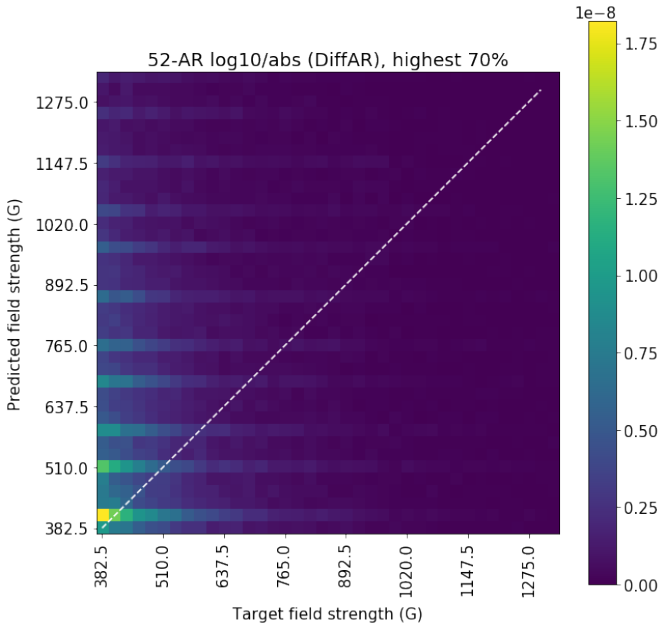
(b) SameAR heatmap, highest 50% of flux bins only

Figure 19: Heatmaps for 52AR log10/abs SameAR experiment.

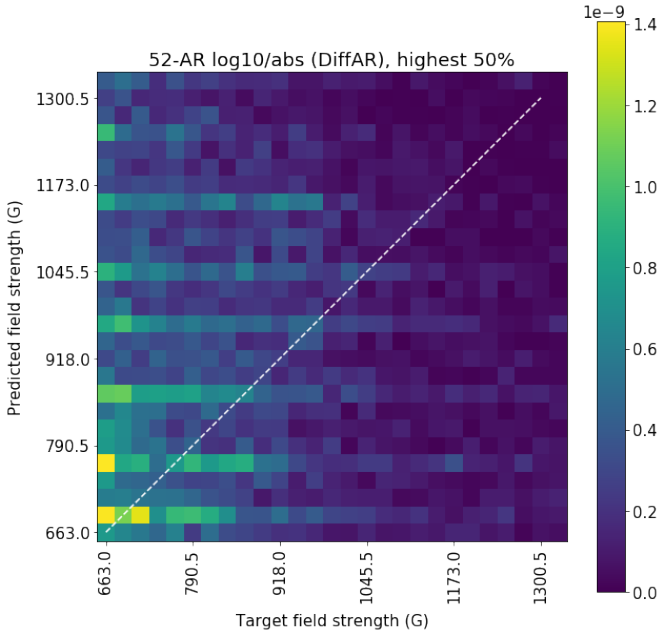
notable activity along this line indicates reasonably accurate correlation between target and predicted flux levels.

Heatmaps for the DiffAR test are shown in Figure 20. The characteristics of these heatmaps correspond to the general performance of this test relative to the SameAR test and the 10-AR tests–there is less activity along the dashed line than we saw for the 52-AR log10/abs SameAR test, but more than we saw for the 10-AR DiffAR test. The difference between heatmap behavior for this test (52 log10/abs DiffAR) and its SameAR counterpart indicates that there is still a slight performance gap between SameAR and DiffAR performance at this data volume, although the gap has nearly disappeared in the quantitative metrics.

Figure 21 shows target versus predicted B_{abs} for this experiment. As with the 10-AR experiment, we again see the points sit close to the line for lower values of target B_{abs}, and then under-prediction worsens for higher values. This effect is more pronounced in the DiffAR test (Figure 21(b)), but this is likely because there are higher values of B_{abs} present in that test dataset (note the difference in values on the axes compared to Figure 21(a)).

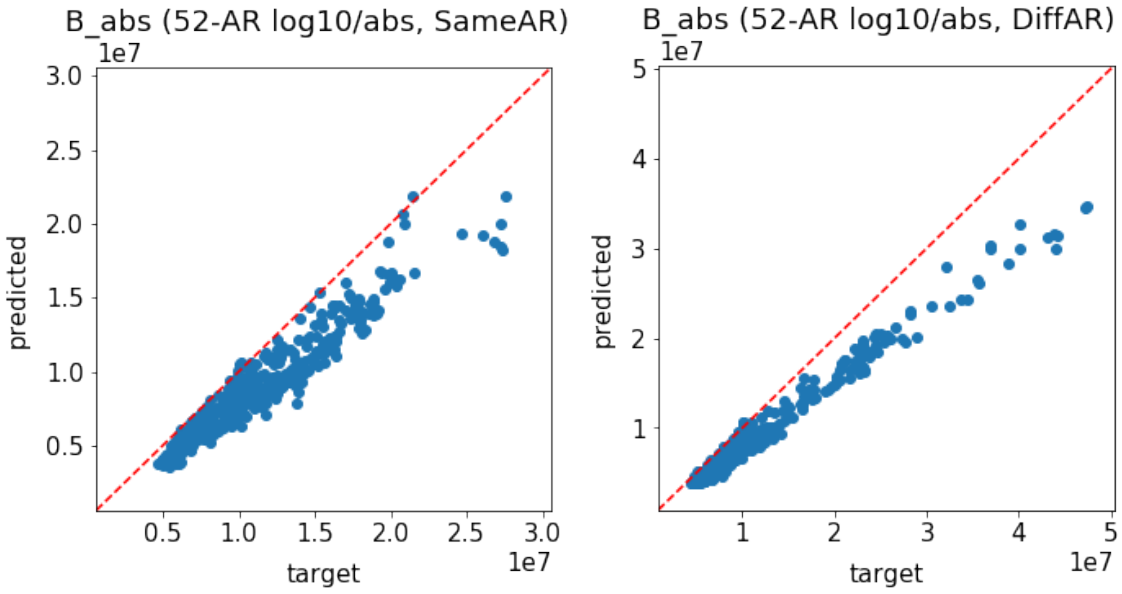


(a) DiffAR heatmap, highest 70% of flux bins only



(b) DiffAR heatmap, highest 50% of flux bins only

Figure 20: Heatmaps for 52AR log10/abs DiffAR experiment.



(a) (b)

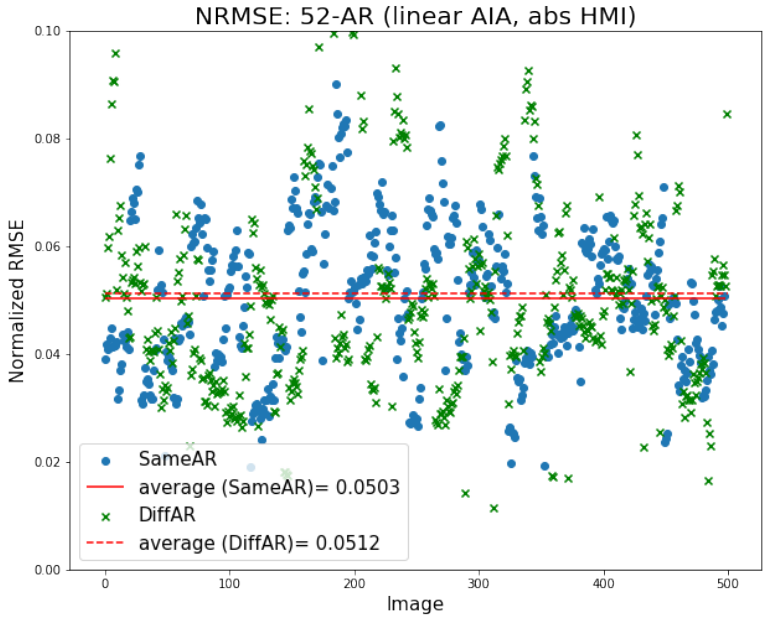
Figure 21: Target vs. predicted total absolute flux (52-AR log10 AIA/abs HMI experiment)

4.4.2 Linear AIA

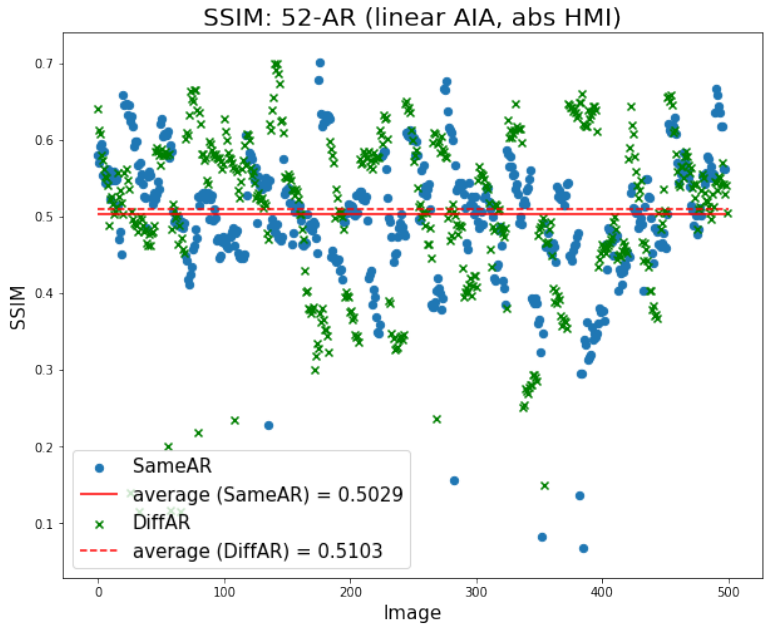
Error and similarity for the experiment with absolute value HMI and linearly scaled AIA data are shown in Figure 22. Changing the scaling of the AIA data from log10 to linear appears to have had very little effect on network performance. NRMSE for both SameAR and DiffAR experiments are nearly identical to those seen in Section 4.4.1. There is a slight improvement in SSIM for both SameAR and DiffAR–it has increased to 0.5029 and 0.5103 for SameAR and DiffAR, respectively, compared to 0.4987 and 0.5024 in the log10 experiment. It is interesting that DiffAR results are slightly better than SameAR in this case, but only by a small margin that is likely not significant.

Figure 23 shows examples of predicted magnetograms for the SameAR experiment and Figure 24 shows predicted magnetograms for the DiffAR experiment. Qualitatively we see these predictions share many of the same characteristics seen in Figures 17 and 18 above, validating the strong similarity in network performance noted in the NRMSE and SSIM results.

Figure 25 shows the target versus predicted field strength heatmaps for the SameAR test, and Figure 26 shows the same for the DiffAR test. The characteristics of these heatmaps and any conclusions that can be drawn from them are the same as those seen in the previous 52-AR experiment using log10 scaled AIA data, again due to the similarity in performance between these two experiments.



(a) NRMSE



(b) SSIM

Figure 22: NRMSE and SSIM for 52-AR (linear/abs) experiment.

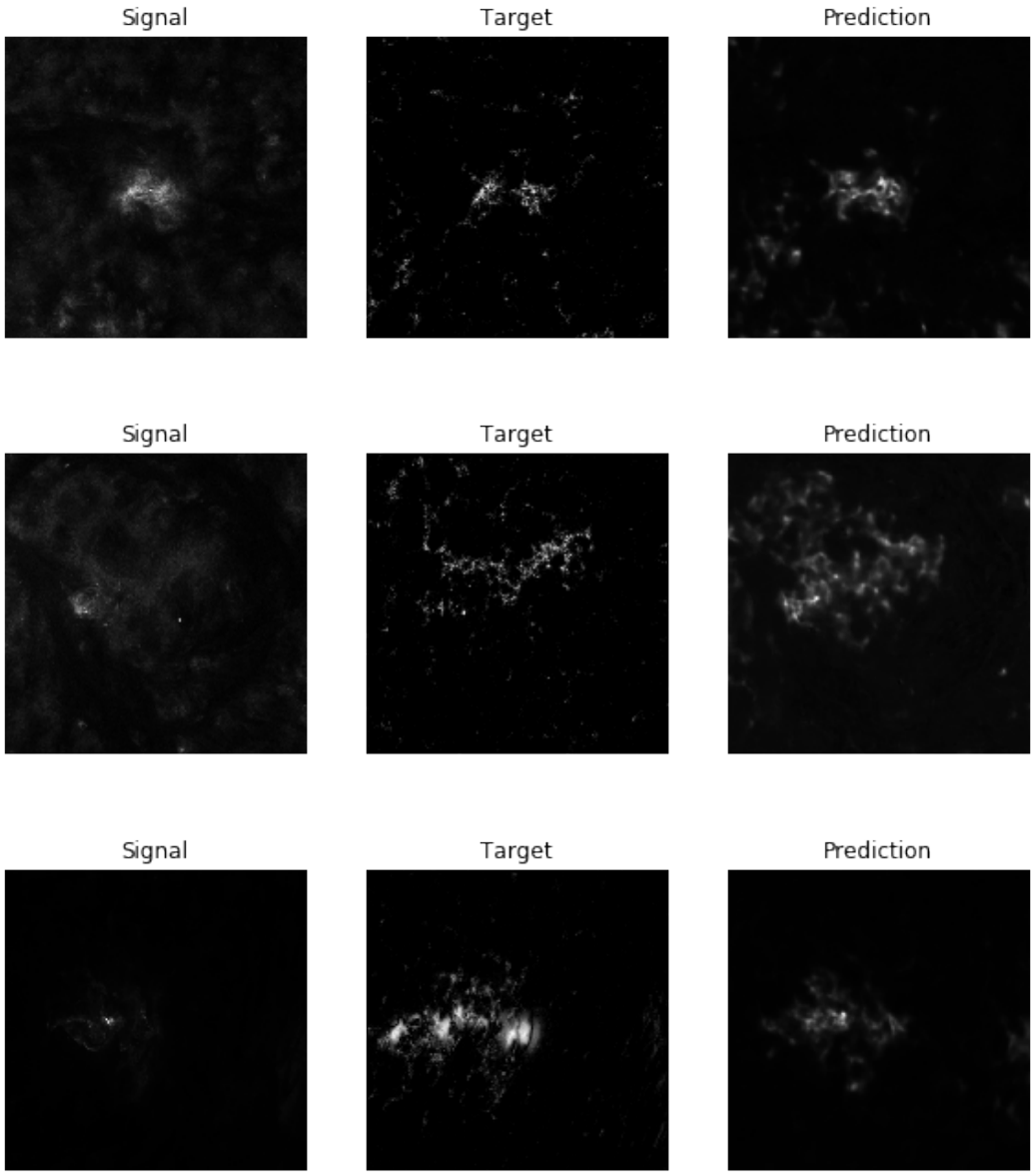


Figure 23: Predicted unsigned magnetograms for the 52-AR linear/abs experiment for SameAR (test data from the same ARs as training).

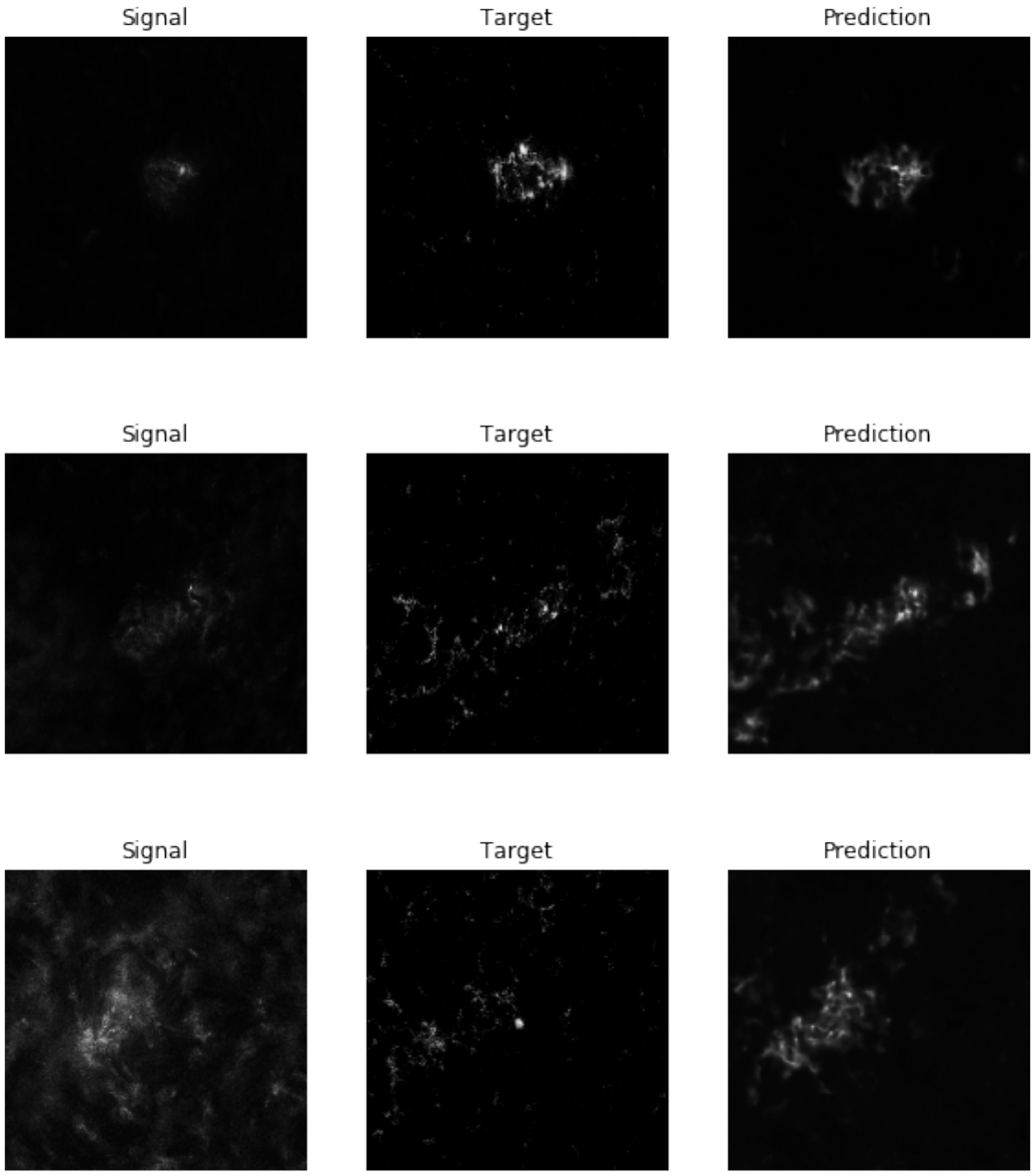
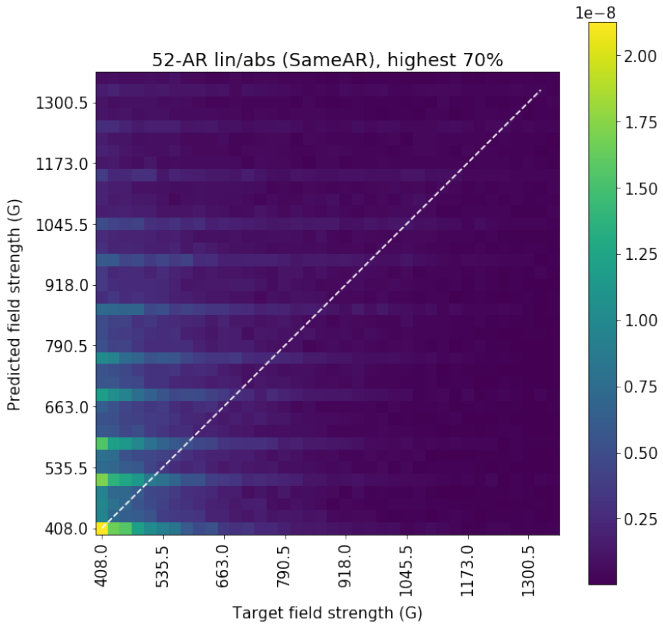
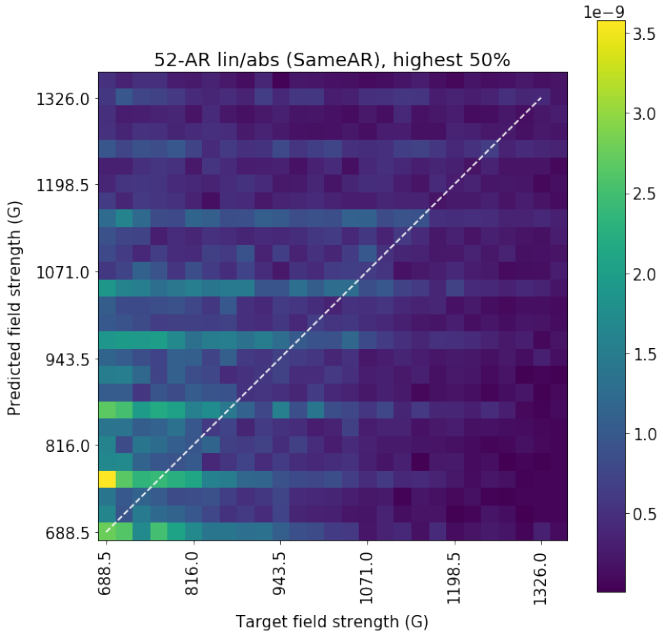


Figure 24: Predicted unsigned magnetograms for the 52-AR linear/abs experiment for DiffAR (test data from new ARs).

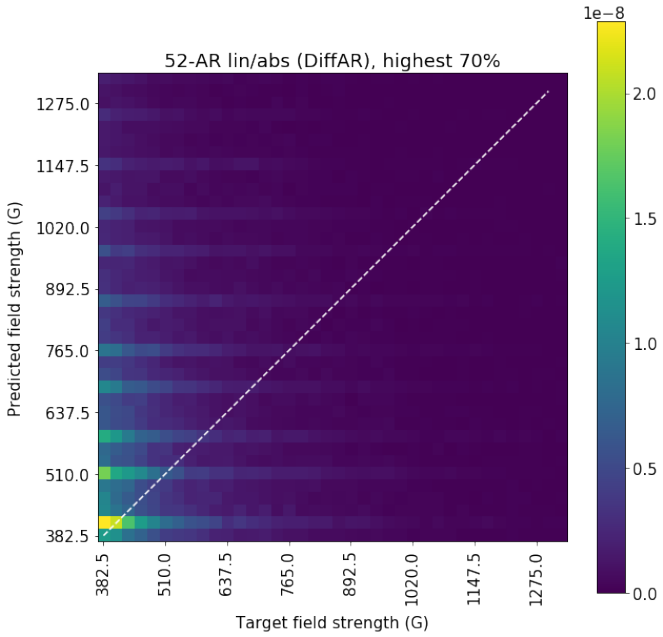


(a) SameAR heatmap, highest 70% of flux bins only

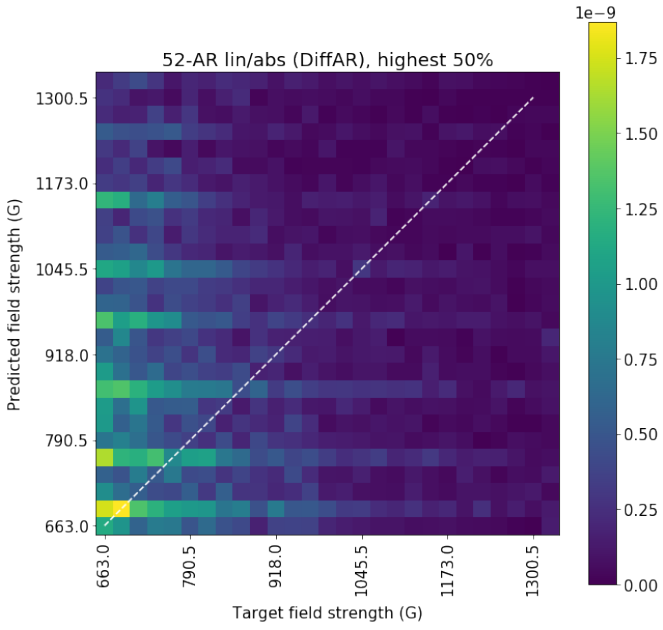


(b) SameAR heatmap, highest 50% of flux bins only

Figure 25: Heatmaps for 52AR (linear/abs) SameAR

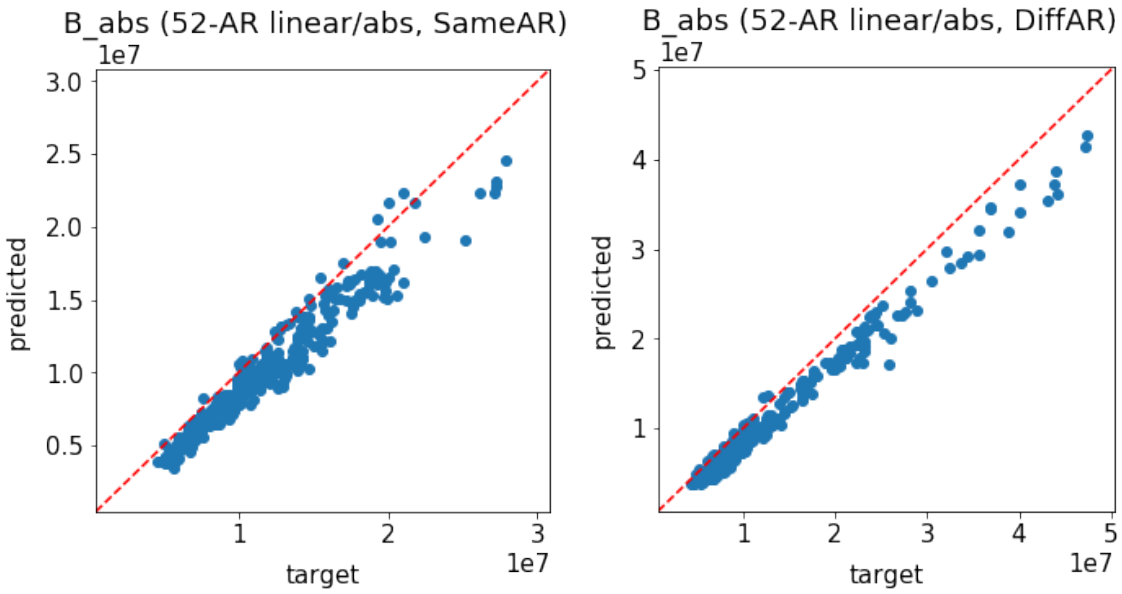


(a) DiffAR heatmap, highest 70% of flux bins only



(b) DiffAR heatmap, highest 50% of flux bins only

Figure 26: Heatmaps for 52AR (linear/abs) DiffAR



(a) (b)

Figure 27: Target vs. predicted total absolute flux (52-AR linear AIA/abs HMI experiment)

The exception is Figure 26(b), which shows less activity along the reference line than its log10 counterpart (Figure 20(b)).

Figure 27 shows target versus predicted B_{abs} for this experiment. These show very similar trends to those we saw in the B_{abs} comparisons for the log10 AIA, absolute value HMI experiment- we see slight under-prediction that is more pronounced the higher the target B_{abs} value is, and the presence of larger B_{abs} in the DiffAR test dataset exaggerates this effect. Both of the 52-AR experiments using absolute value HMI data show this mild under-prediction of the total flux, but overall the correspondence between target and predicted B_{abs} is good. This supports the idea that these methods can be used to predict a larger range of flux

values accurately.

Overall the results of the two 52-AR experiments using unsigned magnetograms as target are very similar. Comparing heatmaps between the two experiments suggests that log10-scaled AIA data may provide a slight advantage, and some of the example images from the log10 test appear higher in quality. However, as this is not reflected in the quantitative results as well, these results do not suggest a significant advantage to using one method of AIA scaling over the other for signal images when working with unsigned magnetograms as target images.

4.5 52-AR Experiments, Signed Magnetograms

This section covers results for experiments using training data drawn from 52 active regions and using signed magnetogram data. We begin by looking at a limited selection of results from the initial experiments where no attention was paid to the location of active regions, in order to show how these results led to investigating the effects of AR location on the characteristics of predicted magnetograms. Then, results from the experiments with training datasets balanced by hemisphere are discussed in more detail. For all signed magnetogram experiments, pixel-wise localization accuracy was not high enough for analysis of target versus predicted flux heatmaps to be meaningful. Some sample heatmaps for these experiments are included in Appendix I, but they will be omitted from the main discussion. We will instead use plots of target versus predicted total flux values, and can now include the total signed flux, the total positive flux, and the total negative flux to this discussion in addition to the total unsigned flux as was shown for the unsigned experiments.

4.5.1 Initial Experiments with an Unbalanced Training Dataset

Here we briefly show the results of the initial experiment with log10 AIA data as signal and signed HMI data as target, before any balancing of HMI data with respect to hemisphere. Tables 3 and 4 show the NRMSE and SSIM for these experiments ('Unbalanced, log10 AIA' and 'Unbalanced, linear AIA'). We see

a significant improvement in NRMSE over the results of the experiment using unsigned HMI data as target–this trend holds for the other experiments with signed magnetograms as well. This is also the first instance we see of the large SSIM decrease between experiments using signed magnetograms and the previous experiments with absolute value magnetograms.

Figure 28 shows example comparisons between target and predicted images, for unbalanced models using both linear and log10 scaled AIA data as signal. These examples have been chosen to illustrate something interesting that was noted while examining images from these models–these networks would predict structure well for certain images, but would reverse the polarity of the magnetic activity. In fact, these networks would nearly always predict positive activity on the right and negative on the left, regardless of what orientation was present in the target image. Noticing this is what initially led us to investigate the idea of balancing the training datasets based on which hemisphere an active region is located in. Reversing the polarity of a predicted image like this is the most obvious effect of the unbalanced training data–as more training images that followed that polarity convention were presented to the network, it learns to assume that that is true for all images. The other models using signed magnetogram data we discuss below will also occasionally reverse the polarity of predicted magnetic activity, but none do so as uniformly as the initial unbalanced models.

Other than the inaccurate polarities, these examples show similar qualitative

markers to those seen in Section 4.4–overall structure is predicted reasonably well but with a loss of detail and blurry appearance, though the blurring here is more pronounced than many of the unsigned experiments. Both this experiment and the subsequent experiments involving signed magnetograms tend to show qualitatively better predictions (more accurate structure, less blurring) for certain ARs– namely those where the strongest magnetic activity is clustered close together in one area of the image. These also happen to be the types of ARs that show the polarity switch the most clearly, so Figure 28 shows only higher-than-average quality predictions from that experiment. Later sets of example prediction images for other experiments have been chosen to show a range of image qualities and AR structures.

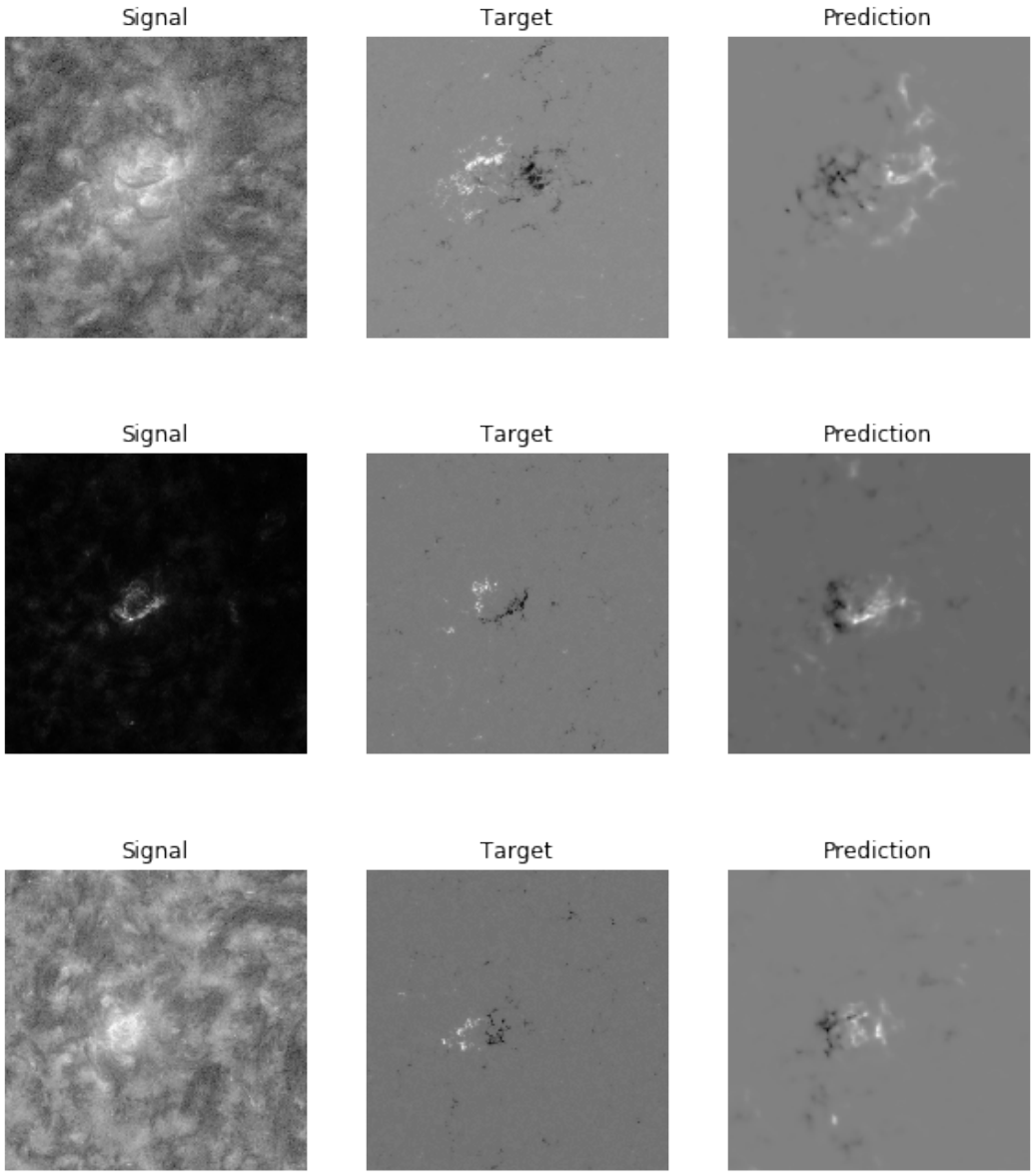


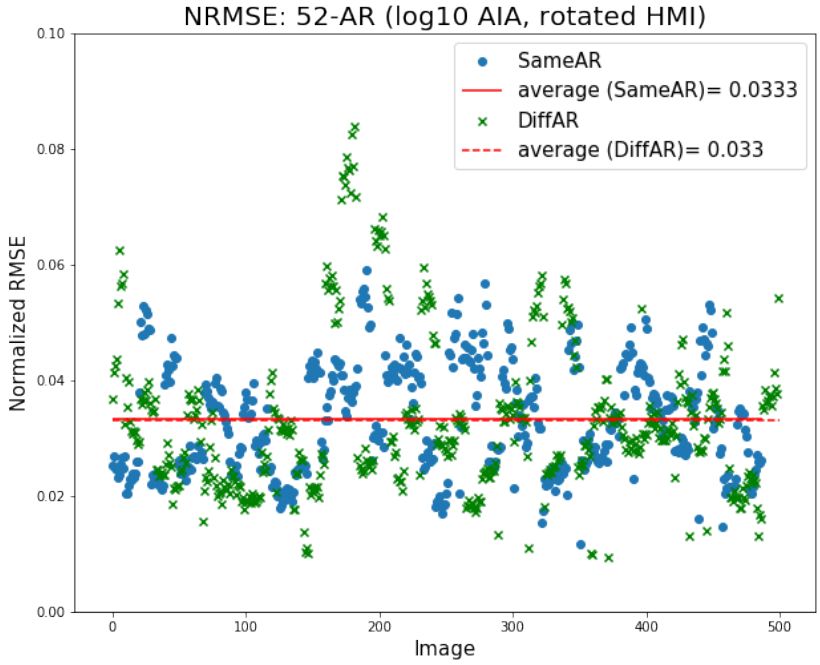
Figure 28: Predictions from the first 52-AR experiments that used signed HMI data not balanced with respect to hemisphere.

4.5.2 Data Augmentation by Rotation

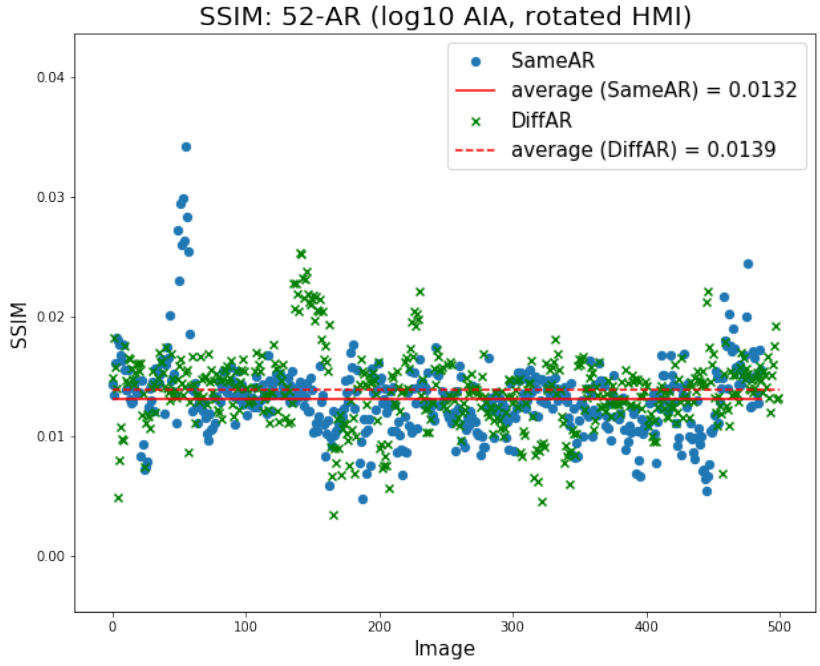
Here we discuss the results of the experiments where data from the unbalanced experiments above was randomly rotated, to artificially remove any bias towards either hemisphere and simulate an equal number of ARs chosen from the north and south hemispheres.

Log10 AIA: Error and similarity for the experiments using randomized rotation of the training data are shown in Figure 29. We see that the NRMSE is very similar to the unbalanced dataset, but slightly higher for DiffAR. Some increase in error is expected–by not making one hemisphere more likely to show up in the data than the other, we have made it harder to predict what activity has which sign. SSIM has decreased slightly for the SameAR test compared to the unbalanced experiment, but increased slightly for DiffAR. Overall, this model's performance is very similar to that of the unbalanced log10 experiment.

Figures 30 and 31 show example images for SameAR and DiffAR, respectively. These images appear qualitatively poorer (more blurry, less similar in structure to the target images) than the previous 52-AR models that used unsigned HMI data. As expected from a training dataset with no bias towards either hemisphere, predictions from this network do not appear to favor any particular sign orientation over the other, unlike what we saw with the initial unbalanced model. Some examples show it predicting the polarities on the correct side of the image (e.g.,



(a) NRMSE



(b) SSIM

Figure 29: Results for model trained on rotated signed HMI data and log10 scaled AIA data.

the second row of Figure 30), while some have reversed polarities similar to what we saw with the unbalanced model (e.g., the third row of Figure 31). Several predictions also have no clear polarity split–instead of mostly positive on one side of the image and negative on the other as in the target image, areas of positive and negative will be scattered throughout the predicted image. The first rows of both Figure 30 and Figure 31 show this characteristic of scattered polarity. Overall these examples support our assumption that there is not an easy way to tell the sign of magnetic activity based solely on the corresponding AIA signal image.

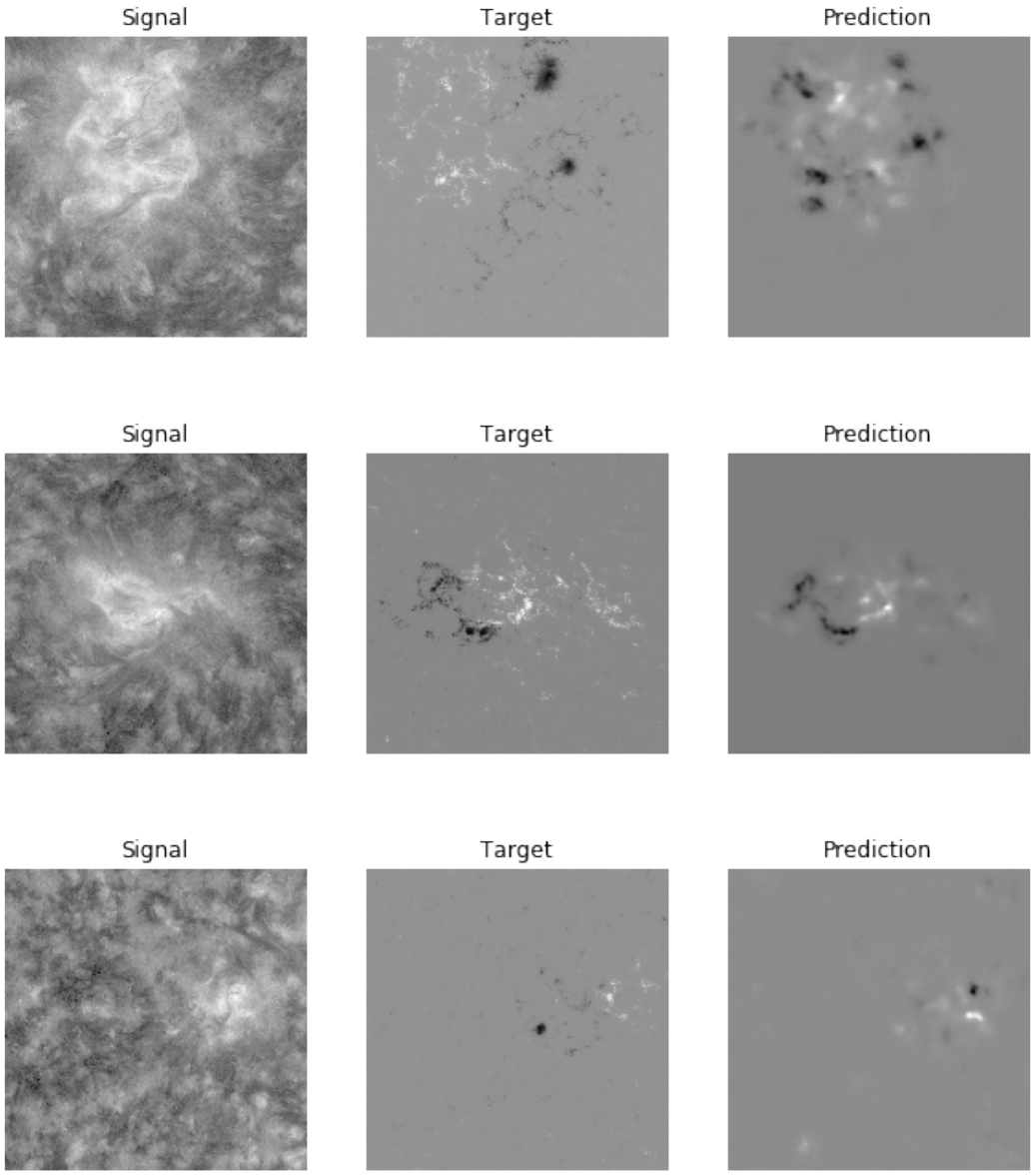


Figure 30: Predictions (SameAR) from the 52-AR experiment with training data randomly rotated to remove any hemisphere bias, using signed HMI and log10 AIA

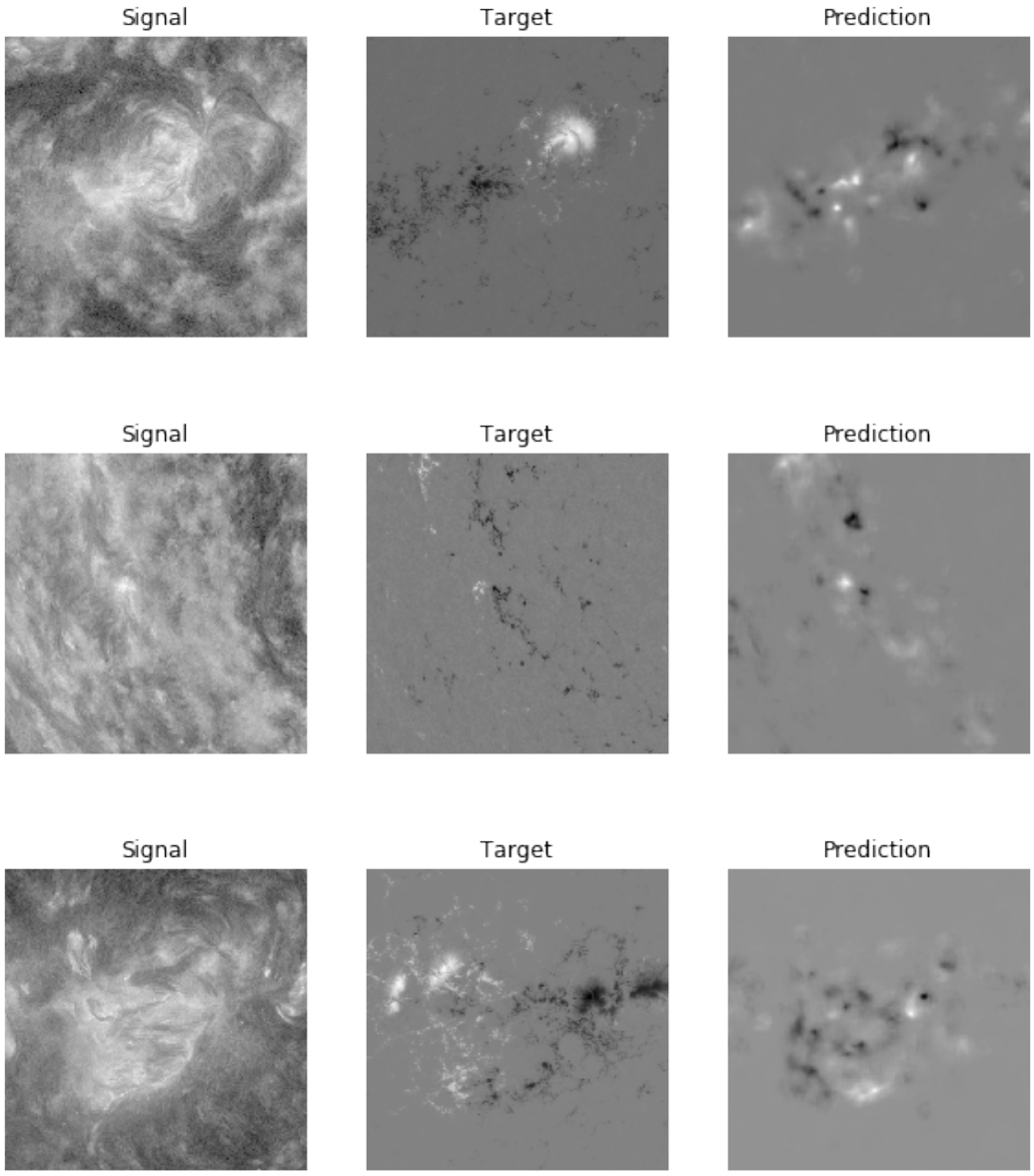


Figure 31: Predictions (DiffAR) from the 52-AR experiment with training data randomly rotated to remove any hemisphere bias, using signed HMI and log10 AIA

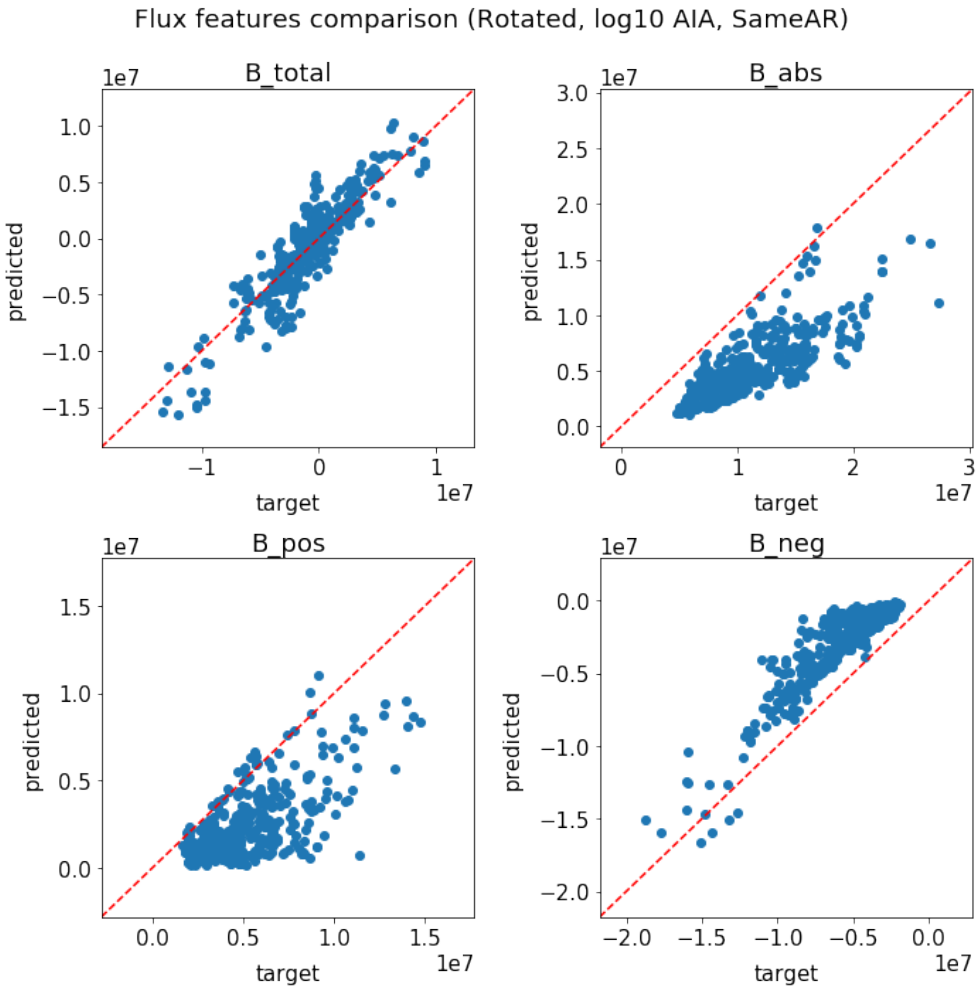


Figure 32: Target vs. predicted flux values (52-AR log10 AIA / rotated signed HMI, SameAR)

Figure 32 shows target versus predicted flux values for the SameAR test. Compared to the unsigned experiments, we are seeing more under-prediction of B_{abs} (top right), and the points are more spread out as opposed to falling mostly in a line as they did for the unsigned experiments, indicating a less accurate relationship between target and predicted values. Total signed flux, B_{total} in the top left, does follow the reference line very well. This is likely due to under-predicting

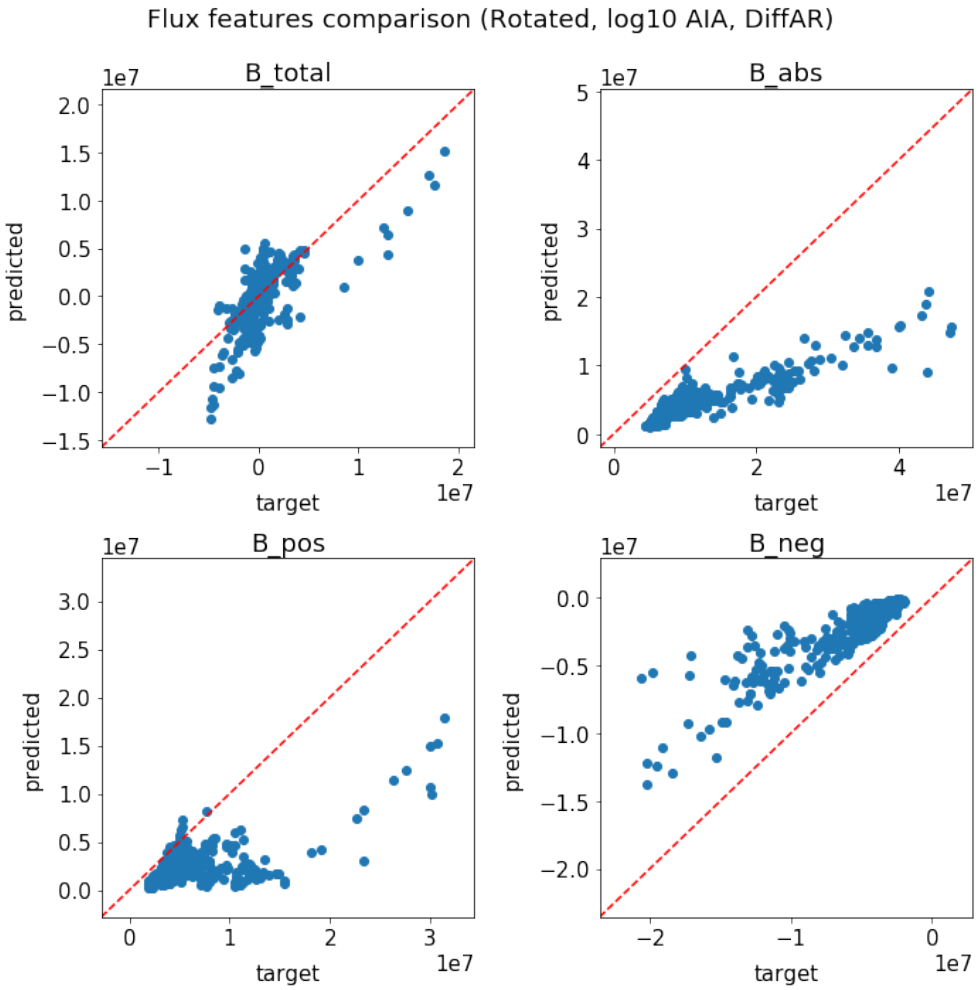


Figure 33: Target vs. predicted flux values (52-AR log10 AIA / rotated signed HMI, DiffAR)

positive flux B_{pos} in the lower left while over-predicting negative flux B_{neg} in the lower right, and the sum of these cancelling out to give a B_{total} that is more or less accurate.

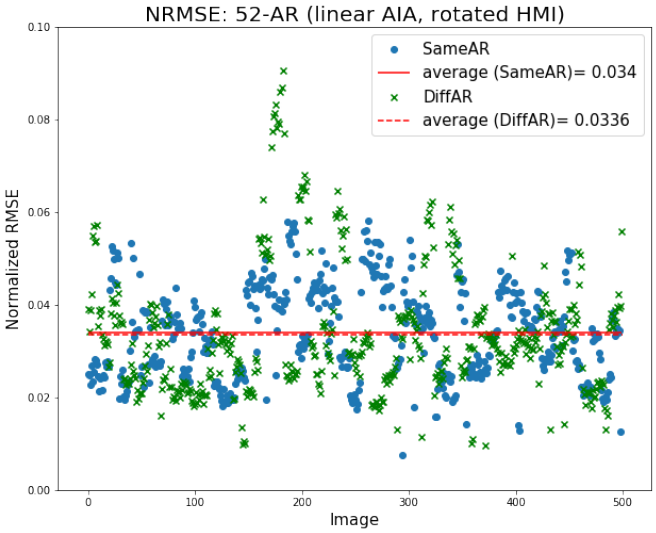
These trends are similar but worsened for the DiffAR test, shown in 33. Under-prediction of B_{abs} is quite severe here, as is under-prediction of B_{pos} and over-prediction of B_{neg}.

Linear AIA: Figure 34 shows NRMSE and SSIM for the model trained on rotated HMI data and linearly scaled AIA data. NRMSE values are extremely similar to the previous experiment that used log10 AIA data, and show the same lack of performance gap between SameAR and DiffAR as in the other 52-AR experiments.

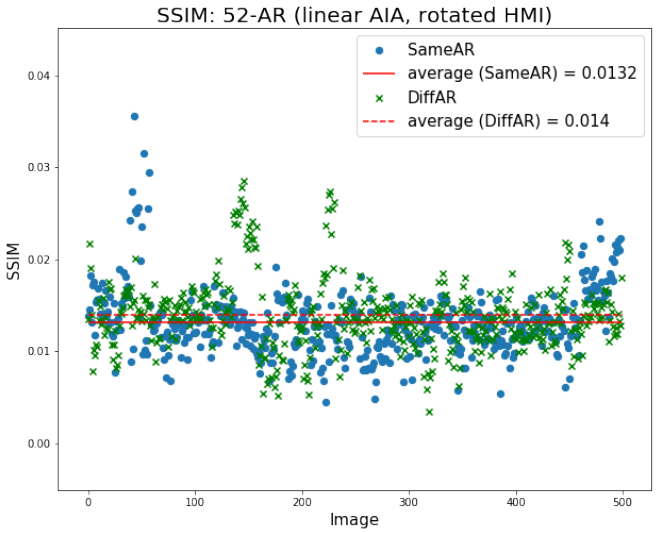
Figures 35 and 36 show comparison images for SameAR and DiffAR respectively. We see mostly the same characteristics as with the log10 model, including a similar mix of correctly placed polarities, reversed polarities, and polarities mixed throughout the image.

Overall, these balanced experiments confirm that the prediction task becomes difficult when there is no statistical bias provided by the training data for what the sign of predicted magnetic activity should be.

The generally poor quality of these balanced models' predictions indicate that training a single model on data from both hemispheres is likely not the best approach to finding a model that performs well on arbitrary active regions. Similar to the 52-AR experiments using unsigned HMI data, we also see no clear evidence of any advantage to using log10 AIA over linear AIA or vice versa as signal images in this case.



(a) NRMSE



(b) SSIM

Figure 34: Results for model trained on rotated signed HMI data and linearly scaled AIA data.

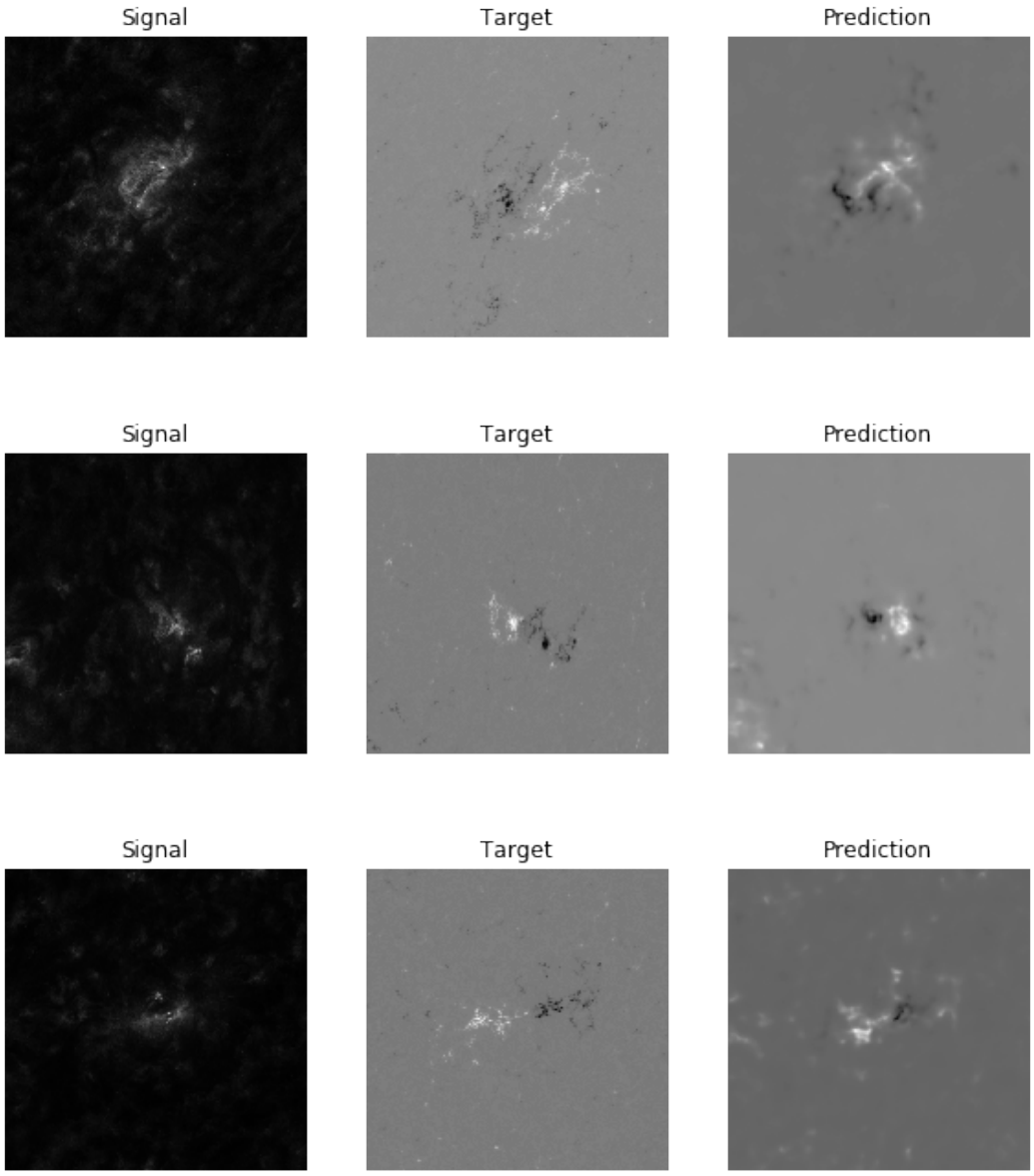


Figure 35: Predictions (SameAR) from the 52-AR experiment with training data randomly rotated to remove any hemisphere bias, using signed HMI and linear AIA

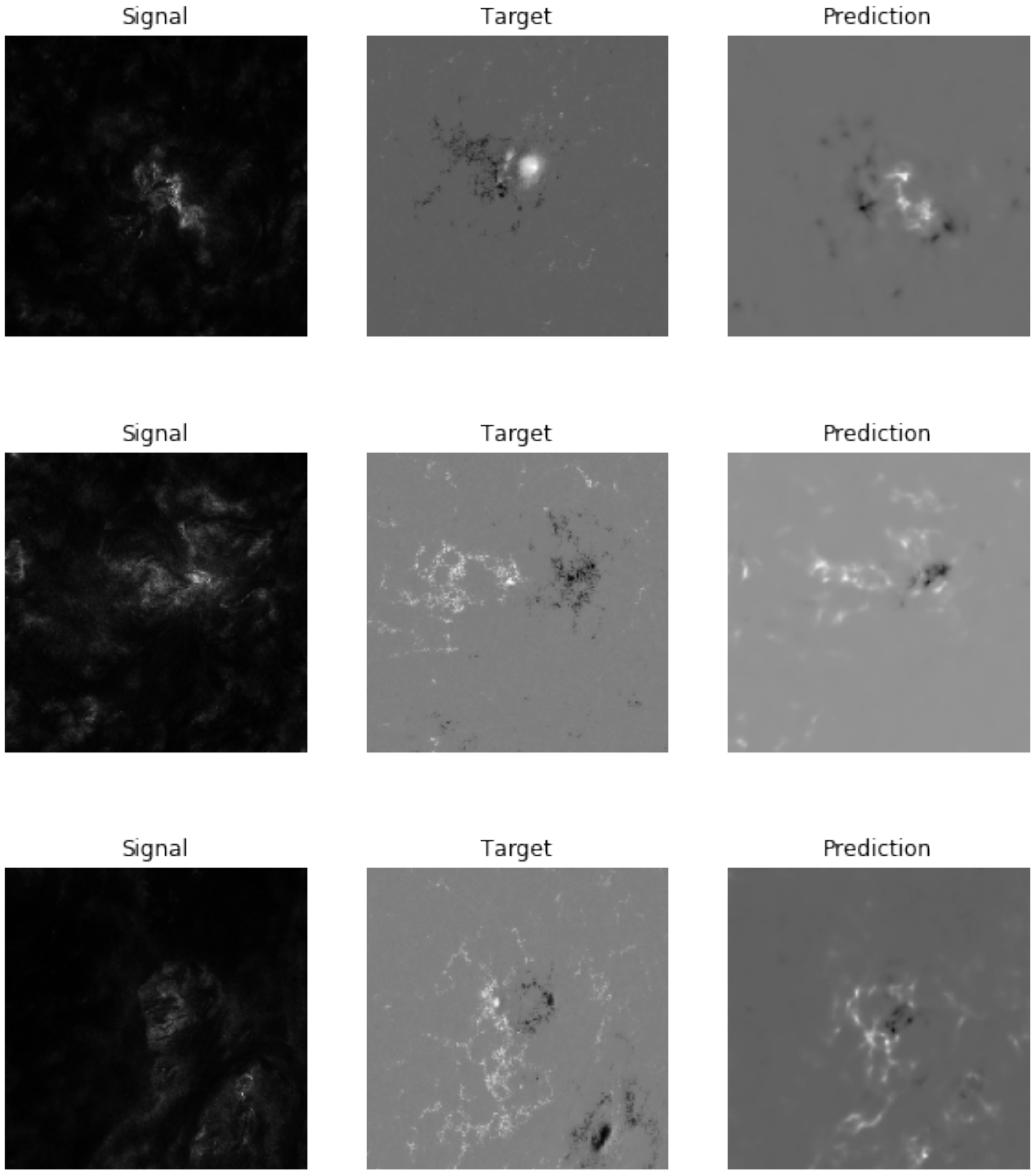


Figure 36: Predictions (DiffAR) from the 52-AR experiment with training data randomly rotated to remove any hemisphere bias, using signed HMI and linear AIA

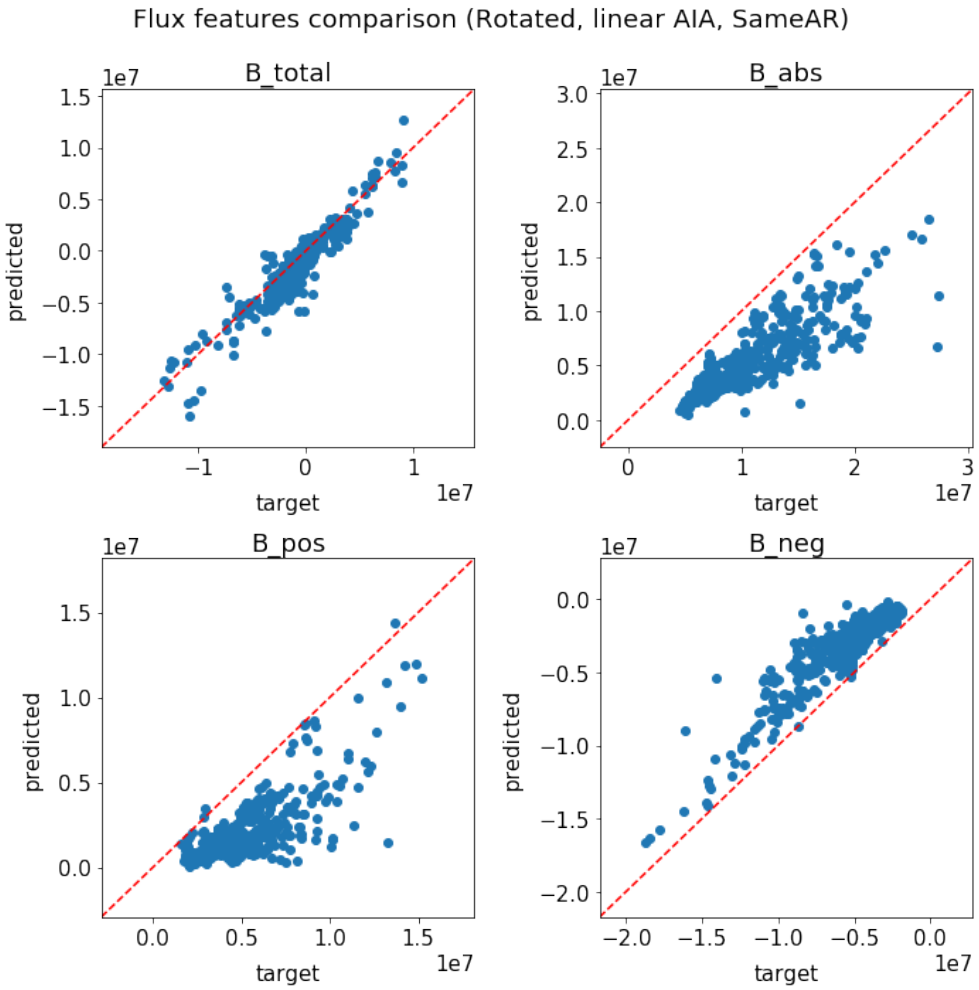


Figure 37: Target vs. predicted flux values (52-AR linear AIA / rotated signed HMI, SameAR)

Figures 37 and 38 show target versus predicted flux feature values for the SameAR and DiffAR tests respectively. These show the same general trends as the flux feature comparisons for the log10 experiment, although the SameAR test here has less severe under-prediction of B_{abs}. There is still over-prediction of B_{neg}, but less so than in the log10 experiment. The DiffAR test in Figure 38 looks very similar to its log10 counterpart in Figure 33, and worse overall than the SameAR

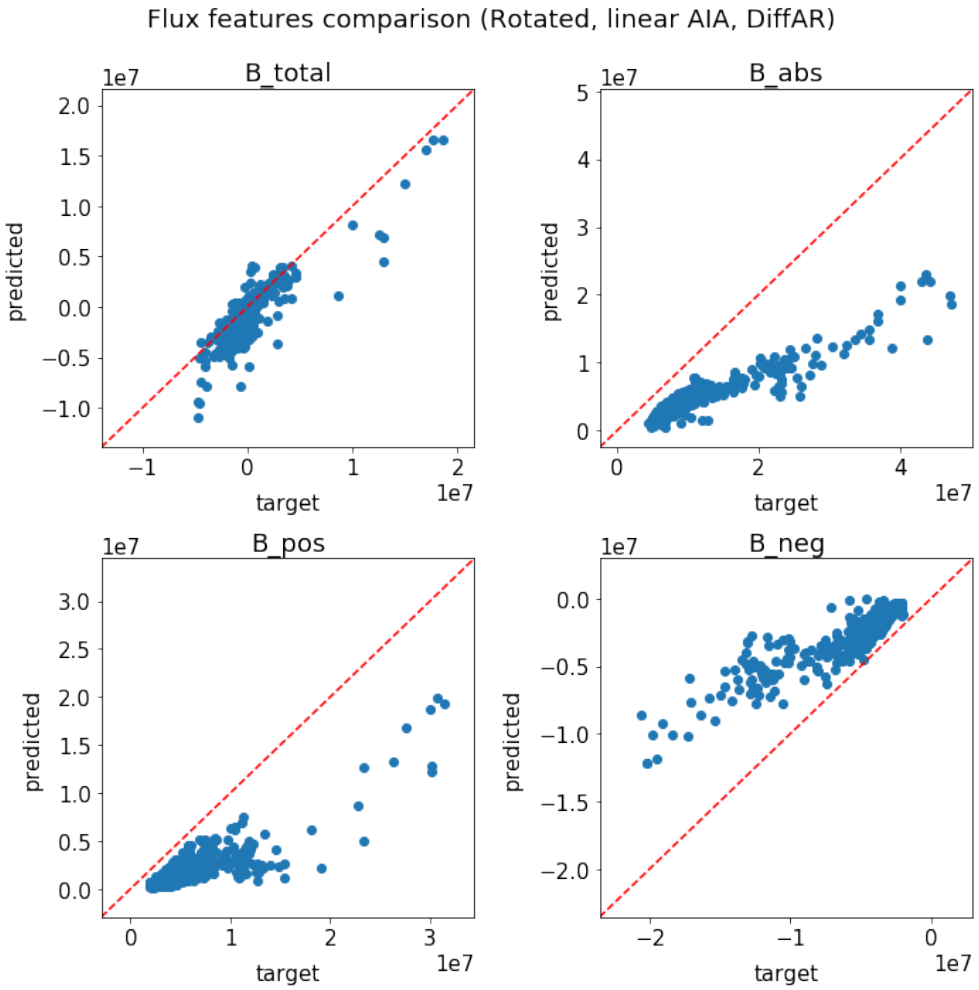


Figure 38: Target vs. predicted flux values (52-AR linear AIA / rotated signed HMI, DiffAR)

test in Figure 37.

Overall, we are seeing generally poorer correlation between target and predicted flux feature values than we saw in the unsigned experiments, as expected given the lower performance of these models using training data drawn from both hemispheres.

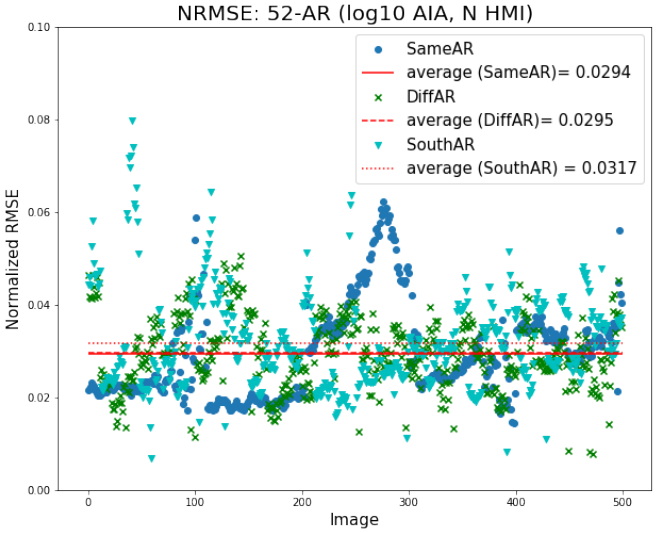
4.5.3 Single Hemisphere

We will now discuss results from the experiments using models trained on active regions from only the northern hemisphere. In addition to the SameAR and DiffAR test datasets used in the previous experiments, here we include results from a SouthAR test dataset, where image pairs are sourced from only the southern hemisphere and rotated 180° prior to being introduced to the network.

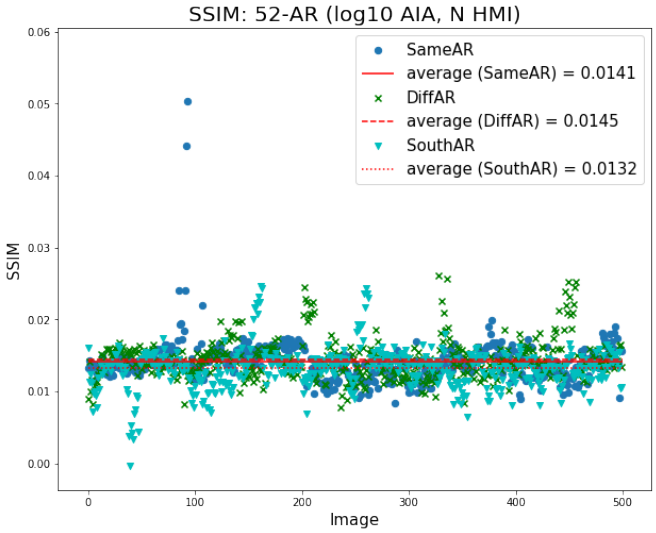
Log10 AIA: Figure 39 shows NRMSE and SSIM for the model trained using north-only HMI data and log10 AIA data. The NRMSE values of 0.0294 and 0.0295 for SameAR and DiffAR respectively are improvements compared to those seen in both the unbalanced and rotated signed HMI experiments–this is as expected, since the prediction task should become easier with data that almost entirely follows the polarity convention of a single hemisphere. For this experiment using log10 AIA data as signal, SSIM has improved over the previous signed experiments for the SameAR test. However, DiffAR SSIM is lower, matching that of the unbalanced DiffAR test.

Figure 39 also shows these results for SouthAR test data. Error for SouthAR has increased compared to DiffAR, and SSIM has decreased slightly as well. However, this SouthAR performance does have better NRMSE than the DiffAR performances of both the unbalanced and rotated experiments.

Figures 40 and 41 show example images from SameAR and DiffAR respec-



(a) NRMSE



(b) SSIM

Figure 39: Results for model trained on single-hemisphere HMI data and log10 scaled AIA data

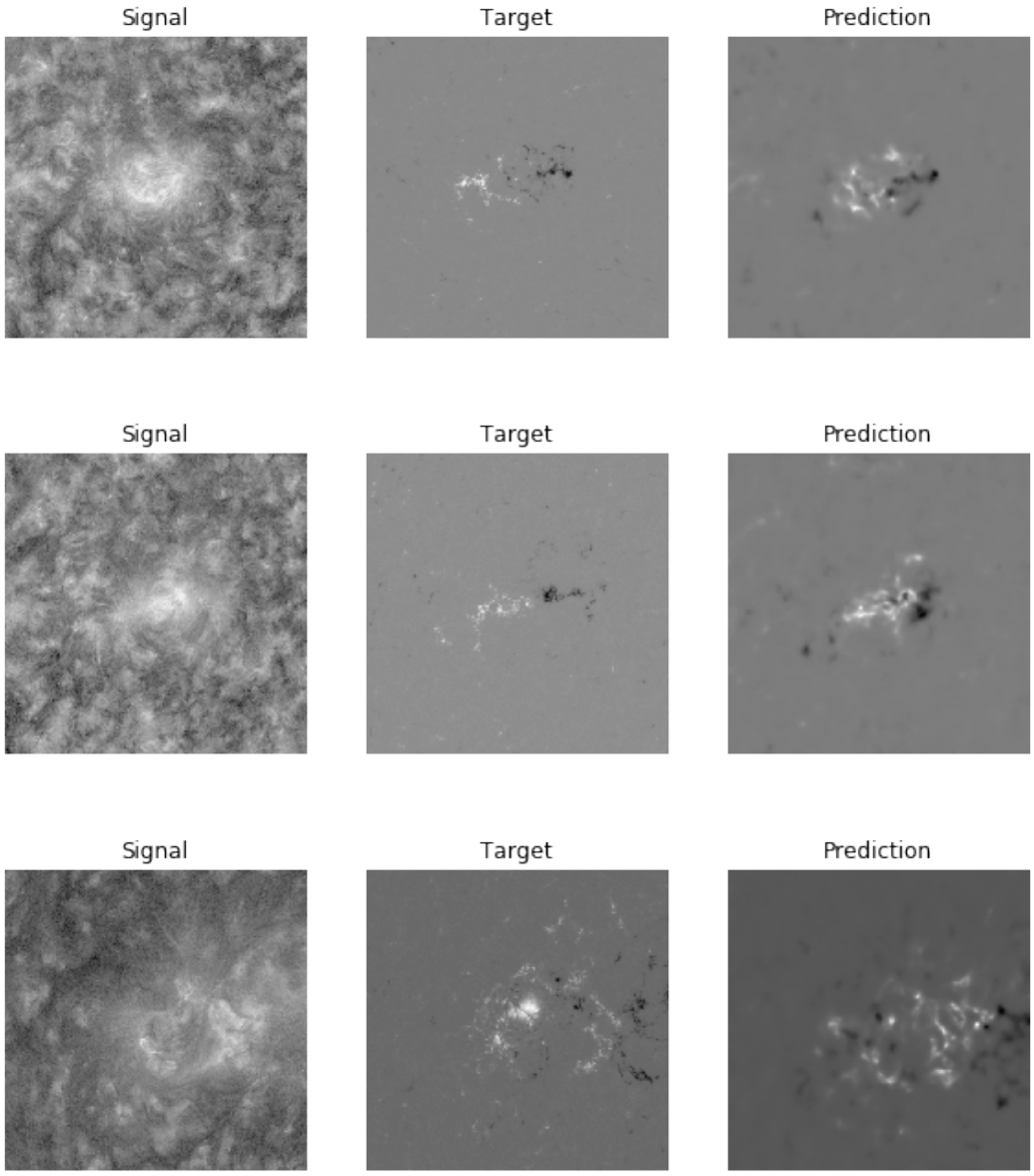


Figure 40: Predictions (SameAR) from the 52-AR experiment with training data chosen from only the north hemisphere, using signed HMI and log10 AIA data

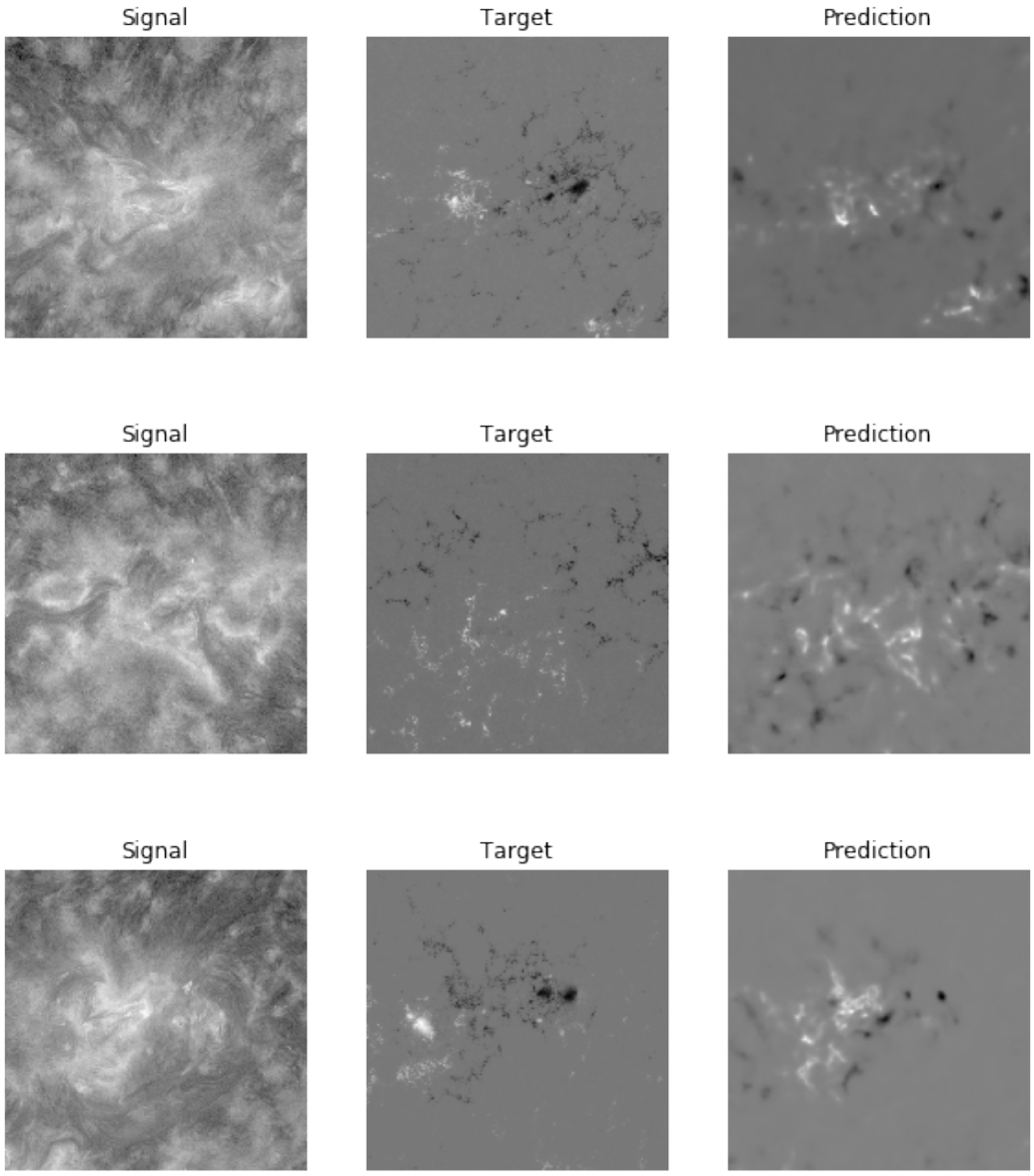


Figure 41: Predictions (DiffAR) from the 52-AR experiment with training data chosen from only the north hemisphere, using signed HMI and log10 AIA data

tively. The network is now generally putting the correct signed polarities in the correct place in the image and capturing the overall structure of the active regions more effectively than any of the previous experiments using signed magnetograms. However, these predictions are still missing smaller areas of activity and experiencing the blurriness seen in the previous experiments. This supports the improvement in NRMSE and SSIM versus the other signed experiments–using only data from one hemisphere has removed much of the ambiguity about the sign of the magnetic activity, and has improved overall image quality.

Figure 42 shows example predictions from the SouthAR test dataset. These images show a decline in quality versus SameAR and DiffAR results that is consistent with the error and similarity discussed above. We see occasional instances of misplaced polarity, and the usual trend of blurred details even when overall structure is reasonably sound. Overall, the results of the SouthAR test support the idea that training a separate model for each hemisphere is the best approach for predictions of signed magnetograms.

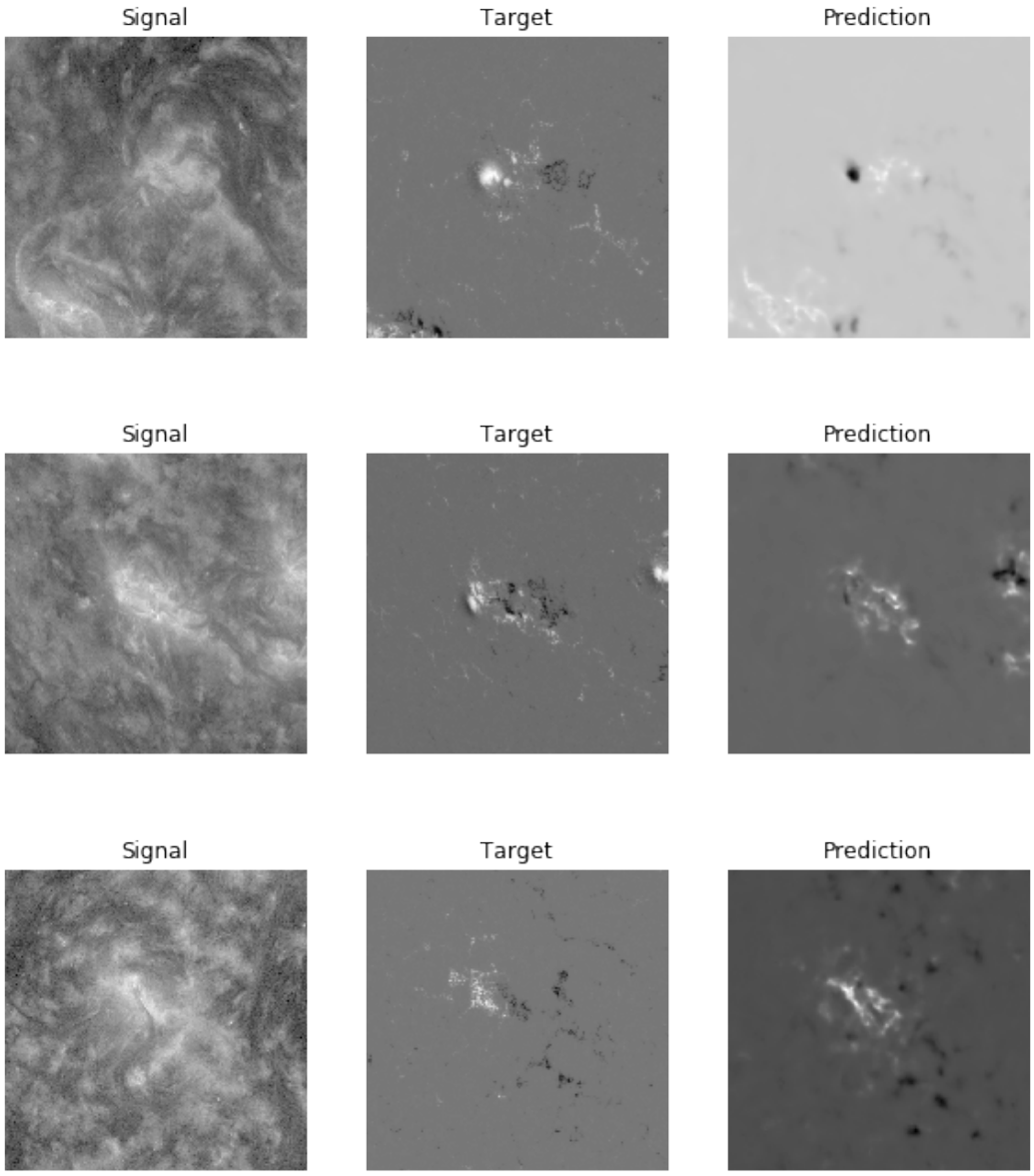


Figure 42: Predictions (SouthAR) from the 52-AR experiment using a model trained on only the northern hemisphere, using signed HMI and log10 AIA data

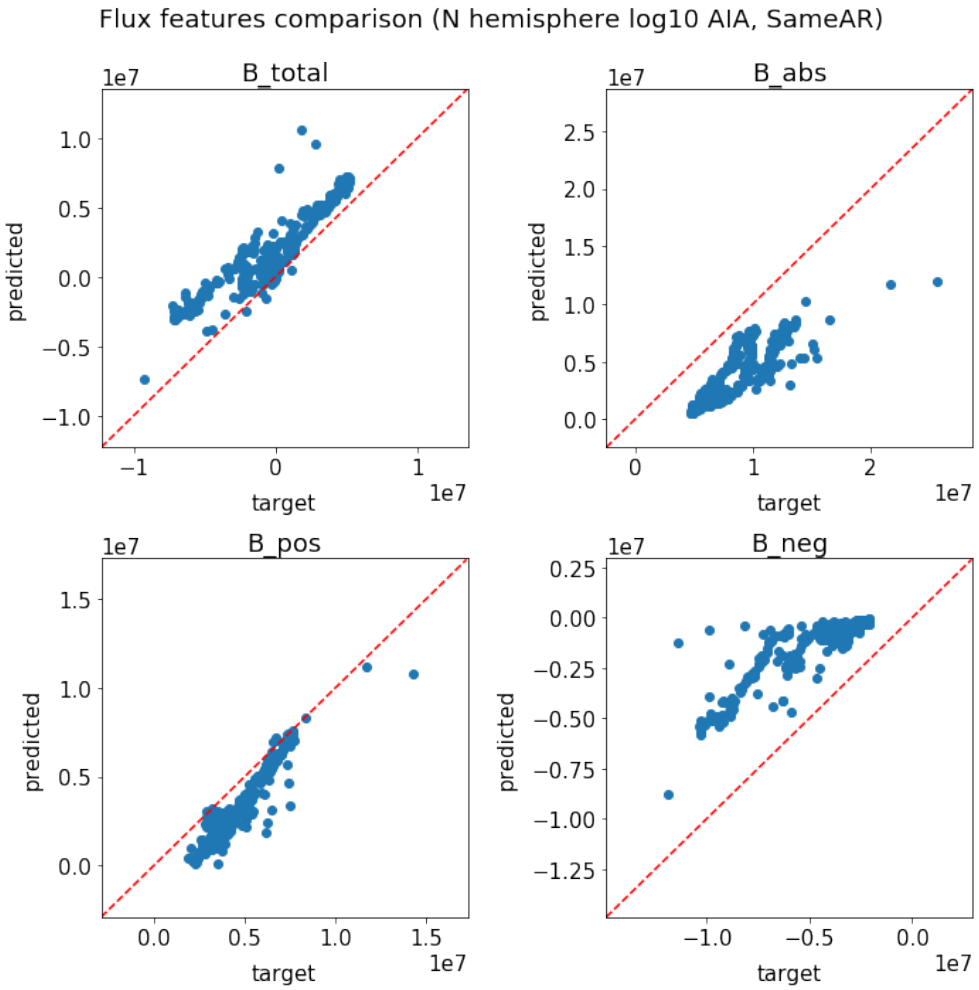


Figure 43: Target vs. predicted flux values (52-AR log10 AIA / north hemisphere HMI, SameAR)

Figures 43, 44, and 45 show target versus predicted flux feature values for the SameAR, DiffAR, and SouthAR tests respectively. Compared to the rotated HMI experiments, prediction of total positive flux B_{pos} has improved- it is now closer to the reference line than it was for the rotated experiment. B_{abs} is still largely under-predicted, although the DiffAR test here does this less severely than the DiffAR tests from the rotated models. All three tests for this model appear to

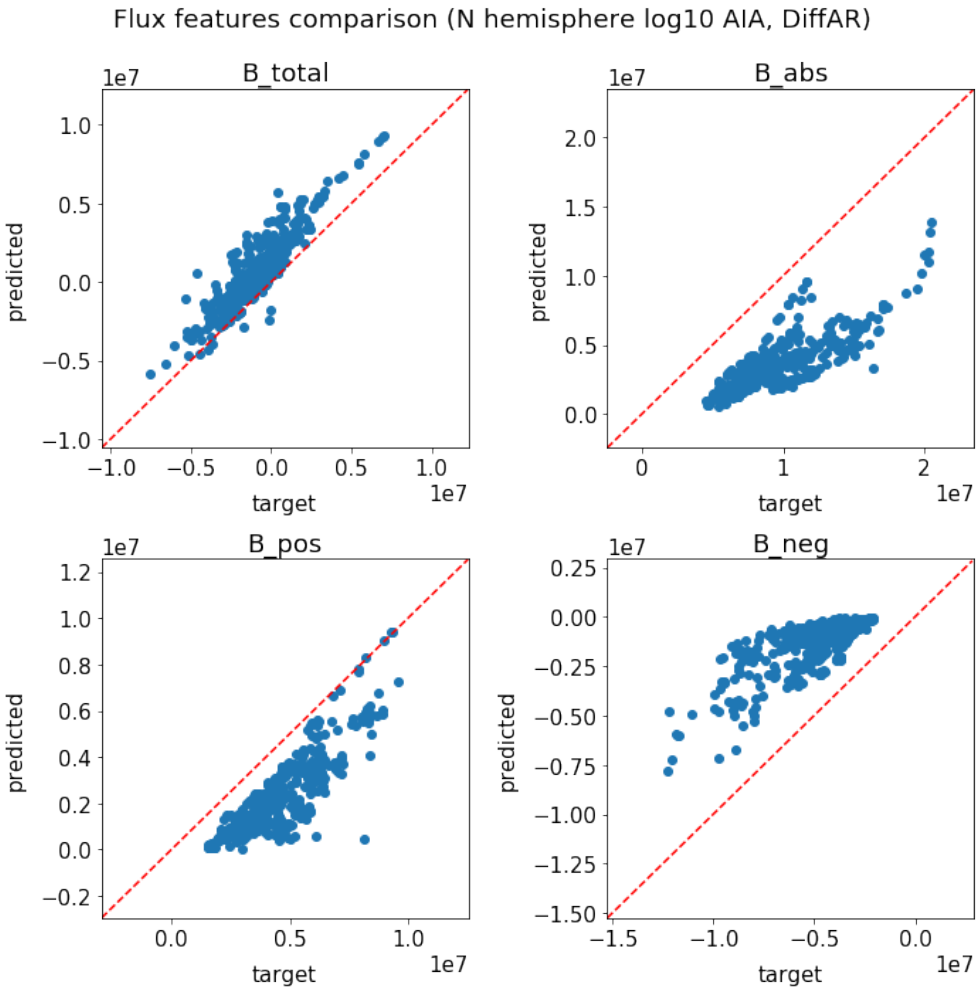


Figure 44: Target vs. predicted flux values (52-AR log10 AIA / north hemisphere HMI, DiffAR)

predict positive flux more accurately while struggling with B_{neg} the most out of all of the features, especially the SouthAR test. This model is under-predicting positive flux, but is doing so less severely than it is over-predicting negative flux, and so B_{total} ends up slightly over-predicted.

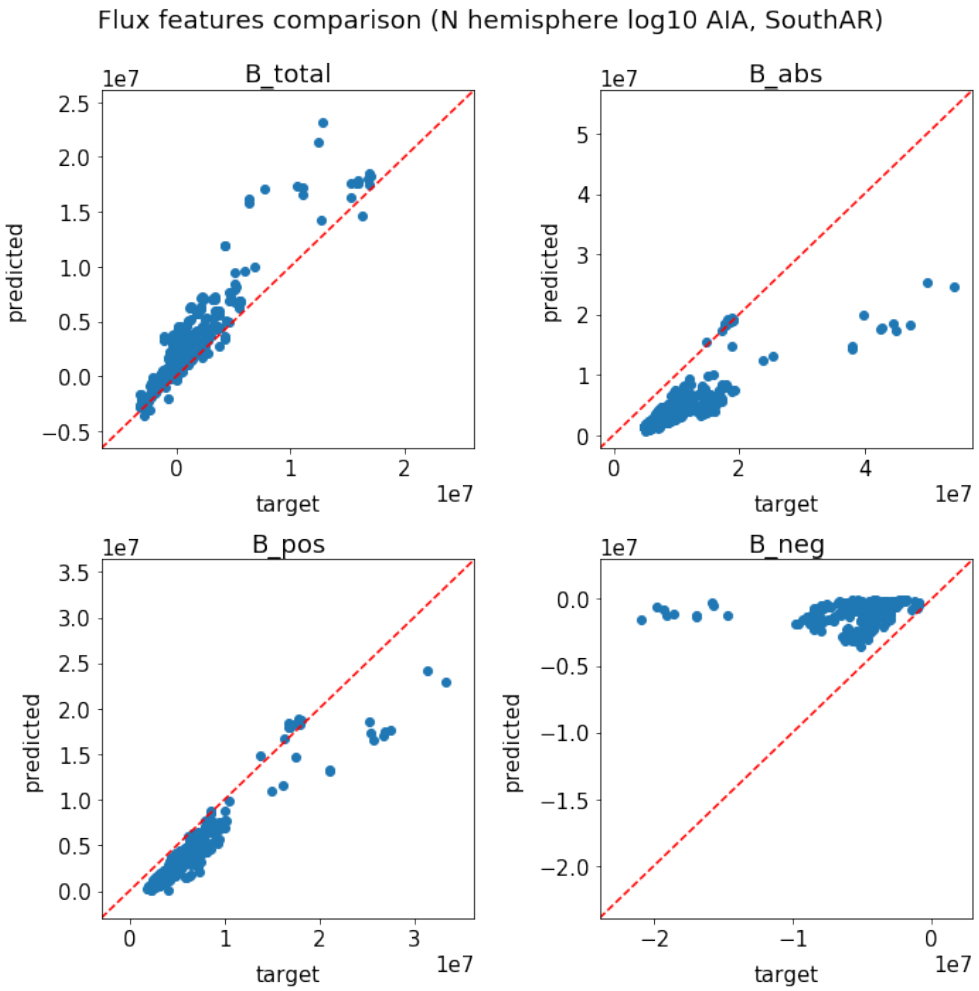
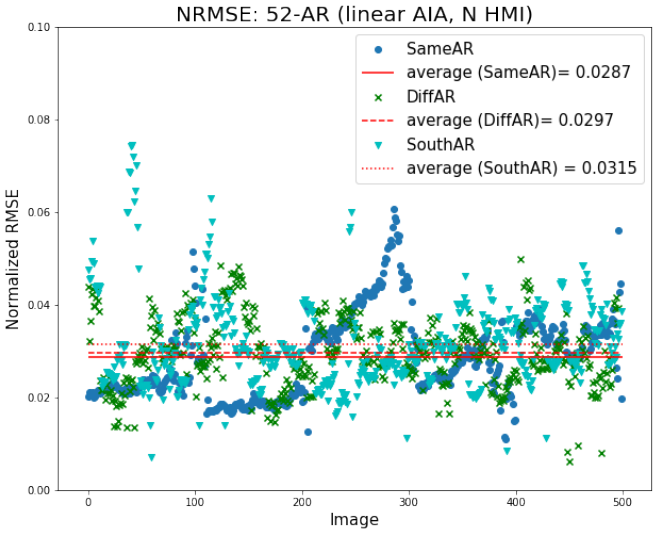


Figure 45: Target vs. predicted flux values (52-AR log10 AIA / north hemisphere HMI, SouthAR)

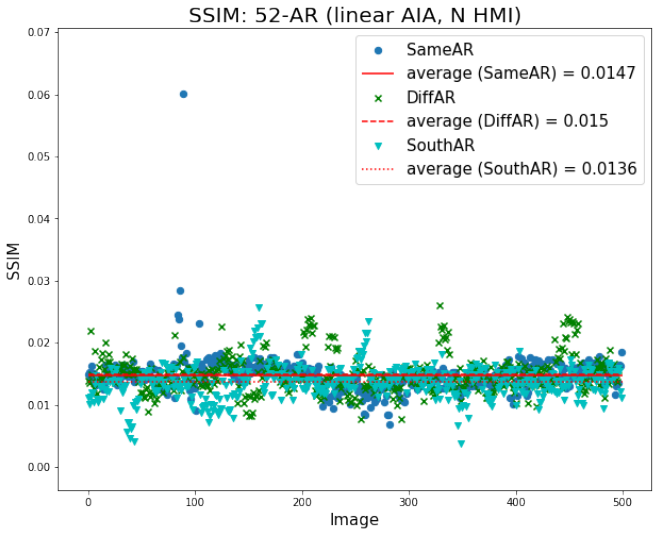
Linear AIA: Figure 46 shows NRMSE and SSIM for the experiment using linearly scaled AIA data. This experiment shows a more uniform improvement over the unbalanced and rotated models, with both lower NRMSE and higher SSIM across both SameAR and DiffAR tests. The consistency of this improvement could indicate that linearly scaled AIA is the better signal data when working with a single hemisphere, but the difference between these results and those from the single hemisphere log10 model are still small enough that it is difficult to say this decisively. Similar to the log10 AIA experiment, the SouthAR NRMSE has improved compared to the DiffAR NRMSE of both the unbalanced and rotated experiments, but is worse than the SameAR and DiffAR results from this same model.

The results of both single hemisphere experiments support the idea that training on a single hemisphere and then rotating any test data that comes from the other hemisphere can improve performance over a model trained on both hemispheres. However, SouthAR performance is still poorer than the DiffAR performance for the same model. This indicates that training one model for each hemisphere and sorting test data accordingly is the best approach to make quality predictions for arbitrary ARs, especially since the large quantity of available data makes training multiple models an accessible solution.

Figures 47, 48, and 49 show example images from the SameAR, DiffAR, and SouthAR tests, respectively. These images generally appear more structurally



(a) NRMSE



(b) SSIM

Figure 46: Results for model trained on single-hemisphere HMI data and linearly scaled AIA data

sound and less blurry than those seen in the log10 experiment, supporting the idea that linearly scaled AIA may have a slight edge in the single hemisphere tests. The blurriness in these single-hemisphere examples is slightly more pronounced and structure is generally less accurate than the best of the unsigned experiments, but this is as expected to some degree due to the increased difficulty of a signed prediction. Leaning on Hale's law to assist in assigning the correct polarities has certainly provided better performance than the other signed experiments, but not all active regions will follow the polarity orientation expected for their hemisphere perfectly and predicting only the magnitude of activity appears to still be the easier task.

Figures 50, 51, and 52 show target versus predicted flux feature values for the SameAR, DiffAR, and SouthAR tests respectively. We see generally better relationships between target and predicted values than the log10 experiment, and B_{neg} especially has improved for the SameAR and DiffAR tests. As with the log10 experiment, we see generally better prediction for B_{pos} than for B_{neg}, which leads to over-prediction for B_{total}. B_{abs} is under-predicted, as follows from the trends for total positive and total negative flux. SouthAR is again notably worse than the other two tests, supporting multiple models as the best approach to make predictions for arbitrary active regions.

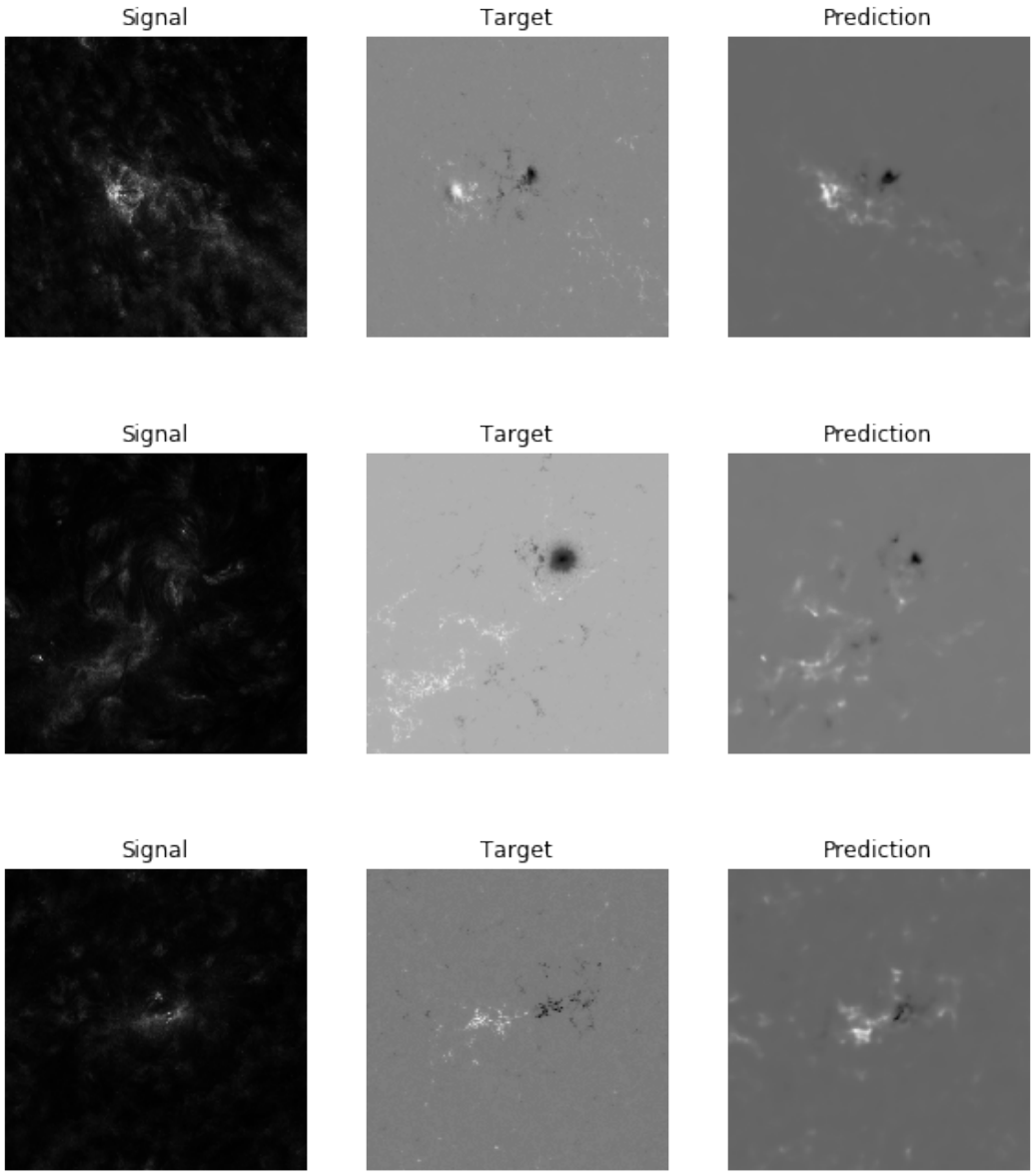


Figure 47: Predictions (SameAR) from the 52-AR experiment with training data with training data from only the northern hemisphere, using signed HMI and linear AIA data

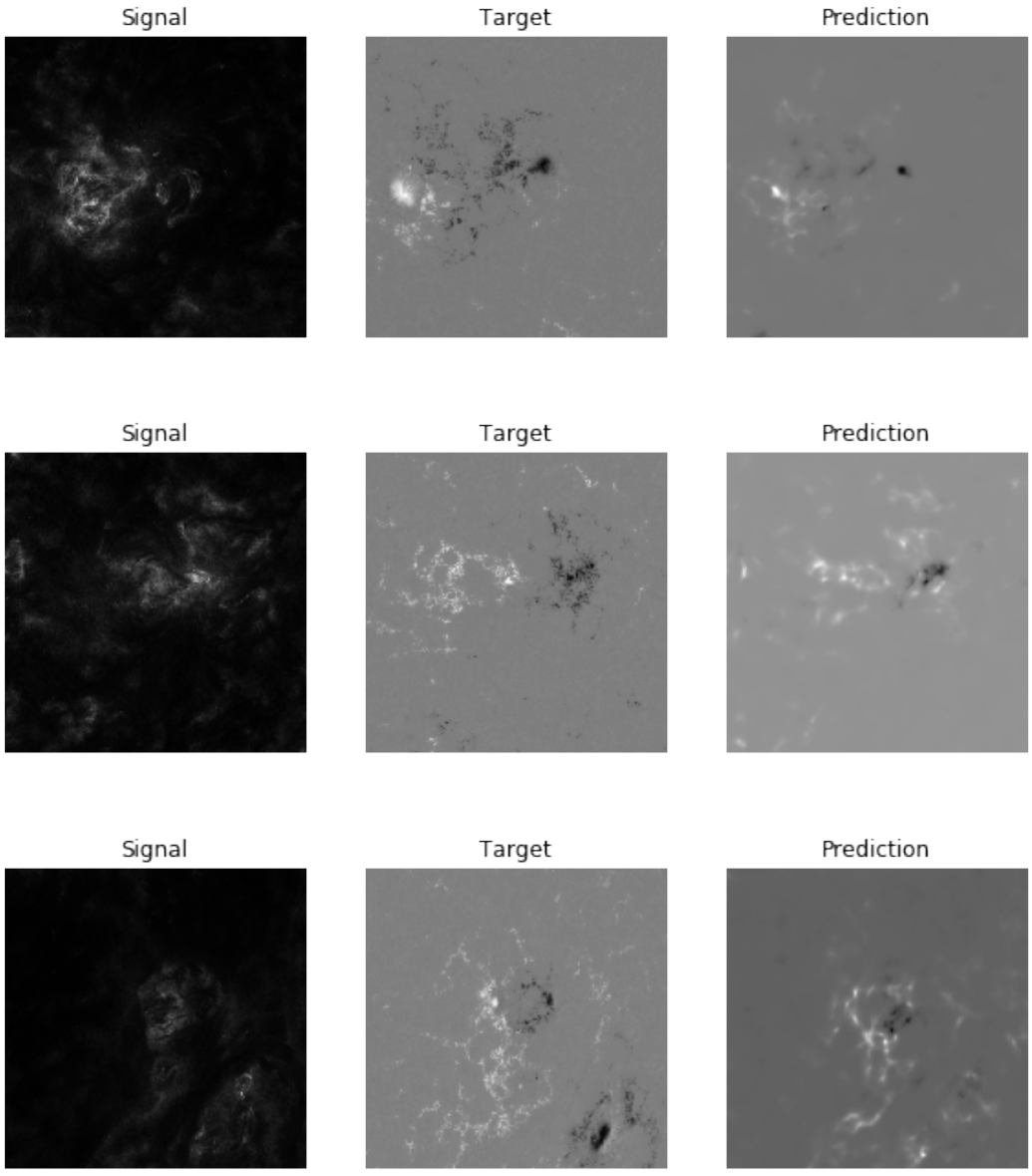


Figure 48: Predictions (DiffAR) from the 52-AR experiment with training data from only the northern hemisphere, using signed HMI and linear AIA data

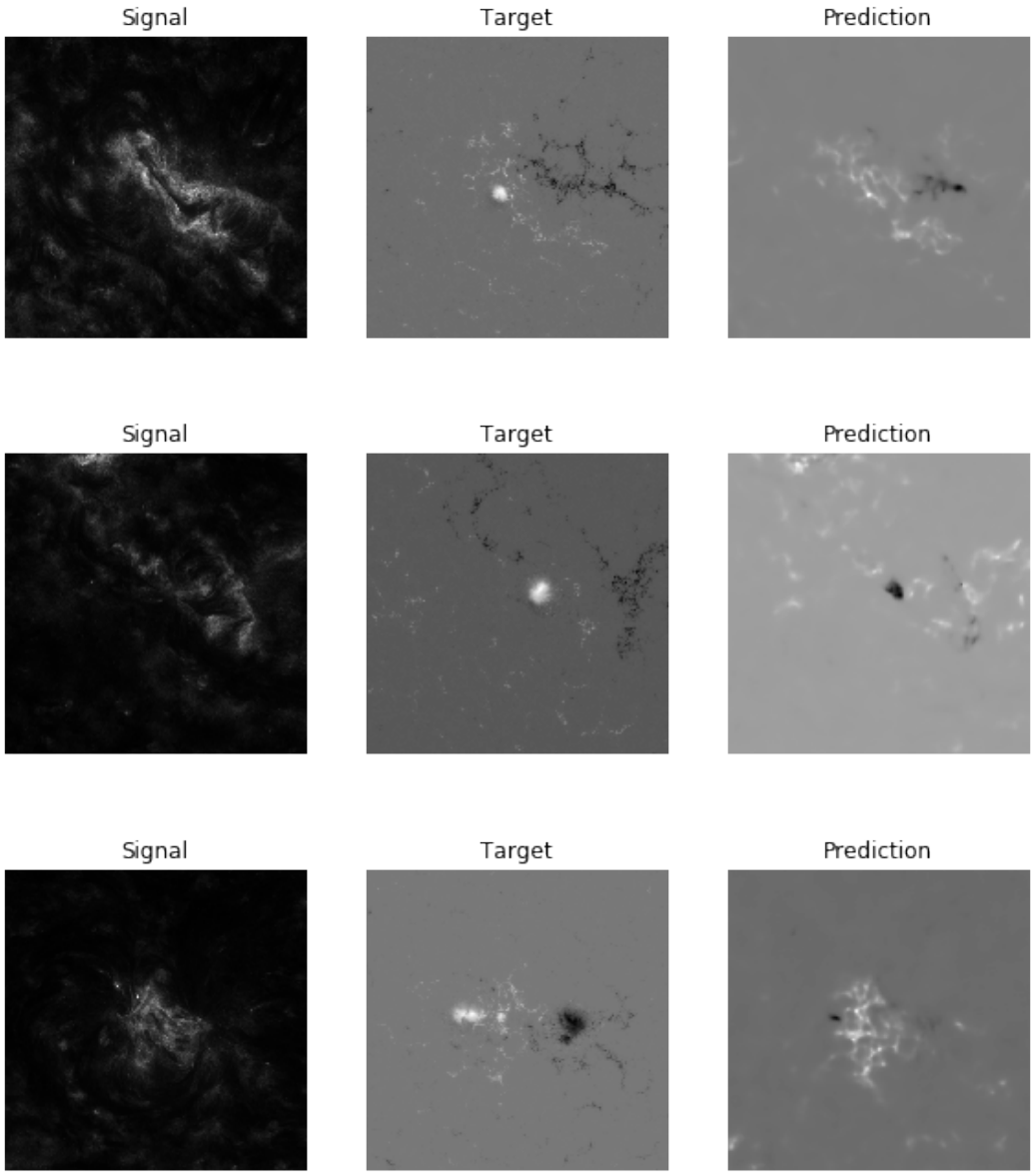


Figure 49: Predictions (SouthAR) from the 52-AR experiment with training data from only the northern hemisphere, using signed HMI and linear AIA data

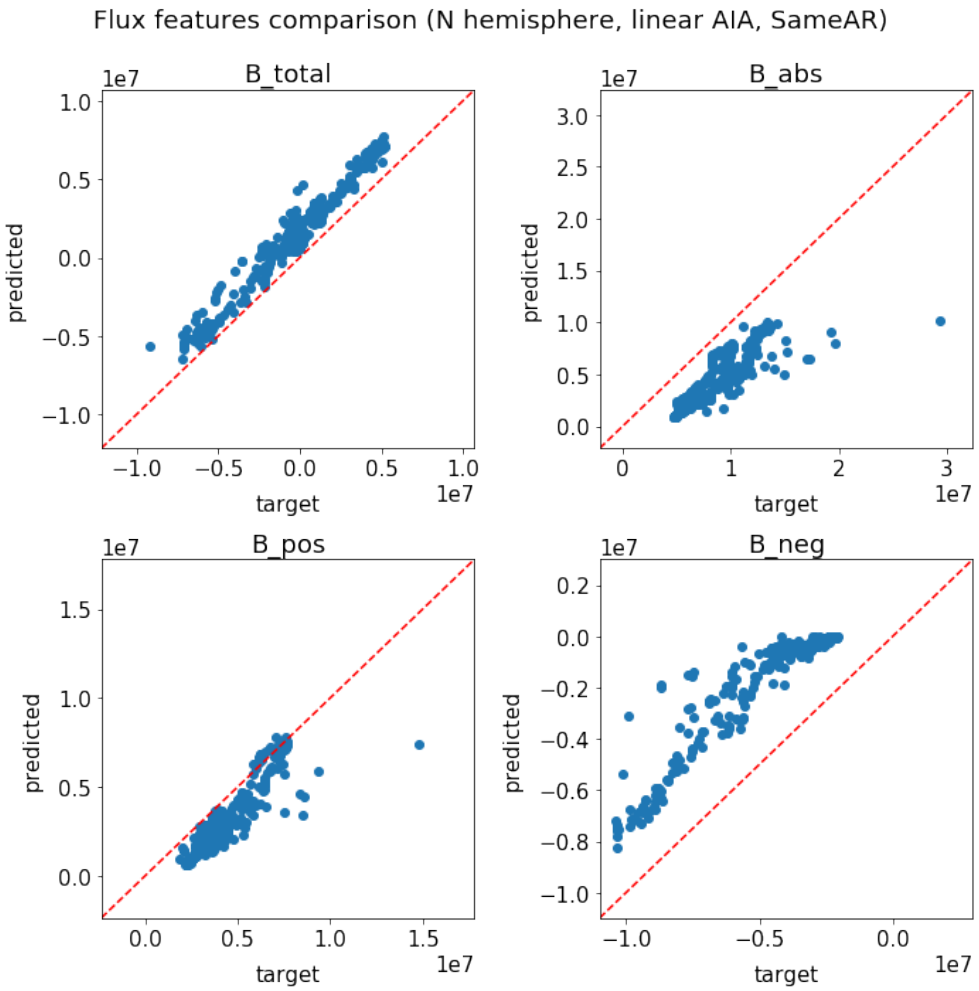


Figure 50: Target vs. predicted flux values (52-AR linear AIA / north hemisphere HMI, SameAR)

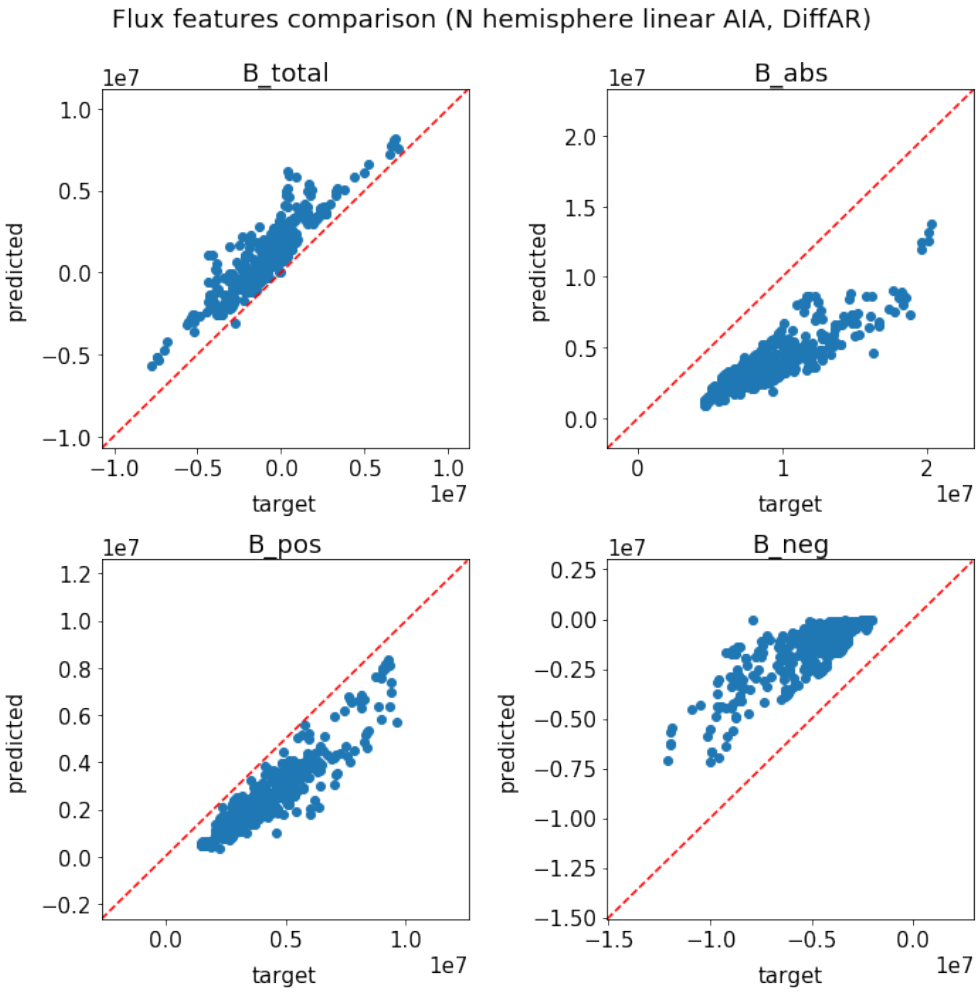


Figure 51: Target vs. predicted flux values (52-AR linear AIA / north hemisphere HMI, DiffAR)

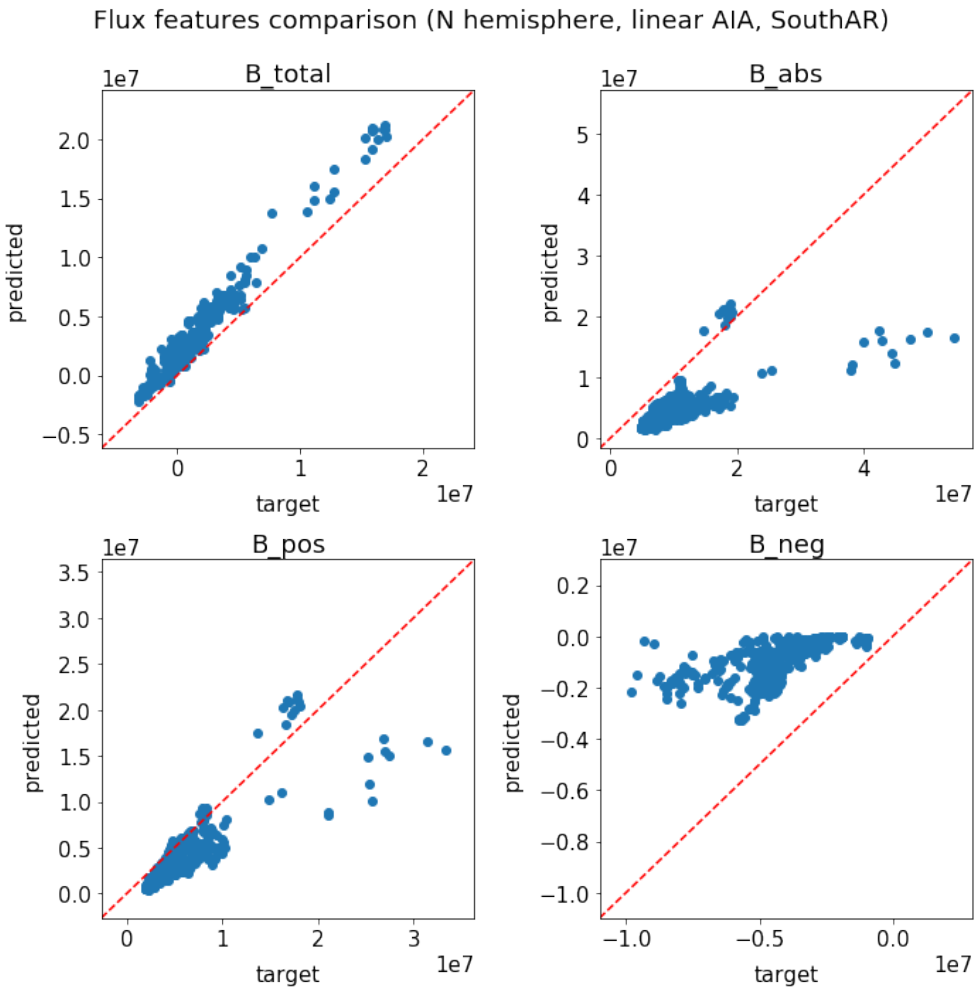


Figure 52: Target vs. predicted flux values (52-AR linear AIA / north hemisphere HMI, SouthAR)

5 CONCLUSIONS AND FUTURE WORK

This thesis explored a variety of approaches to using a U-Net FCNN architecture to predict line-of-sight magnetograms from 304 Å EUV images of solar active regions. The first experiment on a single active region showed that these predictions can be made successfully on this data for a single active region at similar volumes and constraints to those seen in past applications of U-Net for biomedical image processing [3], but do not extend successfully to arbitrary active regions at this scale.

We then increased the size and variety of the training data significantly, and were able to find suitable training dataset characteristics (both volume and variety) and network parameters to enable the network to provide predictions for arbitrary active regions that are similar in quality to predictions for active regions seen in training. A training dataset of ~ 500 image pairs sourced from ~ 52 active regions is a suitable data volume to achieve generalization. We also saw that models trained on unsigned magnetograms have good correspondence between target and predicted magnetic flux strength, although these models have a slight tendency to under-predict the sum of the total absolute magnetic flux (especially for target images with higher B_{abs}).

After earlier experiments that predicted only the magnitude of magnetic activity, we were able to extend to making signed predictions as well. The sign

of magnetic activity proved more difficult to predict than magnitude, and performance was generally poorer when training data was selected from both hemispheres. Sourcing training data from only one hemisphere in order to exploit prior knowledge of the effect that location has on the general structure of an active region (Hale's law) improved performance. However, these single-hemisphere models did not perform at the same level on data from the other hemisphere that had been rotated–this implies that rotating images from one hemisphere may not be an ideal way to imitate the characteristics of the other hemisphere. Another balanced experiment with data drawn equally from both hemispheres may be worthwhile because of this, but as performance from models trained on both hemispheres was generally worse overall than the single hemisphere models, it is not a priority at this time.

Overall, these signed experiments showed that if one wanted to build a system for making the best possible predicted signed magnetograms for arbitrary active regions, one should train separate models on data from each hemisphere, and separate incoming test data accordingly.

The models trained on signed magnetograms were more likely to under-predict the total unsigned flux B_{abs} compared to the unsigned models. They showed good correlation (though often slight over-prediction) between target and predicted total signed flux, B_{total}. These models all tended to predict positive flux more accurately than negative flux as well, but the best-performing signed HMI model

(north hemisphere HMI, linear AIA) noticeably improved prediction for B_{neg} over the other signed experiments.

While there is certainly room for improvement in the signed models' ability to correlate target and predicted flux, these approaches show potential for predicting flux values accurately for a much wider range of flux values than that allowed by [8].

In most experiments, no clear performance difference was observed overall when using linearly scaled versus log10 scaled AIA data as signal. The single hemisphere experiments saw slight improvement when using linearly-scaled AIA over log10, but the difference was fairly small.

Although the experiments conducted here have revealed suitable training dataset characteristics to generalize the modality transfer and an approach that can give uniform performance regardless of active region location, the overall quality of the predictions can still be improved–in future work, several possible methods of doing this could be explored.

The blurriness and tendency to miss small details that were universal in the predicted images presented here are common problems when using FCNNs to generate images. One method to improve image quality by addressing this blurring specifically would be to make this U-Net the generator network in a generative adversarial network (GAN) system [18], as several previous works involving image translation for SDO data have done ([8], [19], and [17]). We would add a second

neural network (called a discriminator network) trained to tell the difference between a true HMI magnetogram and the predicted images made by the generator network. The two networks would then be trained simultaneously in a zero-sum game where the generator attempts to learn how to make predictions good enough to fool the discriminator, and the discriminator attempts to tell the difference between real images and increasingly more realistic fake images. This approach is often used as a way to improve quality and reduce blurring when using CNN's to generate images [18].

Another method for improving prediction quality could be to introduce a weighted cost function, wherein certain pixels in the image are given higher priority than others during training. This approach was used by [1] to improve U-Net's ability to segment cells that were touching in microscopy images; by giving the pixels at the borders of cells that are touching a larger weight, prediction error at those pixels is more severely penalized than the rest of the image. When predicting magnetograms, one could give the pixels along the magnetic neutral line higher importance than the pixels of the rest of the image. As there is evidence that there is a relationship between the characteristics of the magnetic neutral line (length, curvature, number of separate fragments, etc.) and the structure and magnitude of the magnetic activity in an active region [**?**], training with an emphasis on correctly predicting this line could lead to improvement in the structure of predicted magnetograms.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Lecture Notes in Computer Science*, p. 234–241, 2015.

[2] W. D. Pesnell, B. J. Thompson, and P. Chamberlin, "The solar dynamics observatory (SDO)," *Solar Physics*, vol. 275, no. 1-2, pp. 3–15, 2012.

[3] C. Ounkomol, S. Seshamani, M. M. Maleckar, F. Collman, and G. R. Johnson, "Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy," *Nature methods*, vol. 15, no. 11, pp. 917–920, 2018.

[4] P. Liewer, J. Qiu, and C. Lindsey, "Comparison of helioseismic far-side active region detections with stereo far-side euv observations of solar activity," *Solar Physics*, vol. 292, no. 10, pp. 1–30, 2017.

[5] J. Zhao, "Time-distance imaging of solar far-side active regions," *The Astrophysical Journal Letters*, vol. 664, no. 2, p. L139, 2007.

[6] J. Schou, J. Borrero, A. Norton, S. Tomczyk, D. Elmore, and G. Card, "Polarization calibration of the helioseismic and magnetic imager (hmi) onboard the solar dynamics observatory (sdo)," in *The Solar Dynamics Observatory*. Springer, 2010, pp. 327–355.

[7] M. L. Kaiser, T. Kucera, J. Davila, O. S. Cyr, M. Guhathakurta, and E. Christian, "The stereo mission: An introduction," *Space Science Reviews*, vol. 136, no. 1, pp. 5–16, 2008.

[8] T. Kim, E. Park, H. Lee, Y.-J. Moon, S.-H. Bae, D. Lim, S. Jang, L. Kim, I.-H. Cho, M. Choi, and et al., "Solar farside magnetograms from deep learning analysis of stereo/euvi data," *Nature Astronomy*, vol. 3, no. 5, p. 397–400, 2019.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[12] S. Khan, H. Rahmani, S. A. Shah, and M. Bennamoun, "A guide to convolutional neural networks for computer vision," *Synthesis Lectures on Computer Vision*, vol. 8, no. 1, p. 1–207, Feb 2018.

[13] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[14] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[15] K. Barczynski, H. Peter, L. P. Chitta, and S. K. Solanki, "Emission of solar chromospheric and transition region features related to the underlying magnetic field," *Astronomy & Astrophysics*, vol. 619, 2018.

[16] R. Galvez, D. F. Fouhey, M. Jin, A. Szenicer, A. Muñoz-Jaramillo, M. C. Cheung, P. J. Wright, M. G. Bobra, Y. Liu, J. Mason, and et al., "A machine-learning data set prepared from the nasa solar dynamics observatory mission," *The Astrophysical Journal Supplement Series*, vol. 242, no. 1, p. 7, 2019.

[17] E. Park, Y.-J. Moon, J.-Y. Lee, R.-S. Kim, H. Lee, D. Lim, G. Shin, and T. Kim, "Generation of solar uv and euv images from sdo/hmi magnetograms by deep learning," *The Astrophysical Journal Letters*, vol. 884, no. 1, p. L23, 2019. [Online]. Available: https://iopscience.iop.org/article/10.3847/2041-8213/ab46bb/pdf

[18] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[19] A. Dash, J. Ye, and G. Wang, "High resolution solar image generation using generative adversarial networks," 2021.

[20] J. R. Lemen, D. J. Akin, P. F. Boerner, C. Chou, J. F. Drake, D. W. Duncan, C. G. Edwards, F. M. Friedlaender, G. F. Heyman, N. E. Hurlburt *et al.*, "The atmospheric imaging assembly (aia) on the solar dynamics observatory (sdo)," *Solar Physics*, vol. 275, no. 1-2, pp. 17–40, 2012.

[21] P. H. Scherrer, J. Schou, R. Bush, A. Kosovichev, R. Bogart, J. Hoeksema, Y. Liu, T. Duvall, J. Zhao, C. Schrijver *et al.*, "The helioseismic and magnetic imager (hmi) investigation for the solar dynamics observatory (sdo)," *Solar Physics*, vol. 275, no. 1-2, pp. 207–227, 2012.

[22] T. Woods, F. Eparvier, R. Hock, A. Jones, D. Woodraska, D. Judge, L. Didkovsky, J. Lean, J. Mariska, H. Warren *et al.*, "Extreme ultraviolet variability experiment (eve) on the solar dynamics observatory (sdo): Overview of science objectives, instrument design, data products, and model developments," *Solar Physics*, vol. 275, no. 1-2, pp. 115–143, 2012.

[23] L. E. Boucheron, J. Grajeda, T. Vincent, and E. Wuest, "Active region magnetogram image dataset for studies of space weather," *In preparation*, 2021.

[24] S. J. Orfanidis, *Introduction to signal processing*. Prentice-Hall, Inc., 1995.

[25] Y. Liu, J. Hoeksema, P. Scherrer, J. Schou, S. Couvidat, R. Bush, T. Duvall, K. Hayashi, X. Sun, and X. Zhao, "Comparison of line-of-sight magnetograms taken by the solar dynamics observatory/helioseismic and magnetic imager and solar and heliospheric observatory/michelson doppler imager," *Solar Physics*, vol. 279, no. 1, pp. 295–316, 2012.

[26] P. Boerner, C. Edwards, J. Lemen, A. Rausch, C. Schrijver, R. Shine, L. Shing, R. Stern, T. Tarbell, C. J. Wolfson *et al.*, "Initial calibration of the atmospheric imaging assembly (aia) on the solar dynamics observatory (sdo)," in *The Solar Dynamics Observatory*. Springer, 2011, pp. 41–66.

[27] P. Charbonneau and O. R. White, "Hale's sunspot polarity law: High altitude observatory," Apr 1995. [Online]. Available: https://www2.hao.ucar.edu/Education/Sun/hales-sunspot-polarity-law

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[29] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[30] A. Al-Ghraibah, L. E. Boucheron, and R. T. McAteer, "An automated classification approach to ranking photospheric proxies of magnetic energy build-up," *Astronomy & Astrophysics*, vol. 579, June 2015.

APPENDIX I: Additional Results

This appendix contains figures and short discussions mentioned in the main thesis body but deemed not especially relevant to the main discussion and analysis.

Clipped AIA Data

The 10-AR experiment and the 52-AR experiment using log10-scaled AIA data were conducted again using AIA data that had been clipped to the range [1, 2550] prior to other preprocessing; Table 6 shows the results of this. Although we assumed that this clipping would improve performance by preserving the relative brightness between different AIA images, we in fact saw a noticeable decrease in SSIM and general image quality and larger performance gaps between SameAR and DiffAR tests. NRMSE, however, was mostly comparable to that seen in the original experiments.

Table 6: Results of experiments using clipped AIA data

Experiment	**NRMSE**		**SSIM**	
	SameAR	DiffAR	SameAR	DiffAR
10-AR	0.0484	0.0521	0.5228	0.4961
10-AR, clipped AIA	0.0486	0.0550	0.4462	0.4362
52-AR (log10/abs)	0.0504	0.0512	0.4987	0.5024
52-AR (log10/abs), clipped AIA	0.0500	0.0533	0.3919	0.3761

The maximum value of 2550 was chosen based on $\mu_{max} + 3\sigma$, where μ_{max} is the mean of the maximum values present in the AIA dataset and σ is the mean standard deviation of the images. The 52-AR dataset has $\mu_{max} = 2205$ and $\sigma = 69$. This gives a suitable maximum value of 2412, which we increased to 2550 in order to simplify noise calculations. 23% of images had a higher maximum value than 2550 and thus were clipped– the performance decrease may be due in part to a loss of information from a higher percentage of images being affected by clipping (versus HMI data, where only 0.005% of images were clipped).

Also note that σ above represents average standard deviation of the image values in general, not the standard deviation of the maximum values across all images– that value, σ_{max}, was equal to 2837. This indicates a very large spread of maximum values in the AIA images, and the overall maximum value across all images was $\sim$ 100000. Another possible contributor to the performance decrease could be that those 23% of images with higher maximum values than 2550 pushed μ_{max} far higher than the maximum value present in many of the other 77% of images. When those 77% are log10 scaled and then normalized assuming

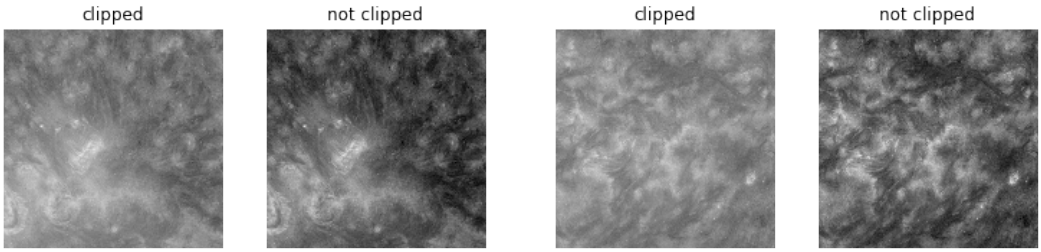


Figure 53: Comparisons between clipped and non-clipped AIA images

a maximum value significantly larger than their actual maximum value, the values in the final `uint8` image end up occupying a much smaller range than they would if they were normalized according to each image's individual original range.

In other words, many of the clipped AIA images had lower contrast than their non-clipped counterparts if log10 scaled (and were darker than the non-clipped versions if they were instead linearly scaled, though no experiments involving clipped linearly scaled data were completed). This may have made it more difficult for the network to extract relevant information from them. Figure 53 shows some examples of this. Figure 54 shows example DiffAR predictions from the 52-AR experiment using clipped AIA data– we see that these are generally blurrier and poorer in structure than the experiment that used AIA data that was not clipped (Figure 18).

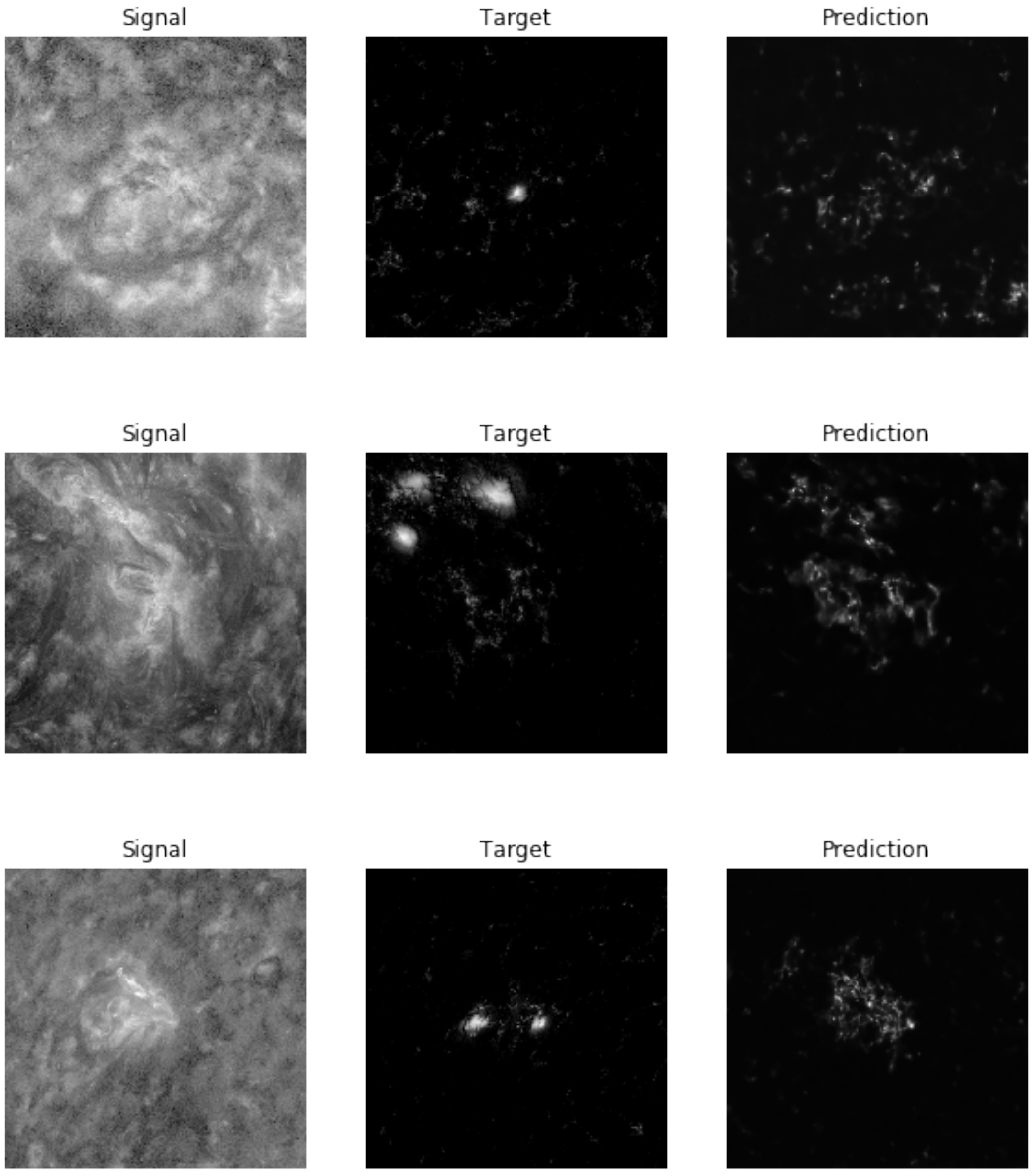


Figure 54: Target vs. predicted magnetograms using clipped AIA data as signal

Heatmaps for Signed HMI Experiments

Figures 55 and 56 show target versus predicted flux heatmaps for the experiment using HMI data from only the northern hemisphere as target, and linearly scaled AIA data as signal. We can see that localization accuracy in this experiment is not high enough for the heatmap to show meaningful structure, even when looking at subsets of only the highest flux values. This was the case for the other experiments using signed HMI data as well, so analysis of these heatmaps was not used in the main discussion.

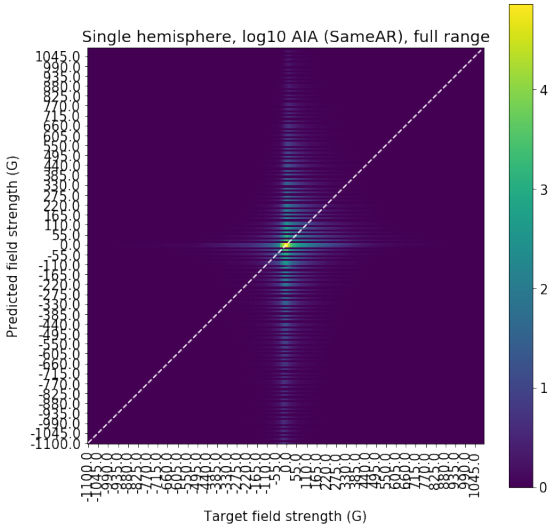


Figure 55: Sample heatmap (full flux range) for experiments using signed HMI data

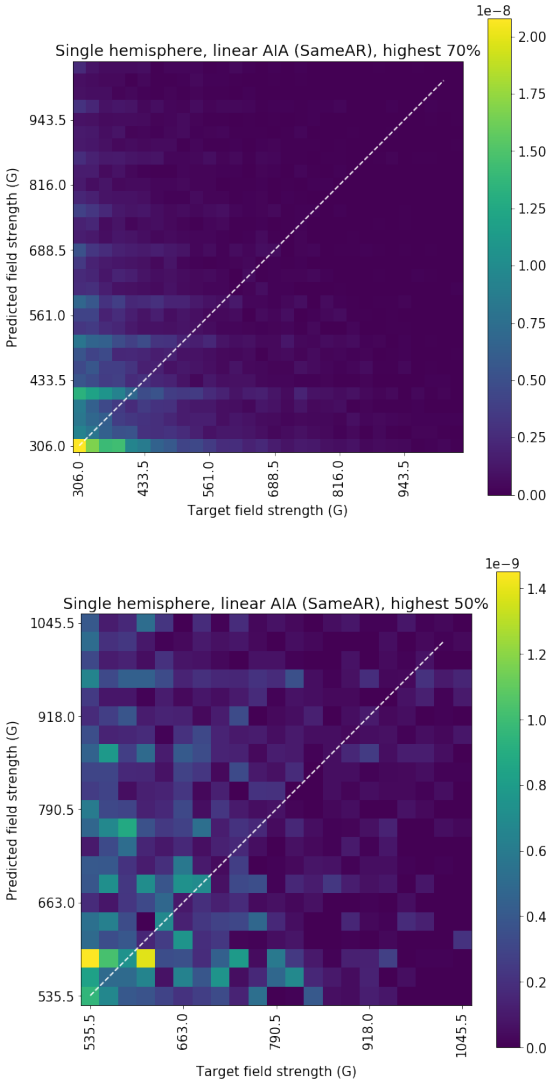


Figure 56: Sample heatmaps (highest 70 and 50%) for experiments using signed HMI data