Research in Astronomy and Astrophysics, 23:085004 (11pp), 2023 August https://doi.org/10.1088/1674-4527/acd595

Attention-Based Deep Learning Model for Image Desaturation of SDO/AIA

Xinze Zhang[1,2], Long Xu[1,3], Zhixiang Ren[3], Xuexin Yu[4], and Jia Li[3,5]
[1] State Key Laboratory of Space Weather, National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China; lxu@nao.cas.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Peng Cheng Laboratory, Shenzhen 518000, China
[4] Department of Automation, Tsinghua University, Beijing 100084, China
[5] State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China
Received 2023 April 4; revised 2023 May 1; accepted 2023 May 7; published 2023 July 5

Abstract

The Atmospheric Imaging Assembly (AIA) onboard the Solar Dynamics Observatory (SDO) captures full-disk solar images in seven extreme ultraviolet wave bands. As a violent solar flare occurs, incoming photoflux may exceed the threshold of an optical imaging system, resulting in regional saturation/overexposure of images. Fortunately, the lost signal can be partially retrieved from non-local unsaturated regions of an image according to scattering and diffraction principle, which is well consistent with the attention mechanism in deep learning. Thus, an attention augmented convolutional neural network (AANet) is proposed to perform image desaturation of SDO/AIA in this paper. It is built on a U-Net backbone network with partial convolution and adversarial learning. In addition, a lightweight attention model, namely criss-cross attention, is embedded between each two convolution layers to enhance the backbone network. Experimental results validate the superiority of the proposed AANet beyond state-of-the-arts from both quantitative and qualitative comparisons.

Key words: techniques: image processing – Sun: atmosphere – Sun: flares

1. Introduction

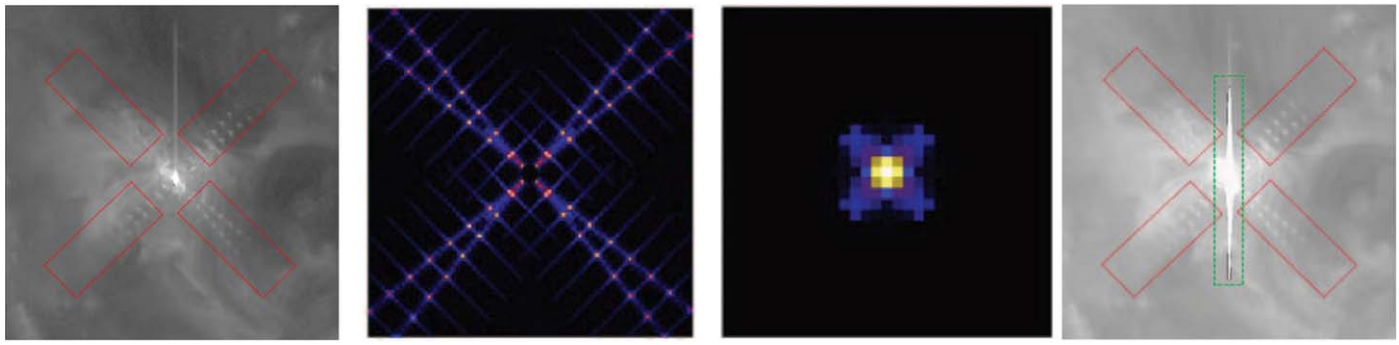
The Atmospheric Imaging Assembly (AIA) (Lemen et al. 2012) onboard the Solar Dynamics Observatory (SDO) (Pesnell et al. 2012) captures full-disk solar images over seven extreme ultraviolet (EUV) (94 Å, 131 Å, 171 Å, 193 Å, 211 Å, 304 Å, 335 Å) wave bands with a temporal cadence of 12 s and an angular resolution of 0.″5, providing an unprecedented high-definition observation of the solar atmosphere, especially fine-grained dynamic evolution of solar activities.

However, in case of big solar flares, the incoming photoflux may exceed the threshold of the charge-coupled device (CCD) of SDO/AIA, resulting in saturation/overexposure of the flare's core region. The imaging system of SDO/AIA is characterized by these two processes, diffraction and diffusion which can be more or less explained by Figure 1. The diffraction replicates the core peak to generate diffraction fringes as shown in Figure 1(a), which is resulted by the convolution with the point-spread function (PSF) of SDO/AIA as shown in Figure 1(b). The diffusion causes the diffusion effect in local area when the input signal goes through the central part of the PSF as shown in Figure 1(c). The diffraction artifact would become apparent against the background in case of high intensity in image core frequently inducing saturation. More precisely, the saturation consists of the primary saturation and blooming/secondary saturation due to two different reasons. The former refers to the fact that the CCD pixels cannot accommodate additional charge for incoming photoflux while the latter names the fact that the primary saturation causes charge to spill into their neighbors. The overall effect of saturation is to flatten and threshold the brightest core of an image anisotropically (north–south direction) as shown in Figure 1(d).

The imaging process described above is formulated as the convolution between incoming photon flux and PSF,

$$I = (A_c + A_d) \otimes f = A_c \otimes f + A_d \otimes f, \quad (1)$$

where f and I are the incoming photon flux and the recorded signal by SDO/AIA respectively, A_c and A_d represent the diffusion component and diffraction component of the PSF respectively, $\otimes$ represents the convolution operator. A_c and A_d are illustrated in Figures 1(c) and (b), where A_c is a core peak, and A_d is the replications of the core peak. It can be observed that diffraction fringes which is the result of $A_d \otimes f$ in Figure 1(a). The effect of diffraction fringes comes from a regular, peripheral diffraction pattern of varying intensity in A_d as shown in Figure 1(b). In particular, diffraction fringes would become more apparent against the background when increasing the peak of f. The other term $A_c \otimes f$ in (1) results in image saturation which is split into the primary saturation and the second saturation/blooming. As discussed in (Guastavino et al. 2019), the blooming cannot be restored while the primary saturation may present in the diffraction fringes due to diffraction effect. In detail, the signal f is coherently and linearly scattered to other regions presenting as diffraction fringes due to diffraction (Guastavino et al. 2019) given by



(a) Diffraction effect (b) Point spread function (PSF) (c) Central part of PSF (d) Anisotropic saturation

Figure 1. An example of diffraction and diffusion effects, where diffraction fringes scatter to other regions beyond the core peak, while diffusion effect causes some degree of blur in the core peak due to the central part of the PSF.

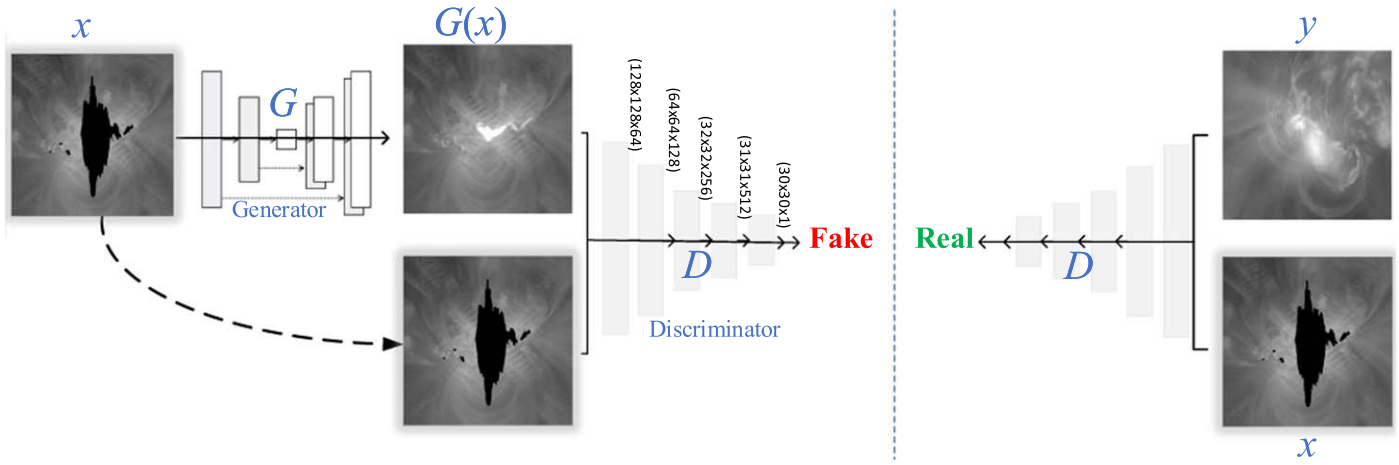


Figure 2. The overview of the proposed model.

$A_d \otimes f$ as shown in Figure 1(a). Thus, the lost signal can be partially retrieved from diffraction fringes (Schwartz et al. 2014; Torre et al. 2015).

The recovery of lost signal from degraded signal is traditionally described as an inverse problem. To resolve an inverse problem, an extra constraint is additionally required. Usually, the extra constraint is given by typical image priors, like sparsity, non-local and total variation. This processing is well known as the regularization method which optimizes both data fidelity term and regularization (prior) term. In DESAT (Schwartz et al. 2015), lost signal recovery in saturated regions was first formulated into an inverse diffraction issue as

$$I_d = A_d \otimes f + B_d, \tag{2}$$

where I_d is the known recorded image in diffraction regions, B_d is the unknown saturated image background related to diffraction fringes. Then, the regularization method was employed to resolve (2) to recover f. In Schwartz et al. (2015), B_d is estimated from the interpolation of two neighbor unsaturated images which are provided by short-time exposure of SDO/AIA. This short-time exposure of SDO/AIA can be automatically triggered once solar flare occurs. However, in case of large solar flare, the neighbor images of short-time exposure are also saturated, resulting in failure of DESAT. To address this problem, a Sparsity-Enhancing DESAT (SE-DESAT) Guastavino et al. (2019) was proposed to estimate B_d from only current image instead of its neighbors. Nevertheless, desaturated result is limited by the segmentation of diffraction fringes and primary saturation regions and the estimation of background. In addition, the blooming regions cannot in principle be restored in both DESAT and SE-DESAT.

Inspired by great success of deep learning, Mask-Pix2Pix (Zhao et al. 2019), PCGAN (Yu et al. 2021) and MCNet (Yu et al. 2022) have been proposed to desaturate solar images in our previous efforts. Different from DESAT (Schwartz et al.

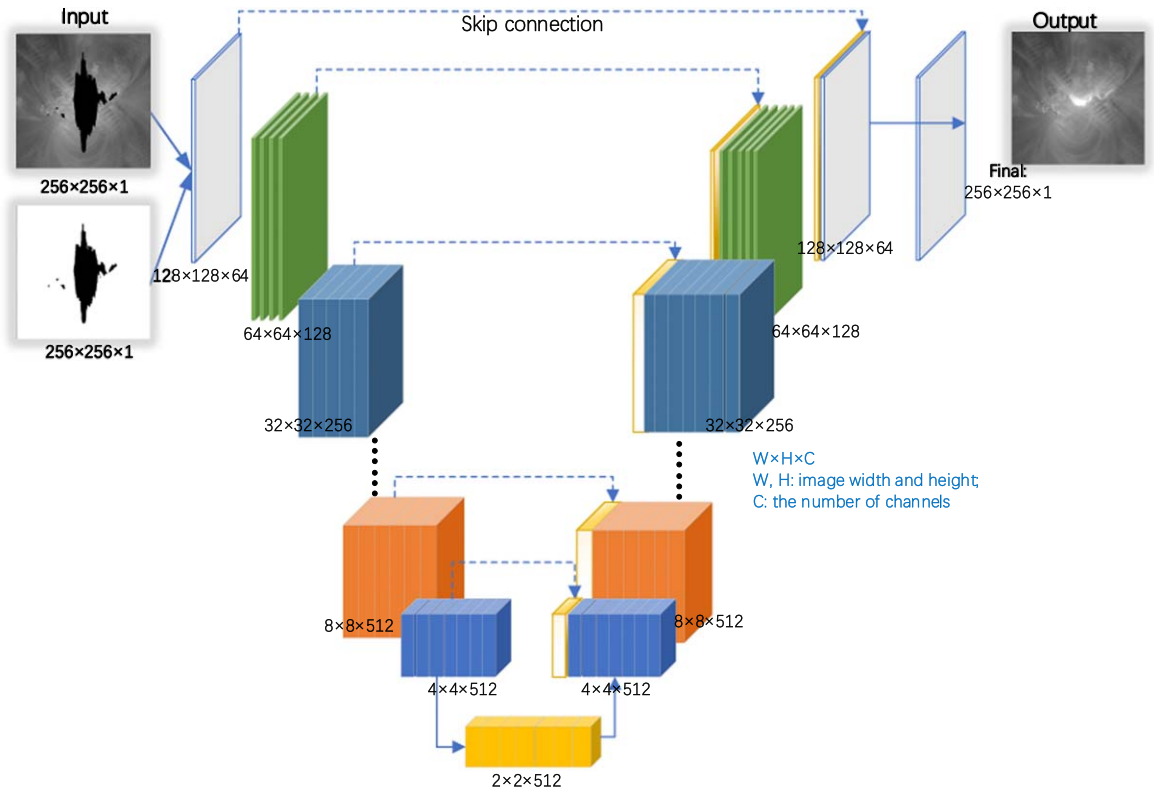


Figure 3. The generator of the AANet, which learns a mapping from I_m and I_d to I_{gt}. The discriminator supervises the learning process of the generator by an adversarial loss for discriminating fake $\{I_d, I_g, I_{gg}\}$ and real $\{I_d, I_{gt}, I_{gtg}\}$ pairs. It finally minimizes the distance between two probability distributions of $\{I_g\}$ and $\{I_{gt}\}$.

2015) and SE-DESAT (Guastavino et al. 2019) which explicitly model the recovery of saturation (desaturation) as an inverse diffraction, our models implicitly describe the desaturation as an image inpainting task with the help of deep learning. In addition, relative to DESAT and SE-DESAT, our models could compensate both the primary and second saturations with advanced image generation techniques of deep learning. Moreover, it is not necessary to segment primary saturation and blooming, and estimate background from the image superposed diffraction fringes, which were however two big challenges for both DESAT and SE-DESAT. Besides, partial convolution (PC) (Liu et al. 2018) was used in PCGAN (Yu et al. 2021) instead of standard convolution for processing invalid pixels within a convolution block.

As discussed in Equation (1) and Figure 1, a peripheral regular diffraction pattern A_d replicates the core peak to generate diffraction fringes distributed outside of the core peak. These diffraction fringes carry information about the core region to scatter outside the core region. They can be utilized to restore the saturated region through an inverse process of diffraction, which has been successfully formulated by convolutional neural networks (CNNs) in our previous efforts (Yu et al. 2021, 2022). However, due to the small receptive field of CNN, we cannot efficiently unitize the diffraction fringes spread throughout the entire image, result in a compromise of desaturation. In this work, considering the non-local property of diffraction fringes, an lightweight attention module, namely criss-cross attention (Huang et al. 2019), is employed to enhance CNNs to exploit global diffraction fringes for desaturation. This attention model has the receptive field of the entire image, so it can efficiently synthesize the information of the entire image through different weights.

The rest of paper is organized as follows. Section 2 introduces the network architecture, convolutions and loss functions of the proposed AANet in details. Experimental results are provided in Section 3. Conclusion and discussion are given in Section 4.

2. Method

The desaturation problem has been formulated as an inverse problem as given in Equation (2). The traditional solution was through regularization method (Guastavino et al. 2019) which was however challenged by the decision of the primary saturation region and the estimation of the background signal. Deep learning has been widely acknowledged as a universal approximator (Cybenko 1989; Hornik 1991), which has

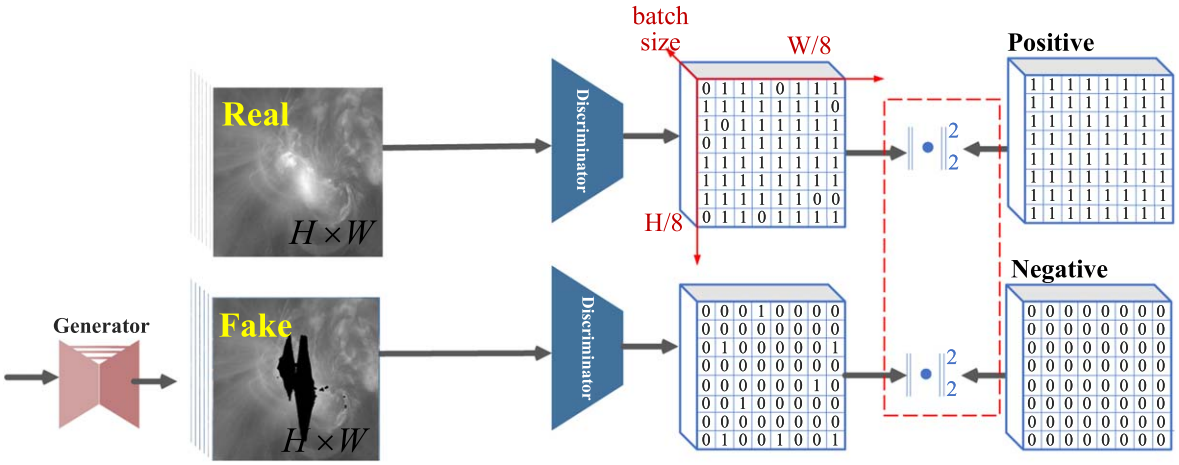


Figure 4. Diagram of PatchGAN for discriminating "real" and "fake."

achieved a big success in a variety of image processing tasks, such as image denoising, enhancement, super-resolution, inpainting, deconvolution and etc. In this section, an attention augmented convolutional neural network (AANet) is constructed to exploit attention mechanism for image desaturation. First, the network architecture of the proposed AANet is presented in details. Second, the loss function is presented and discussed in details for the optimization of the proposed model.

2.1. Network Architecture

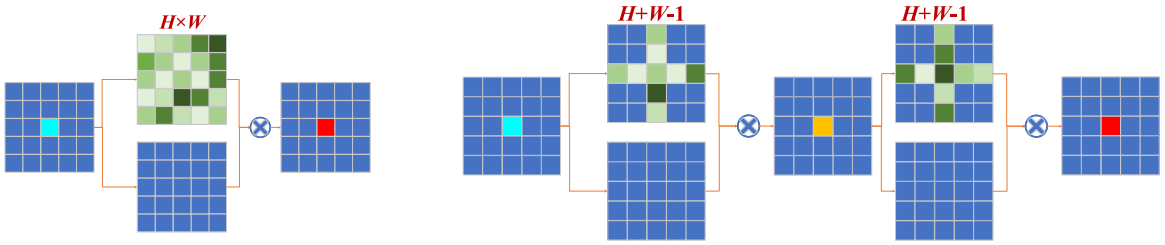
The overall network of the proposed model is a generative adversarial network (GAN) as shown in Figure 2, consisting of a generator and a discriminator. The generator is a U-Net which architecture is shown in Figure 3. It consists of an encoder of eight convolutional layers and a decoder of eight deconvolutional layers. The basic modules of the generator include criss-cross attention (Huang et al. 2019), PC, regional composite normalization (RCN) (Wang et al. 2021), and ReLU/LeakyReLU. They are stacked repeatedly in the generator. In addition, the skip connection connects the encoder and the decoder at each layer. The detailed parameters are listed on the left side of network architecture, where the name of each module and volume of each module are provided. Moreover, the PC instead of the standard convolution is employed for processing invalid pixels in a convolution block. As illustrated in Figure 3, a mask image is provided to indicate "normal" and "saturated" pixels in an image for guiding the PC. In the encoder, invalid region gradually becomes smaller along with the mask updating in Liu et al. (2018). Finally, all pixels become valid, and the mask image converges to an all-ones matrix. During convolution process, the mask image is provided to image branch to guide extracting image features.

The discriminator is a general CNN consisting of convolution layers. Specifically, it is a PatchGAN (Isola et al. 2017; Zhu et al. 2017) which means that each image is divided into small patches (e.g., 8×8) rather than a whole for discriminating real/false (real: positive or false: negative). The output of discriminator is a $H/8 \times W/8$ binary matrix where "0" and "1" indicate the probability of each patch being real or false. From Figure 4, mean square error (MSE) is computed to measure the loss of the discriminator.

2.2. Attention

Concerning image inpainting, attention mechanism is of great importance for exploring global/contextual information in an image, which is equivalent to exploring non-local prior in traditional image processing. In the literature, there have been many attention models, including non-local (Wang et al. 2018), SENet (Hu et al. 2020), GCNet (Cao et al. 2019), CCNet (Huang et al. 2019) and transformer (Vaswani et al.v 2017).

In this work, a lightweight attention, namely criss-cross attention (CCNet) (Huang et al. 2019) is employed for low computational complexity. It exploits contextual information to further augment image convolutional features for recovery of saturated region of an image. The diagram of a CCNet is given in Figure 5, where an input image is first passed through convolution layers to produce feature maps. Then, these feature maps are fed to a criss-cross attention module for further enhancing image features, producing new feature maps. The criss-cross attention aggregates contextual information for each pixel in its criss-cross path, which means only the pixels in the



(a) Non-local attention module (b) Criss-cross attention module

Figure 5. Diagrams of non-local and criss-cross attention modules. (Computational complexity is $\mathcal{O}(H \times W)(H \times W)$) and $\mathcal{O}(H \times W) \times (H + W - 1)$, respectively.)

same row and column as the current pixel are concerned. It is worth pointing out that two times of aggregation of criss-cross path are implemented in a CCNet. Thus, for each pixel, a CCNet actually aggregates the contextual information of all pixels of an image block. Such a recurrent aggregation is named as recurrent criss-cross attention (RCCA). To explore global contextual information over local feature representation, the non-local attention module (Wang et al. 2018) generates a dense attention map size of $H \times W$ for each pixel as shown in Figure 5(a), while the criss-cross attention module generates a sparse attention map only size of $H + W - 1$. However, going through two criss-cross operations, each pixel of the final output feature map can gather contextual information from all pixels as shown in Figure 5(b).

2.3. Convolutions

Using deep learning, image restoration is accomplished by referring to the degraded image itself and the statistical distribution of massive unimpaired images. In this work, the desaturation of solar image is regarded as an image inpainting task. Solar images are impaired by saturated regions/holes as big flares happen. In image inpainting, convolution across intersection region between valid pixels and invalid pixels need to be designed specifically. First, invalid pixels should be excluded from standard convolution, namely partial convolution (PC) (Liu et al. 2018). Second, the deviation caused by PC should be compensated so that output energy of PC remains the same relative to standard convolution. Given the input degraded image/feature map x, convolution weight w and bias b, standard convolution is described as

$$y(i, j) = b + \sum_{u,v \in \mathcal{R}} w(i, j) \cdot x(i + u, j + v), \tag{3}$$

where $y(i, j)$ denotes the (i,j)-th position of output feature map y, $\mathcal{R}$ confines the receptive field of convolution. For example, 3×3 receptive field is formulated as $\mathcal{R} = \{(-1, -1), (-1, 0),...,(0, 1), (1, 1)\}$. When receptive field of convolution slides across the boundary of impaired hole, both valid and invalid pixels participate in convolution operation.

To solve this issue, the PC (Liu et al. 2018) is introduced, which is described as

$$y(i, j) = \begin{cases} b + \dfrac{\sum_{u,v \in \mathcal{R}} \mathbf{1}}{\sum_{u,v \in \mathcal{R}} m(i + u, j + v)} \sum_{u,v \in \mathcal{R}} w(i, j) \cdot x(i + u, j + v) \cdot m(i + u, j + v), \\ \quad if \sum_{u,v \in \mathcal{R}} m(i + u, j + v) > 0 \\ 0, \qquad \text{otherwise,} \end{cases} \tag{4}$$

$$m(i, j) = \begin{cases} 1, & if \sum_{u,v \in \mathcal{R}} m(i + u, j + v) > 0 \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where m represents the mask image where "0" stands for saturated pixels and "1" stands for normal pixels, it is updated by Equation (5) for each operation. The symbol 1 is a constant matrix with all entries equal to 1, with the same size as m. From Equations (4) and (5), PC only depends on valid pixels by introducing a mask to exclude invalid pixels in convolution, while the deviation caused by invalid pixels is calibrated by scaling the output of PC, and the scaling factor is proportional to the number of valid pixels in receptive field.

2.4. Loss Functions

To optimize a neural network for image generation, a hybrid loss function was usually raised. It includes both pixel-level

and feature-level image losses, providing image high-fidelity and photorealistic effect respectively. The pixel-level loss is represented by the L_1 norm and L_2 norm, i.e., mean absolute error (MAE) and mean square error (MSE) functions. It measures the pixel-level difference between generated image and ground-truth in supervised learning. Inspired by the classical image priors in image processing, image gradient and image smoothness are of great importance to the perception of human visual system. Thus, two additional losses, namely gradient loss (Ma et al. 2020) and total variation loss (Johnson et al. 2016) are included in the loss function of the proposed model. Relative to pixel-level loss, the feature-level loss could well describe the photorealistic property of an image, but ignores the pixel-level difference. In this work, perceptual loss (Johnson et al. 2016), and style loss (Gatys et al. 2016) are employed to measure feature-level difference between generated image and ground-truth. The last but not the least, an adversarial loss (Mao et al. 2017) is included in the loss function, which optimizes a generator through zero-sum game against a discriminator (Goodfellow et al. 2014).

Let I_d be the input degraded image, I_m the initial binary mask image, I_g the generated image, and $I_{\rm gt}$ the ground-truth, the pixel-level loss given by L_1 norm is defined as

$$\mathcal{L}_{\rm rec}(G) = \lambda_h \|(1 - I_m) \odot (I_g - I_{\rm gt})\|_1 + \lambda_v \|I_m \odot (I_g - I_{\rm gt})\|_1, \quad (6)$$

where $\|\cdot\|_1$ denotes L_1 norm, λ_h and λ_v are two weights for combining recovered saturated region and normal region. They are empirically set to 100 and 10, indicating more weight is allocated to the recovered saturated region in Equation (6).

The gradient loss (Ma et al. 2020) is adopted to ensure that the generated image has sharp structures and edges of objects, which is defined as the L_1 loss over image gradient map as

$$\mathcal{L}_{\rm gra}(G) = \lambda'_h \|(1 - I_m) \odot (\nabla(I_g) - \nabla(I_{\rm gt}))\|_1 + \lambda'_v \|I_m \odot (\nabla(I_g) - \nabla(I_{\rm gt}))\|_1, \quad (7)$$

where ∇ represents a gradient operator (Ma et al. 2020) to compute image gradient, λ'_h and λ'_v are two weights (empirically set to 300 and 10), assigning different weights to saturated regions and normal regions.

Total variation loss (Johnson et al. 2016) is included to ensure image smoothness, especially around the boundary between normal and saturated regions. It is defined as

$$\mathcal{L}_{tv}(G) = \sum_{(i,j)\in P,(i,j+1)\in P} \|I_c^{(i,j)} - I_c^{(i,j+1)})\|_1 + \sum_{(i+1,j)\in P,(i,j)\in P} \|I_c^{(i+1,j)} - I_c^{(i,j)})\|_1, \quad (8)$$

where P indicates the region connecting saturated and normal regions.

The adversarial loss (Mao et al. 2017) is adopted to ensure photorealistic effect of generated image from feature-level, which is formulated as

$$\mathcal{L}_{\rm adv}(G, D) = \mathbb{E}_{(I_d,I_{\rm gt})}[\log D(I_d, I_{\rm gt}, \nabla(I_{\rm gt}))] + \mathbb{E}_{(I_d,I_m)}[\log(1 - D(I_d, G(I_d, I_m), \nabla(G(I_d, I_m))))]. \quad (9)$$

The perceptual loss (Johnson et al. 2016) is adopted to capture high-level semantic information and alleviate grid-shaped artifacts in recovered regions (Liu et al. 2018), which is formulated as

$$\mathcal{L}_{perc}(G) = \sum_{i=1}^{T} \|\Psi_i(I_g) - \Psi_i(I_{\rm gt})\|_1 + \sum_{i=1}^{T} \|\Psi_i(I_c) - \Psi_i(I_{\rm gt})\|_1, \quad (10)$$

where $I_c = (1 - I_m) \odot I_g + I_m \odot I_{\rm gt}$ indicates the combination of the recovered region and the normal region extracted directly from the ground-truth. It can be seen that the perceptual loss computes L_1 norm in feature domain for I_g and I_c, respectively, where Ψ_i represents the feature map of the i-th pooling layer of VGG-16 (Simonyan & Zisserman 2015). In this work, the first three pooling layers ($T = 3$) are used in Equation (10).

The style loss (Gatys et al. 2016) has been proved to be effective for capturing image semantic information, which first computes Gram matrix for each feature map of VGG-16, and then calculates L_1 norm of Gram matrix. Therefore, it is defined as

$$\mathcal{L}_{sty}(G) = \sum_{i=1}^{T} \|K_i((\Psi_i(I_g))^T(\Psi_i(I_g)) - (\Psi_i(I_{\rm gt}))^T(\Psi_i(I_{\rm gt})))\|_1 + \sum_{i=1}^{T} \|K_i((\Psi_i(I_c))^T(\Psi_i(I_c)) - (\Psi_i(I_{gt}))^T(\Psi_i(I_{\rm gt})))\|_1, \quad (11)$$

where K_i is a weight for scaling, which is given by $1/H_iW_iC_i$ for the i-th layer of VGG-16, $\Psi_i(I)$ represents the operator of Gram matrix, which outputs a feature map size of $H_i \times W_i \times C_i$.

3. Experimental Results

For evaluating the proposed AANet, experiments are performed to first compare our model with two state-of-the-art desaturation methods, PCGAN (Yu et al. 2021) and MCNet (Yu et al. 2022). Then, the effectiveness of criss-cross attention module is verified by an ablation study. The source code of AANet can be accessed via GitHub (https://github.com/filterbank/AANet).

3.1. Dataset

For training deep learning models, a new large-scale data set beyond the previous one (Yu et al. 2021) is established in this work. In this new data set, raw data of 14 bit-FITS format instead of "png" images are included, so that the high fidelity

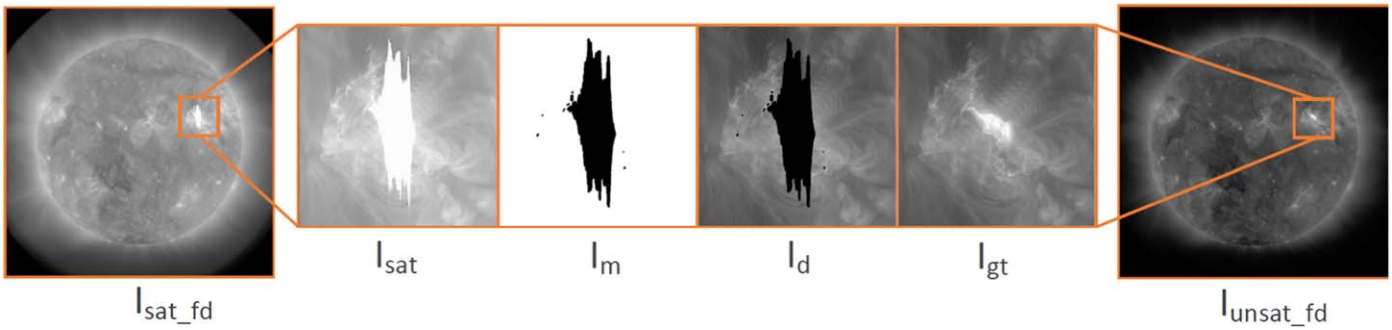


Figure 6. An sample in desaturation data set established by Yu et al. (2021), which is composed of four images: $I_{\rm sat}$, $I_{\rm gt}$, I_m and I_d. $I_{\rm sat}$ is a saturated image and $I_{\rm gt}$ is the nearest unsaturated image of short-time exposure. I_m is a binary mask which indicates saturated and unsaturated pixels of $I_{\rm sat}$ by 1 and 0, respectively. I_d is the simulated degraded image which is obtained by $I_{\rm gt} \odot I_m$.

Table 1
Quantitative Comparison of PSNR/SSIM (Mean/std) between two AANet Models (PCGAN+CC and MCNet+CC) and Two Benchmarks (PCGAN, MCNet)

Methods	PSNR	SSIM
PCGAN	29.21+/−0.1345	0.8079+/−0.0632
PCGAN+CC	30.21+/−0.2066	0.8445+/−0.0671
MCNet	30.40+/−0.1913	0.8497+/−0.0592
MCNet+CC	30.94+/−0.2078	0.8596+/−0.0707

of scientific computing and physical plausibility of computing results can be guaranteed. Each sample of the data set consists of a ground-truth given by the image of short-time exposure without overexposure, a mask image labeling saturated pixels provided by the image of long-time exposure and a manually overexposed image flatten by a preset threshold. We gather M-class and X-class solar flare data at 193 Å of SDO/AIA (Lemen et al. 2012) from 2010 to 2017. The short time exposure images closest to the overexposure one of long time exposure are taken as the ground-truths. In addition, they are normalized by the time of long time exposure. A sample of the data set is shown in Figure 6, where the saturated image $I_{\rm sat}$ is only used to deduce a more realistic mask (denoted by I_m) of saturated region by imposing a threshold on $I_{\rm sat}$, degraded image I_d results from $I_{\rm gt} \odot I_m$ ($\odot$ represents element-multiplication operator), and $I_{\rm gt}$ is given by the image of short-time exposure closest to the long-time overexposed one. During model training, the triplet $\{I_d, I_m, I_{\rm gt}\}$ is fed to our proposed network to optimize the model parameters. The whole data set contains about 18,700 samples. To train and test the network with multiple splittings of the data set, we split the data set into eight equal portions to alternatively select seven of them for training and the rest for testing. Thus, a mean and standard deviation (STD) value on the performance measures can be provided.

3.2. Implementation Details

We evaluate the AANet on our established data set. It should be pointed out that there are two versions of AANet, one is PCGAN plus criss-cross (CC) attention and the other is MCNet plus CC. In our experiments, we employ the well-known data augmentation techniques to augment the training data set, including randomly cropping input image triplet (degraded image, corresponding mask and ground-truth) from 350×350 to 256×256, and randomly rotating (no rotating, 90°, 180° and 270°) and flipping them. The proposed AANet is implemented by the well-know PyTorch package, trained by a NVIDIA GeForce RTX 3090 GPU with batch size of 28 and epoch number of 200. The convolution weights are initialized by the method in He et al. (2015) and optimized by the ADAM algorithm (Kingma & Ba 2014) with $\beta_1 = 0.500$ and $\beta_2 = 0.999$. The initial learning rate is set to $2e-4$, and then decays half at the 100th and 150th epoch successively.

3.3. Comparisons with State-of-the-Arts

The proposed AANet is compared to the three benchmarks, SE-DESAT (Guastavino et al. 2019), PCGAN (Yu et al. 2021) and MCNet (Yu et al. 2022).

For subjective comparison, seven samples of different saturated sizes are selected from the data set as shown in Figure 7. It can be observed that both the AANet and two benchmarks can recover the whole image well with sharp object edges and rich image structures, but the MCNet (Yu et al. 2022) and AANet have more consistent structure with the ground-truth. The MCNet (Yu et al. 2022) and AANet can generate a finer texture structure while PCGAN has a slight blocking artifacts at the peak of saturation sometimes, indicating the benefits of non-local information for image desaturation. Specifically, the MCNet employs validness migratable convolution (VMC) to exploit non-local information by copying surrounding pixels before convolution operation. While the AANet refers to non-local pixels through

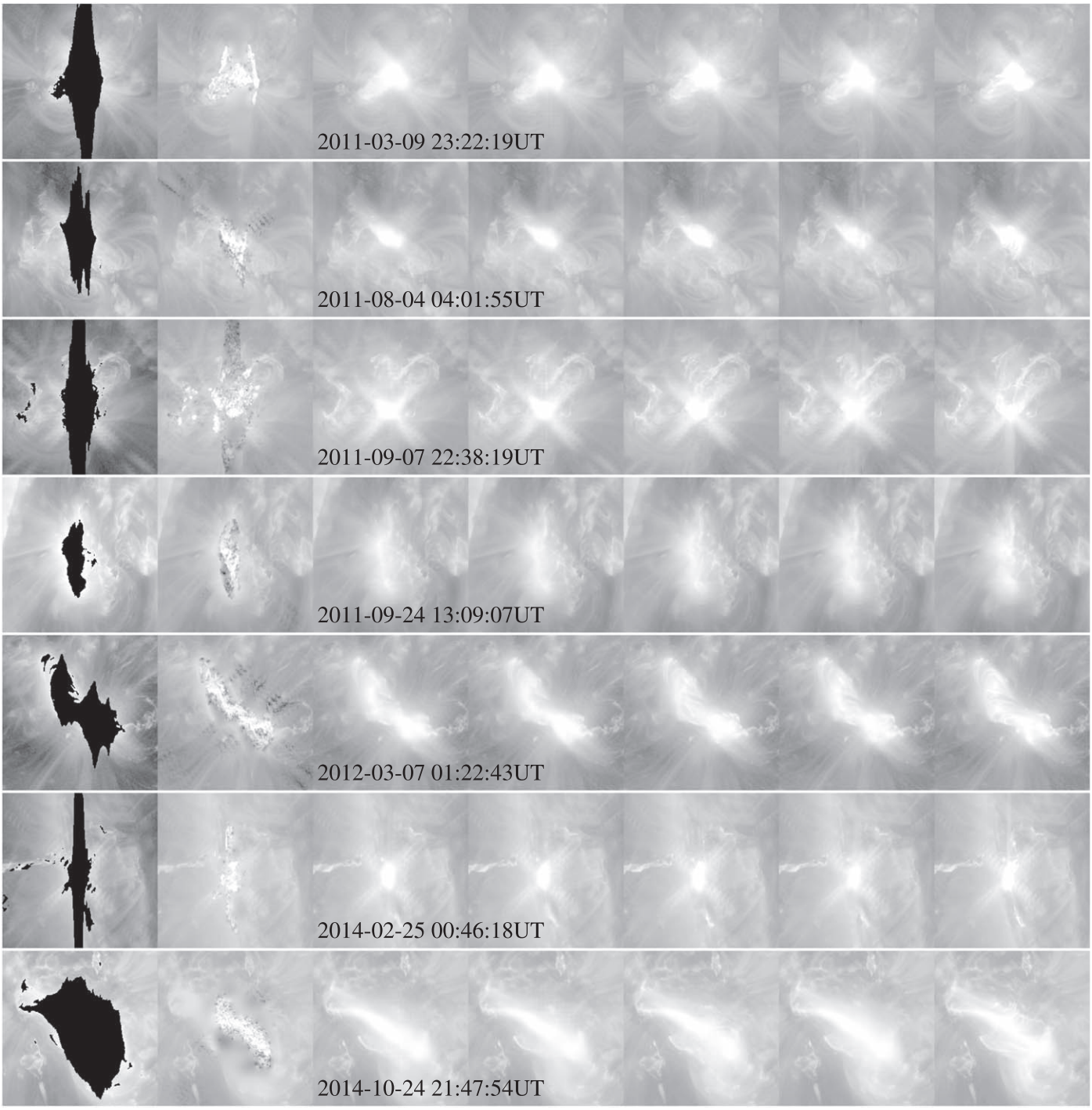


Figure 7. Subjective quality comparison among SE-DESAT (Guastavino et al. 2019), PCGAN (Yu et al. 2021), MCNet (Yu et al. 2022) and AANet ("PCGAN+CC" and "MCNet+CC," where "CC" represents criss-cross).

a lightweight attention module. For objective evaluation, peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) (Wang et al. 2004) are computed over the testing set for each splitting of the data set, and their means and STDs are listed in Table 1. It can be seen that the two AANets outperform the two benchmarks respectively. The two AANets achieve the PSNR improvements of 1.0 dB, 0.54 dB over PCGAN and MCNet respectively. The success of the two AANets is due to the

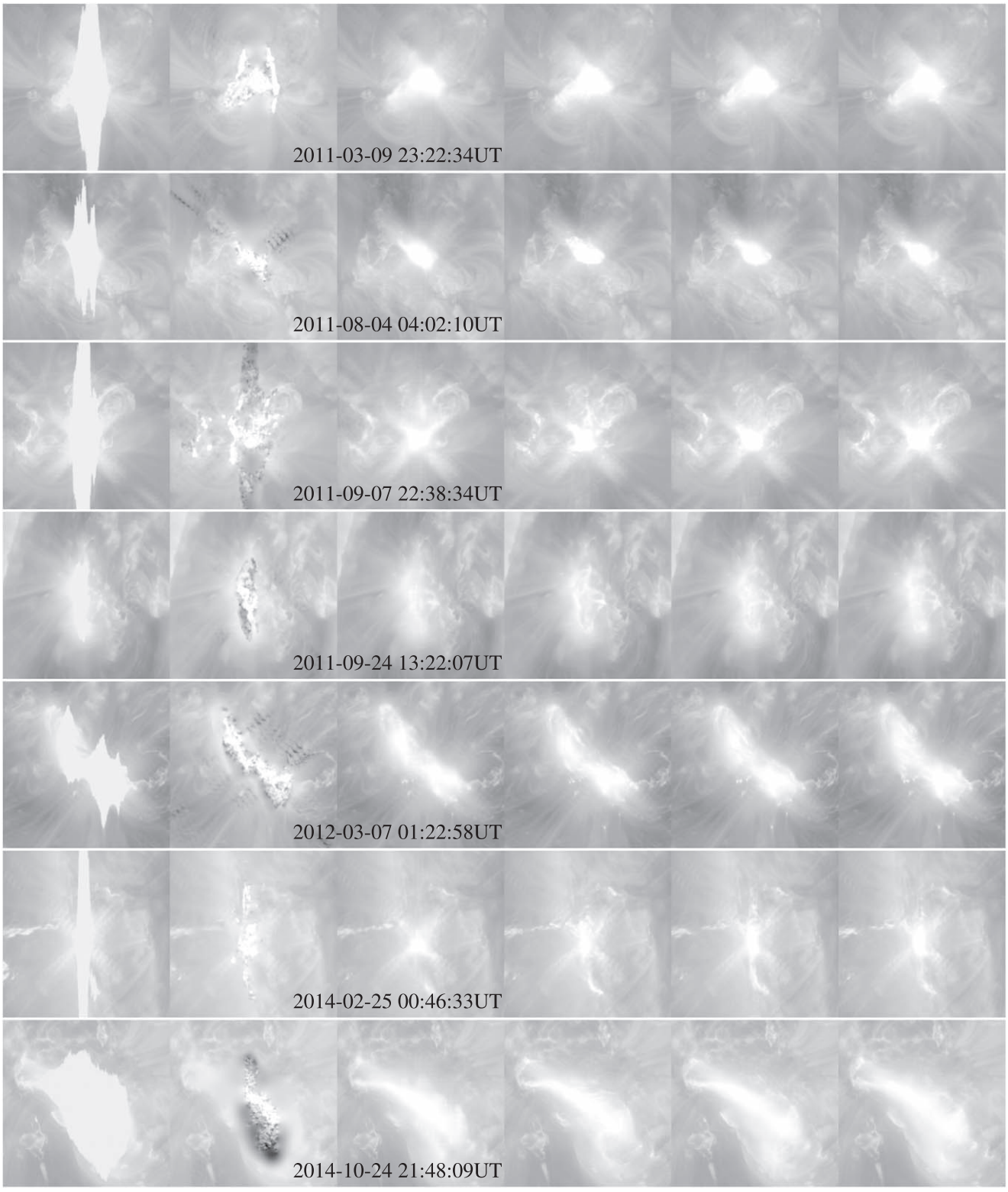


Figure 8. Subjective quality comparison for real saturated images among SE-DESAT (Guastavino et al. 2019), PCGAN Yu et al. (2021), MCNet Yu et al. (2022) and AANet ("PCGAN+CC" and "MCNet+CC," where "CC" represents criss-cross).

Table 2
Quantitative Comparison with State-of-the-art Methods on Testing Set

	PCGAN (Yu et al. 2021)		MCNet (Yu et al. 2022)	
CC Configuration	PSNR	SSIM	PSNR	SSIM
Original	29.21	0.8079	30.40	0.8497
E(1)	29.53	0.8212	30.69	0.8574
E(2)	29.99	0.8418	30.67	0.8506
E(3)	29.64	0.8176	30.67	0.8552
E(4)	29.37	0.8164	30.61	0.8476
E(5)	29.54	0.8345	30.59	0.8530
E(1,2)	29.86	0.8387	30.58	0.8508
E(2,3)	29.57	0.8161	30.67	0.8540
E(1,2,3)	29.39	0.8136	30.75	0.8550
E(1)/D(1)	29.32	0.8143	30.60	0.8521
E(2)/D(2)	30.06	0.8438	30.51	0.8514
E(3)/D(3)	29.57	0.8190	30.59	0.8511
E(1,2)/D(2,1)	30.21	0.8445	30.94	0.8596
E(2,3)/D(3,2)	29.37	0.8155	30.79	0.8539
E(1,2,3)/D(3,2,1)	29.47	0.8149	30.72	0.8528

exploration of non-local information. In addition, the "CC" is more beneficial to PCGAN than MCNet since the latter has explored non-local information through the VMC. We also give the STDs of PSNR and SSIM in Table 1. It can be observed that the STDs of PSNR/SSIM are quite small, indicating the models are stable. In addition, they are comparable among all the tested models.

Comparing AANet with MCNet, they have the different mechanisms of non-local information exploring. The former is through a specially designed convolution operator, namely VMC, and the latter is through the popular attention module. From Table 1, the former is slightly superior to the latter with respect to both PSNR and SSIM. In addition, the attention module is more flexible to explore non-local information. It can be easily embedded in any backbone neural networks, and easily optimized to adapt to the specific task.

We also apply the trained AANet to real saturated images corresponding to long time exposure to evaluate its performance in real scenario. The visual quality comparison among SE-DESAT (Guastavino et al. 2019), PCGAN, MCNet and AANet ("PCGAN+CC," "MCNet+CC") is shown in Figure 8. It can be observed that the saturated region can be repaired more or less by all of the competitive models, where the size of saturated region shrinks obviously. Compared to PCGAN and MCNet, AANet demonstrates more natural and appealing visual effect, especially for large holes.

3.4. Exploring Attention

To study how the attention module is embedded for the best trade-off between efficiency and complexity, we performed a set of experiments by embedding the attention module in the different layers of the network. The backbone network of AANet is a U-Net, where both encoder and decoder can embed the attention module in the different layers. The experiments are listed in Table 2, where "E" and "D" represents encoder and decoder respectively, "CC" represents criss-cross attention, the numbers in parentheses indicate the layers where "CC" module is embedded.

From Table 2, we can conclude: 1) attention module contributes more to the low-level image features, i.e., the shallow layers of a neural network. The first three layers with attention module have obvious improvement; 2) the encoder benefits more than the decoder from attention module since the former can encode original pixel-level information into the compressed features through attention module while the latter can only access the high-level image features. This is consistent with the result of VMC in MCNet (Yu et al. 2022); 3) testing over individual layer and combined layers demonstrate that embedding the attention module into the first two layers achieves a good trade-off between efficiency and complexity.

4. Conclusions and Discussion

This paper proposes a criss-cross attention augmented deep neural network, namely AANet, to repair saturated images of SDO/AIA. The experimental results verify that the attention mechanism really makes a difference in the image desaturation task. Unlike general image denoising or enhancement, the information of saturated regions is completely instead of partially lost in our task. Thus, attention module which borrows information from the non-local regions is significantly important to recover lost information. Compared to the benchmarks, the AANet performs better with respect to both qualitative and quantitative comparisons, which attributes to criss-cross attention module for exploring non-local information efficiently.

Acknowledgments

This work was supported by the National Key R&D Program of China (Nos. 2021YFA1600504 and 2022YFE0133700), the National Natural Science Foundation of China (NSFC) (Nos. 11790305, 11873060 and 11963003).

References

Cao, Y., Xu, J., Lin, S., Wei, F., & Hu, H. 2019, in 2019 IEEE/CVF Int. Conf. Computer Vision Workshops, ICCV Workshops, Seoul, Korea (South) (IEEE), 1971

Cybenko, G. 1989, Math. Control Signals Syst., 2, 303

Gatys, L. A., Ecker, A. S., & Bethge, M. 2016, in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2414

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., et al. 2014, in Proc. 27th International Conf. Neural Information Processing Systems—Volume 2, NIPS'14 (Cambridge, MA: MIT Press), 2672

Guastavino, S., Piana, M., Massone, A. M., Schwartz, R., & Benvenuto, F. 2019, ApJ, 882, 109

He, K., Zhang, X., Ren, S., & Sun, J. 2015, in Proc. IEEE International Conf. Computer Vision, 1026 *(Santiago, Chile)*

Hornik, K. 1991, NN, 4, 251

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. 2020, ITPAM, 42, 2011

Huang, Z., Wang, X., Huang, L., et al. 2019, in 2019 IEEE/CVF Int. Conf. Computer Vision, ICCV, Seoul, South Korea (IEEE), 603

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. 2017, in Proc. IEEE Conf. Computer Vision and Pattern Recognition *(Hawaii, USA)*, 1125

Johnson, J., Alahi, A., & Fei-Fei, L. 2016, in European Conf. Computer Vision *(Amsterdam, Netherlands)* (Berlin: Springer), 694

Kingma, D. P., & Ba, J. 2014, arXiv:1412.6980

Lemen, J. R., Title, A. M., Akin, D. J., et al. 2012, SoPh, 275, 17

Liu, G., Reda, F. A., Shih, K. J., et al. 2018, in Proc. European Conf. Computer Vision (ECCV) *(Munich, Germany)* 85

Ma, C., Rao, Y., Cheng, Y., et al. 2020, in Proc. IEEECVF Conf. Computer Vision and Pattern Recognition *(Seattle, USA)* 7769

Mao, X., Li, Q., Xie, H., et al. 2017, in Proc. IEEE International Conf. Computer Vision, 2794

Pesnell, W. D., Thompson, B. J., & Chamberlin, P. C. 2012, SoPh, 275, 3

Schwartz, R., Torre, G., Massone, A., & Piana, M. 2015, A&C, 13, 117

Schwartz, R. A., Torre, G., & Piana, M. 2014, ApJL, 793, L23

Simonyan, K., & Zisserman, A. 2015, CoRR, arXiv:1409.1556

Torre, G., Schwartz, R. A., Benvenuto, F., Massone, A. M., & Piana, M. 2015, InvPr, 31, 095006

Vaswani, A., Shazeer, N., Parmar, N., et al. 2017, in Annual Conf. Neural Information Processing Systems *(Long Beach, CA)* ed. I. Guyon et al., 5998

Wang, N., Zhang, Y., & Zhang, L. 2021, TIP, 30, 1784

Wang, X., Girshick, R. B., Gupta, A., & He, K. 2018, in 2018 IEEE Conf. Computer Vision and Pattern Recognition *(Salt Lake City, USA)*, 7794

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. 2004, ITIP, 13, 600

Yu, X., Xu, L., Ren, Z., Zhao, D., & Sun, W. 2022, RAA, 22, 065009

Yu, X., Xu, L., & Yan, Y. 2021, SoPh, 296, 1

Zhao, D., Xu, L., Chen, L., Yan, Y., & Duan, L.-Y. 2019, AdAst, 2019, 5343254

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. 2017, in Proc. IEEE international Conf. Computer Vision *(Venice, Italy)*, 2223