



Institute for Information
and Communication Technologies,
Electronics and Applied Mathematics

Convolutional Neural Networks for the Identification, Characterization and Modality Translation of Solar Active Regions from Ground-based Observations

Niels Sayez

Thesis submitted in partial fulfillment
of the requirements for the degree of
Ph.D in Engineering Sciences

Dissertation committee:

Prof. Christophe De Vleeschouwer (UCLouvain, advisor)
Dr. Véronique Delouille (Royal Observatory of Belgium, advisor)
Prof. Laurent Jacques (UCLouvain, secretary)
Prof. Vincent Barra (Université Clermont Auvergne, France)
Prof. Mark Butala (University of Glasgow, Scotland)
Prof. Laurent Francis (UCLouvain, chairman)

May 2025.

— *Neurons¹ ... together ... strong.*

Caesar, *Planet of the Apes* franchise [¹ Maybe... I think.]

Abstract

Our Sun is a magnetically variable star, and this has profound impacts on planets like Earth which possess both an atmosphere and an intrinsic magnetic field. Heliophysics is the study of how the Sun influences the state of the natural space environment, called 'Space Weather'. Space weather is primarily driven by solar flares and coronal mass ejections, which are the most energetic manifestations of solar magnetic activity. These phenomena typically originate from solar active regions, which emerge as sunspots on the Sun's surface. Monitoring sunspots is a key component of heliophysics. This field of study relies heavily on imagery from both ground-based telescopes, such as those at the Royal Observatory of Belgium (ROB), and space-based missions like the Solar Dynamics Observatory (SDO). However, ground-based observations are subject to atmospheric disturbances, which can complicate the analysis of solar activity.

This thesis first presents a fully automated framework for the identification, grouping, and classification of sunspots observed in white light imagery. The aim of this framework is to support the monitoring of solar activity while requiring minimal manual annotation. The proposed system detects sunspots at the pixel level, organizes them into sunspot groups, and classifies these groups according to the McIntosh system. Our study demonstrates that utilizing multiple low-quality, automatically generated pseudolabels per sample during the training of a convolutional neural network (CNN) enables effective and robust sunspot segmentation. The proposed method proves to be highly effective in handling noise and atmospheric disturbances that commonly affect ground-based imagery, such as the images recorded by the Uccle Solar Equatorial Table (USET) at ROB. Regarding the clustering of sunspots, we propose a simple aggregation

procedure to accurately retrieve the same sunspot groups as the ones identified in the USET catalog maintained by ROB. Finally, we propose a CNN architecture that can successfully extract relevant features from sunspot groups and classify them based on labels from the USET catalog. We show that an ensemble learning approach allows to identify ambiguous cases and to recover annotation errors. Such quality control ensures consistent and accurate classification of sunspot groups, enhancing the reliability and robustness of solar activity monitoring.

In the second part of this thesis, we address the task of filling in gaps in the frequency coverage of solar observations. Such gaps originate either from missing historical observation modes or from limited instrument deployment beyond the Earth-Sun line. Hence, we explore the possibility to predict the missing modality from the available one(s). Our investigation explores deep learning models for cross-modal image translation, which synthesize target-modality observations from available source-modality images. Notably, two specific image translation scenarios are considered: (a) CaIIK views of sunspots and their surroundings from white light images, and (b) magnetograms from EUV images. Our study envisions and compares two models, addressing image translation differently. The first relies on generative adversarial networks (GAN) that are widely adopted in various research fields. The other is based on a modulation mechanism to affect the reconstruction based on features explicitly extracted from the source modality. Our work demonstrates that adaptive instance modulation results in reconstructions that are less prone to hallucinations. Increased robustness to hallucinations is an important advantage in solar physics research where spurious features can be highly damaging.

Remerciements

Le travail d'un chercheur en thèse est souvent solitaire, mais il serait malhonnête de dire qu'il ne résulte que de lui-même. C'est pourquoi je tiens à exprimer ma gratitude envers toutes les personnes qui ont contribué, de près ou de loin, à l'aboutissement de ce travail. Bien que n'étant pas succincte, cette liste de personnes ne saurait non plus être exhaustive du fait du nombre de rencontres et échanges effectués au cours de ces années, mais je me lance.

Tout d'abord, j'aimerais adresser mes plus sincères remerciements à mes promoteurs, le Professeur Christophe De Vleeschouwer et le Docteur Véronique Delouille de l'Observatoire Royal de Belgique. Je vous suis reconnaissant pour la confiance que vous avez placée en moi, vos conseils éclairés ainsi que votre exigence et encouragements. Votre supervision à toute épreuve m'aura permis de maintenir une motivation autant dans les mauvais moments que les bons et constitue une contribution majeure à l'aboutissement de ce travail doctoral.

Ma gratitude s'étend également aux professeurs qui ont accepté d'être membres de mon jury de thèse. En premier lieu, j'adresse mes sincères remerciements aux professeurs externes à l'UCLouvain Mark Butala et Vincent Barra qui, malgré le peu ou l'absence d'interaction préalable, ont consacré du temps et donné de précieux conseils pour améliorer ce manuscrit. J'exprime également ma reconnaissance aux professeurs Laurent Jacques et Laurent Francis de l'UCLouvain, respectivement pour m'avoir suivi tout au long de ce parcours doctoral et pour avoir assumé la présidence de mon jury.

Une thèse, ça se mène sur la durée, mais repose bien sûr sur des interactions quotidiennes. C'est pourquoi je tiens particulièrement à saluer tous ceux que j'ai rencontrés à l'UCLouvain et en particulier les membres Xtra du labo de CDV. Tout d'abord, les Dauphins du Stévin: Simon, Benoît, Amir, Pierre-Yves, Stéphanie, Gabriel, Maxime, Antoine, Victor et Anne-Sophie. Je suis particulièrement reconnaissant envers Victor, pour m'avoir permis d'entrevoir un retour possible dans la recherche à l'université après mon petit passage en industrie, au détour d'une rencontre fortuite en gare, et Anne-Sophie, pour m'avoir encouragé à oser me lancer dans un doctorat quand l'opportunité s'est présentée. Les dauphins ont ensuite laissé place aux membres du Concatenate Group qui, malgré un nom vaguement sérieux de facade, a pour mascottes un Yorkshire nain nommé Zoé et une voiture rouillée nommée Titine (si si, je vous jure). Dès lors, j'adresse un chaleureux merci à Abolfazl, Nana, Benoît, Baptiste, Florine, Vladimir (le papa de Zoé), Alexandre, Antoine (le papa de Titine), Gilles, Mathieu, Ali, Sarra, Elias et enfin Dany, notre membre honoraire. Chacune de ces personnalités hautes en couleurs a été incontournable dans l'achèvement de cette thèse, de par nos discussions foireuses (souvent !) et sérieuses (rares !) et surtout par la cohésion et le soutien indéfectible entre membres du groupe à chaque instant. Je ne pourrai jamais oublier les CCBs, les quiz cinema/Wikipedia, les Digestoguessr et autres Vendrediplodocus.

L'Observatoire n'est bien sûr pas en reste avec Sabrina et Laure, membres de l'équipe d'observation solaire à qui j'exprime ma gratitude pour leur aide dans la manipulation des données exploitées et leur expertise en héliophysique essentielle dans ce travail. Un grand merci également au CISM pour les ressources de calcul mises à disposition des chercheurs, ainsi qu'à l'agent masqué qui veille à leur usage correct.

Enfin, je souhaite conclure en rendant hommage à tous mes amis et membres de ma famille qui, malgré le fait qu'ils se demandaient certainement "Mais qu'est-ce qu'il fait? De quoi il parle?", n'ont jamais cessé de m'entourer et me soutenir dans ce long parcours, de jour comme de nuit. À tout ceux qui pensent "Il va encore penser qu'il usurpe un diplôme. ..." je leur dirai "Aaaaah! ... P'tet pas cette fois." On en discutera. Merci à tous. Bisous.

Contents

Contents	v
List of Publications	xi
List of Figures	xi
List of symbols	xv
1 Introduction	1
1.1 Research Questions	3
1.2 Contributions and Thesis Outline	4
2 Heliophysics Background	7
2.1 Structure of the Sun	8
2.1.1 <i>Solar Interior</i>	9
2.1.2 <i>Solar Atmosphere</i>	11
2.2 Active regions	13
2.2.1 <i>Solar Dynamo</i>	14
2.2.2 <i>Magnetic Features</i>	16
2.2.3 <i>Sunspot Group Lifecycle and Classification</i>	19
2.3 Observation Tools and Datasets	22
2.3.1 <i>Ground-based Observatories</i>	25
2.3.2 <i>Space-based Observatories</i>	26
3 Deep Learning Background	29
3.1 Deep Learning Basics	30
3.2 Deep Learning for Visual Content Analysis	35
3.2.1 <i>Architectures for Visual Content</i>	35

3.2.2	<i>Visual Data Augmentation</i>	40
3.3	Deep Learning for Image Generation	41
3.3.1	<i>Use Cases</i>	42
3.3.2	<i>Frameworks/Paradigms</i>	45
3.3.3	<i>Quantitative Evaluation of Generative Models</i>	52
4	Study of sunspot groups in continuum imagery	57
4.1	Motivation and related works	60
4.1.1	<i>Sunspot segmentation</i>	60
4.1.2	<i>Grouping of individual sunspots into sunspot groups</i>	61
4.1.3	<i>Sunspot Classification</i>	62
4.2	Data	63
4.2.1	<i>USET input dataset</i>	63
4.2.2	<i>Segmentation dataset</i>	64
4.2.3	<i>Clustering dataset</i>	65
4.2.4	<i>Classification dataset</i>	65
4.3	Methods	66
4.3.1	<i>Pipeline overview</i>	66
4.3.2	<i>Segmentation</i>	67
4.3.3	<i>Clustering</i>	71
4.3.4	<i>Classification</i>	72
4.4	Results	75
4.4.1	<i>Sunspot Segmentation</i>	76
4.4.2	<i>Clustering</i>	78
4.4.3	<i>Classification</i>	81
4.5	Discussion and Perspectives	88
5	Cross-modal Translation of Active Regions Observations	93
5.1	Motivations and related work	95
5.1.1	<i>Cross-modal I2I translation in solar physics</i>	96
5.1.2	<i>Our contributions</i>	97
5.2	Data	98
5.2.1	<i>Translation from WL to Ca II K images</i>	98
5.2.2	<i>Translation from EUV images to magnetograms</i>	98
5.3	Methods	100
5.3.1	<i>Conditional adversarial network</i>	101
5.3.2	<i>Non-adversarial strategies</i>	103
5.4	Experiments	106
5.4.1	<i>Train-test data splitting</i>	106

- 5.4.2 *Performance metrics* 107
 - 5.5 Results 108
 - 5.5.1 *Generic reconstruction quality metrics* 108
 - 5.5.2 *WL-to-CallK* 110
 - 5.5.3 *AIA-304Å-to-HMI* 114
 - 5.6 Discussion and Perspectives 117
- 6 **Conclusion** 121
 - 6.1 Research Questions recap and Contributions 122
 - 6.2 Perspectives 124
 - 6.3 Final Thoughts 126
- ★ **Appendices** 127
 - A USET observation equipment 127
 - B Clustering method details 128
 - C Optimal Clustering parameters grid search 129
 - D Sunspot Group Database entry bounding box computation . . 130
 - E Classification precision, recall, support and F1 score 131
 - F Classification results details 132
 - G Attempts at I2I translation using diffusion-based models . . . 133
 - H I2IwFiLM versus DiffI2I 135
 - I Quality metrics 137
 - J Weber contrast 138
- Bibliography** 141

List of Publications

Topic-related Publications

The work described in this thesis has been published in peer-reviewed journals and presented at international conferences. These publications directly reflect the research achievements and key findings discussed in the following chapters. The content of Chapter 4 corresponds to the first publication, and Chapter 5 presents the research from the second publication.

- **Niels Sayez**, Christophe De Vleeschouwer, Véronique Delouille, Sabrina Bechet, and Laure Lefèvre. SunSCC: Segmenting, Grouping and Classifying Sunspots from Ground-Based Observations Using Deep Learning. In *Journal of Geophysical Research: Space Physics*, 128(12): e2023JA031548, 2023.
- **Niels Sayez**, Christophe De Vleeschouwer, Véronique Delouille, Sabrina Bechet, and Laure Lefèvre. Mitigating hallucination with non-adversarial strategies for image-to-image translation in solar physics. Under review in *Astronomy & Astrophysics: Numerical methods and codes*, 2025.

Unrelated Publications

Beyond the solar physics-related works, several other peer-reviewed publications were produced during the doctoral training. While they explore different research directions, they represent additional contributions to the field of deep learning for computer vision.

- **Niels Sayez** and Christophe De Vleeschouwer. Accelerating the creation of instance segmentation training sets through bounding box annotation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 252–258, 2022.
- Seyed Abolfazl Ghasemzadeh, Gabriel Van Zandycke, Maxime Istasse, **Niels Sayez**, Amirafshar Moshtaghpour, and Christophe De Vleeschouwer. Deepsportlab: a unified framework for ball detection, player instance segmentation and pose estimation in team sports scenes. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 379. BMVA Press, 2021.

List of Figures

2.1	Schematic of the Solar system formation process.	8
2.2	Cross-sectional schematic of the Sun from the inner core to the corona.	9
2.3	Illustration of the Sun's rotation profile highlighting differential rotation.	11
2.4	Three-panel depiction of solar activity showing cyclic characteristics.	15
2.5	Illustration of the magnetic cycle alternating between poloidal and toroidal configurations.	16
2.6	Active region displaying photospheric and chromospheric structures.	17
2.7	Cross-sectional diagram of sunspot formation mechanism. . . .	18
2.8	Illustration of the McIntosh classification scheme for sunspot groups.	22
2.9	Illustration of the Mount Wilson classification scheme for sunspot groups.	23
2.10	Artifacts affecting ground-based solar observations.	26
2.11	Fleet of space missions observing the Sun.	27
3.1	Basic building blocks of deep learning: neurons, layers and activation functions.	31
3.2	Illustration of kernel convolution operation and popular kernels. . . .	36
3.3	Illustration of a CNN encoding a single-channel image into low-dimensional feature maps.	37
3.4	Illustration of the U-Net architecture.	38
3.5	Common data augmentation techniques for visual data.	41
3.6	Illustration of common visual tasks addressed with deep learning. . . .	42

3.7	Encoder-Decoder architecture overview.	46
3.8	Generative Adversarial Networks paradigm overview.	47
3.9	Diffusion Model paradigm overview.	49
4.1	SunSCC pipeline for sunspot segmentation, clustering, and clas- sification.	68
4.2	Pseudolabel generation pipeline.	69
4.3	Reduced McIntosh classification scheme used in SunSCC.	73
4.4	F1-score of sunspot detection as a function of threshold value for each sunspot size bin.	79
4.5	Trade-off between sunspots remaining undetected and sunspots being merged in a single entity.	80
4.6	Test set confusion matrices of classification with majority vote for the classification of the Z, p and c components.	84
4.7	Classification accuracy along angular distance and Train/Test class distributions	86
4.8	Examples of failure cases close to the limb.	87
4.9	Example of annotation error in USET catalog revealed by SunSCC.	88
4.10	Per-component test set confusion matrices for correct and in- correct majority votes.	89
4.11	Examples of ambiguous cases.	90
5.1	Sample pair of WL and CalI-K segments used for whitelight-to- calcium translation	99
5.2	Sample pair of AIA-304Å and HMI images used for EUV-to- Magnetogram translation	100
5.3	Diagram of Pix2PixCC model and adversarial training for WL- to-CalIK translation.	102
5.4	First training stage of I2lWFiLM.	104
5.5	Second training stage of I2lWFiLM.	105
5.6	Weber contrast between plage pixels intensity and background pixels intensity along angular distance bins in WL and CalIK images.	111
5.7	Highest-quality CalIK reconstructions of ARs near the center of the solar disc and near the limb.	112
5.8	Subpar CalIK reconstructions of ARs near the center of the solar disc and near the limb.	113
5.9	Comparison of Pix2PixCC and I2lWFiLM performances using generic image reconstruction metrics	113

5.10 Plage reconstruction performance of Pix2PixCC and I2lwFiLM models using application-specific metrics. 115

5.11 Example of reconstructed HMI Bz using Pix2PixCC model and I2lwFiLM. 116

5.12 Δ SSIM histograms for I2lwFiLM and Pix2PixCC models computed on AIA-304Å-to-HMI test set. 117

A.1 USET facility at the Royal Observatory of Belgium. 128

C.1 Search for optimal kernel longitudinal and latitudinal bandwidths values for the proposed clustering method. 130

G.1 Attempts using DiT architectures. 133

G.2 Attempts using LBBDM architecture. 134

G.3 Attempts using DiT architectures. 135

H.1 Second training stage of DiffI2I model learning to recover $\mathcal{E}^{\mathcal{P}}$ vectors from embedding space $\mathcal{E}^{\mathcal{W}}$ 137

List of symbols

Deep Learning-related Notations

Acronyms

cGAN	Conditional Generative Adversarial Network - A GAN variant controlling generation with auxiliary information. (p. 48)
CNN	Convolutional Neural Network - A deep learning model that uses convolution operations to extract features for pattern recognition in grid-like data like images. (p. 2)
DM	Diffusion Model - Framework that transforms random noise into structured data by progressively removing noise through a learned reverse diffusion process. (p. 2)
ED	Encoder-Decoder - Framework compressing input data into a compact latent representation and reconstructs the desired output from this encoding. (p. 45)
FiLM	Feature-wise Linear Modulation. (p. 39)
GAN	Generative Adversarial Network - Dual-network framework where a generator and a discriminator compete adversarially, with iterative training refining the generator to produce realistic synthetic data. (p. 2)
GT	Ground Truth - The reference label of a sample (p. 32)
GVP	Guidance Vector Prediction. (p. 105)
I2I	Image-to-Image (p. 44)

LR	Learning Rate - A hyperparameter that controls the magnitude of weight updates during the training of a deep learning model. (p. 34)
MLP	Multi-Layer Perceptron - Deep learning model composed of several layers of logical neurons, where each neuron in one layer is connected to every neuron in the next. (p. 32)
NN	Neural Network (p. 30)
ReLU	Rectified Linear Unit - An activation function (p. 30)
VAE	Variational Autoencoder - A generative model learning to encode input data into a latent space and reconstruct it, using variational inference to approximate the data distribution. (p. 2)
ViT	Vision Transformer - A deep learning model that applies transformer architecture to image processing by treating image patches as sequences. (p. 35)

Losses and Metrics

ΔSSIM	Difference between $\text{SSIM}_{\text{model}}$ and baseline $\text{SSIM}_{\text{flat}}$. (p. 108)
λ_i	Hyperparameter weighting the importance of the i -th loss function. (p. 33)
$\mathcal{L}$	Loss function (p. 33)
$\mathcal{L}_{CC}$	Correlation coefficient based loss function. (p. 102)
$\mathcal{L}_{FM}$	Feature-matching loss function. (p. 102)
$\mathcal{L}_{LSGAN, D}$	Least-squares adversarial loss function for the discriminator of a GAN. (p. 103)
$\mathcal{L}_{LSGAN, G}$	Least-squares adversarial loss function for the generator of a GAN. (p. 101)
IoU	Intersection over Union - Metric evaluating spatial predictions in object detection and segmentation tasks (p. 34)
MS-SSIM	Multi-Scale Structural Similarity Index Measure (p. 107)
MSE	Mean Square Error (p. 33)
PPR	Plage Pixel Ratio (p. 108)
PSNR	Peak Signal-to-Noise Ratio (p. 43)

$\widehat{\text{RMSE}}$	Root Mean Squared Error (p. 107)
SSIM	Structural Similarity Index Measure (p. 43)
SSIM_{flat}	SSIM between a flat image and the target CaIIK image. (p. 108)
SSIM_{model}	SSIM between the generated and target CaIIK images. (p. 108)

Images, Sets and Vectors

$\widehat{\mathcal{E}}^{\mathcal{P}}$	Guidance vector predicted from the input image $\mathcal{I}_{\mathcal{W}}$. (p. 105)
$\widehat{\mathcal{I}}_{\mathcal{C}}$	Reconstructed Ca II K image. (p. 101)
$\widehat{\mathcal{P}}$	Pair of source-reconstructed target samples of I2I translation task. (p. 101)
$\mathcal{E}^{\mathcal{P}}$	Guidance vector extracted from the pair $\mathcal{P}$. (p. 105)
$\mathcal{I}_{\mathcal{C}}$	Original Ca II K image. (p. 101)
$\mathcal{I}_{\mathcal{W}}$	Original white light image. (p. 101)
$\mathcal{M}_I$	Set of pseudo-labels associated to image I (p. 70)
$\mathcal{P}$	Pair of corresponding source-target samples of I2I translation task. (p. 101)
$\mathcal{WL}_{\mathcal{C}}$	WL images used for clustering and classification tasks (p. 65)
$\mathcal{WL}_{SA}$	WL images used for segmentation training, using automatically generated annotation (pseudolabels) (p. 65)
$\mathcal{WL}_{SM}$	WL images used for segmentation validation, manually annotated (p. 65)
m_{τ_j}	Mask obtained with threshold τ_j (p. 70)
BBox	Bounding Box (p. 64)

Misc.

D	Discriminator network of a GAN. (p. 101)
G	Generator network of a GAN. (p. 101)
$\mathbf{M}(i, j)$	Confusion between class i and class j in the confusion matrix $\mathbf{M}$. (p. 83)

$I_{\bullet}$	Isolated boxes of type $\bullet$, $\bullet$ can be T for Target boxes or M for Mean-shift boxes. (p. 78)
$U_{\bullet}$	Unmatched boxes of type $\bullet$, $\bullet$ can be T for Target boxes or M for Meanshift boxes. (p. 78)
$\#I_{\bullet}$	Size of the set of isolated boxes of type $\bullet$. (p. 78)
$\#U_{\bullet}$	Size of the set of unmatched boxes of type $\bullet$. (p. 78)

Heliophysics-related Notations

Acronyms

AIA	Atmospheric Imaging Assembly - Four-telescopes array onboard the SDO mission observing the Sun in extreme ultraviolet. (p. 27)
AR	Active Region - Region of intense intricate magnetic field originating from the solar interior and reaching the surface and atmosphere. (p. 1)
BL	Babcock-Leighton - Mechanism transitionning toroidal magnetic field lines to poloidal configuration. (p. 14)
BMR	Bipolar Magnetic Region - Internal magnetic field lines crossing the solar surface showing opposite polarity at their emergence and entry points. (p. 14)
CaII K	Calcium II K line - Spectral line at 393.4 nm. (p. 23)
CCD	Charge-Coupled Device - Digital camera technology. (p. 25)
CME	Coronal Mass Ejection - A massive cloud of plasma and magnetic field ejected from the solar corona into space. (p. 1)
DDB	Drawing Database - Historical records of photospheric observations collected at ROB. (p. 64)
EUV	Extreme Ultra Violet - Wavelength range from 10 to 120 nm. (p. 27)
GONG	Global Oscillation Network Group - Worldwide network of identical telescopes providing helioseismic data and magnetograms. (p. 25)
H α	Hydrogen-alpha line - Spectral line at 656.3 nm. (p. 23)
HMI	Helioseismic and Magnetic Imager - Instrument onboard SDO providing filtergrams/magnetograms of the photosphere. (p. 63)

LOS	Line-of-sight. (p. 96)
MDI	Michelson Doppler Imager - Instrument onboard the SOHO mission measuring velocity and producing magnetograms of the photosphere (now superseded by the SDO/HMI instrument). (p. 60)
ROB	Royal Observatory of Belgium. (p. 2)
SDO	Solar Dynamics Observatory - Mission dedicated to continuous study of the Sun, measuring its EUV irradiance and producing multi-wavelengths high-resolution observations and magnetograms. (p. 2)
SGDB	Sunspot Group Database - A comprehensive catalog containing all sunspot groups historically observed at ROB, along with their associated properties and characteristics. (p. 64)
USET	Uccle Solar Equatorial Table - Solar observation facility at ROB equipped with four full-disk imaging telescopes for sunspot drawings, photospheric and chromospheric ($H\alpha$, CaII K) observations. (p. 25)
WL	White Light - Observations of the Sun in the visible spectrum. (p. 23)

McIntosh Classification

Zpc	3-components code for McIntosh scheme for sunspot group classification. (p. 20)
asym	Class label for asymmetric penumbrae, merging classes a and k (p. 73)
frag	Class label for fragmented sunspot groups, merging classes i and c (p. 73)
LG	Class label for Large Groups, merging classes D , E , and F (p. 72)
sym	Class label for symmetric penumbrae, merging classes s and h (p. 73)

Misc.

$\sin(\theta)$	Angular distance between the center of the Sun and the observed point. (p. 110)
$R_{\odot}$	Physical radius of the Sun (p. 8)
R	Radius of the solar disc in an observation. (p. 84)
r	Distance between the center of an AR and the solar disc center. (p. 84)

1

Introduction

HELIOPHYSICS is the discipline that studies how the Sun drives space weather, defined as the physical and phenomenological state of natural space environment. Variations of the internal magnetic activity of the Sun leads to observable 'active regions' (AR) as intense and intricate magnetic fields emerge from the solar interior and interact with the solar atmosphere. ARs appear at the solar surface mainly in the form of 'sunspots' and 'faculae' that develop in the solar atmosphere as 'plages' in lower altitude and as 'coronal loops' in the higher part of the atmosphere named 'solar corona'. Sunspots are the main source location of eruptive events such as 'solar flares' and 'coronal mass ejections' (CME), the two main drivers of space weather. Solar flares release energy in the form of photons emitted across the whole wavelength spectrum, while CMEs are massive clouds of plasma ejected from the solar corona into space. Strong flares affect the ionised upper layer of the Earth's atmosphere and disturb telecommunication and global positioning systems, while CMEs in certain magnetic field configuration may cause rapid changes in the geomagnetic field and also induce electrical currents at the Earth's surface, potentially affecting electric transmission grids.

Therefore, scientific curiosity and the need to protect our technological infrastructure have driven systematic monitoring of solar activity and the development of tools able to report and predict space weather events. He-

1 | Introduction

liophysics research often relies on imagery from both ground-based telescopes, such as those at the Royal Observatory of Belgium (ROB) , and space-based missions like the Solar Dynamics Observatory (SDO). However, ground-based observations are prone to noise from atmospheric disturbances, complicating the analysis of solar activity. Namely, the images recorded at ROB or any other ground observatory are subject to atmospheric turbulence (also known as seeing) and other disturbances such as clouds, aircrafts or dust on the optics. On the other hand, space-based observations are not immune to their own set of challenges, such as the limited number of instruments deployed and the impossibility to maintain or replace sensors embarked on the spacecrafts.

The observational capabilities have evolved from early telescopes to digital cameras and multi-wavelength instruments. While this technological progression has enabled increasingly detailed solar observations, it also created inherent gaps in our historical records. Many solar phenomena, particularly those visible only in certain spectral ranges or occurring in specific layers of the solar atmosphere, are completely absent from historical datasets before the introduction of appropriate instruments.

Deep learning has revolutionized image analysis over the past decade, achieving state-of-the-art performance across a wide range of computer vision tasks such as object detection, segmentation, and image generation. This success story began with Convolutional Neural Networks (CNNs), which introduced the ability to automatically learn hierarchical features from images. The impact of CNNs quickly spread beyond computer vision to transform various domains including medical imaging, autonomous driving, and satellite imagery analysis. The field has continued to evolve with the emergence of Transformers, which have enhanced deep learning capabilities by better handling long-range dependencies and context understanding. Parallel to these developments, remarkable progress has been made in generative modeling. The introduction of Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and more recently, Diffusion Models (DM) has pushed the boundaries of what's possible in image restoration, image synthesis and image-to-image translation. Advanced architectures like Stable Diffusion exemplify this progress, demonstrating unprecedented abilities in creating highly realistic images from text descriptions and performing sophisticated image editing tasks.

2 |

The application of these powerful deep learning techniques to heliophysics, however, presents unique challenges. While conventional deep learning tasks benefit from massive datasets, with platforms like ImageNet providing over 14 million annotated images for training, heliophysics research operates in a regime of significant data scarcity. Available samples often number only in the thousands, a limitation that stems from both the specialized nature of solar observations and the substantial expertise required for accurate annotation. Moreover, unlike many general computer vision applications where approximate results might be acceptable, heliophysics demands exceptionally high precision and physical accuracy in its analyses. These constraints make the development of deep learning solutions for solar physics particularly challenging, requiring careful consideration of both architectural choices and training strategies.

1.1 Research Questions

This thesis aims to leverage deep learning to offer tools that can assist heliophysics studies and improve solar activity monitoring, while addressing several challenges: using ground-based imagery and requiring minimal manual label annotation, providing results that are reliable and physically accurate. This thesis addresses research questions related to two applications using solar observations: the detection and classification of sunspot groups on the surface of the Sun, and physically accurate cross-modality image translation.

For sunspot group detection and classification, traditional computer vision approaches have often struggled with the inherent variability in ground-based observations and atmospheric interference. While deep learning offers promising solutions, the development of reliable automated tools faces significant challenges: ground-based solar observations contain artifacts and varying image quality conditions that complicate feature detection, and the field lacks extensive labeled datasets due to the time-intensive nature of manual annotation. In this context, we investigate whether we can develop an efficient and fully automated tool using ground-based solar observations only, and if such a tool can provide reliable results that overcome artifacts inherent to ground-based imagery and the limited amount of samples with manually annotated sunspots.

Q1 Can a deep learning model reliably detect and characterize sunspot groups from ground-based images despite atmospheric artifacts and weak annotation?

Regarding cross-modality image translation, popular deep learning solutions such as Generative Adversarial Networks have been widely adopted for their ability to generate highly convincing images. However, these approaches often suffer from hallucinations and physical accuracy issues, making them potentially unreliable for scientific applications. This raises the crucial question of whether the predictions of such a generative model can be trusted and meaningfully exploited in heliophysics research, where physical accuracy is paramount. In this context, we explore whether we can develop a model that can predict observations in a target modality from another one available, while ensuring the predictions maintain physical consistency with the input observations.

Q2 Is it possible to fill gaps in historical records by exploiting observations of another correlated modality?

Q3 Do generative models produce physically reliable predictions that can be used in heliophysics research?

Successfully addressing these research questions would enable the reconstruction of observations in specific modalities from other available ones, effectively filling historical gaps when only certain measurements existed.

1.2 Contributions and Thesis Outline

This manuscript targets a dual audience of heliophysicists and machine learning engineers. Special care was taken to bridge knowledge gaps between these communities through dedicated background chapters. This enables deep learning practitioners to understand heliophysics challenges, while helping heliophysicists grasp modern deep learning tools and their limitations. This manuscript comprises six chapters. Following this intro-

duction, two chapters establish the needed background: Chapter 2 examines heliophysics essentials, while Chapter 3 explores deep learning fundamentals.

Chapter 2 introduces the Sun's structure and its magnetic activity, establishing the origin of active regions. It then examines active regions and their manifestations in detail, particularly through sunspots and their classification. The chapter closes with a presentation of current solar monitoring tools and the limitations of ground-based observations.

Chapter 3 presents the deep learning foundations essential to this research. After covering core concepts and terminology, it details specific challenges of visual data processing and their deep learning solutions. Finally, it explores image generation tasks and the architectures designed for these applications.

These introductory chapters pave the way for Chapter 4 and 5, two technical chapters that form the core of this thesis research contributions.

Chapter 4 introduces SunSCC, a fully automated framework for detecting and classifying sunspot groups in support of solar activity monitoring. The framework operates in three stages: first, a CNN segments sunspots in white light images captured at ROB, leveraging automatically generated pseudo-labels for training and handling noise and atmospheric disturbances common in ground-based observations. Second, an innovative clustering method aggregates these sunspots into coherent groups aligned with ROB's historical catalog. Finally, an ensemble of CNN classifiers perform group classification according to a well-established group classification scheme. Consensus among the ensemble systematically identifies and prioritizes groups needing expert review, facilitating the integration of rigorous consistency checks to advance sunspot group monitoring precision.

Chapter 5 addresses a critical challenge in generative deep learning models: their tendency to produce visually convincing outputs that may be physically inconsistent with input data. This limitation is especially problematic in scientific applications requiring accurate cross-modal relationships. The chapter focuses on solar cross-modality image-to-image translation, predicting missing solar observations (e.g., calcium K spectral line)

from alternative observations available (e.g., white light images). Diverging from adversarial networks and noise-based models that introduce unreliable artifacts, this chapter presents a deep learning model prioritizing robust feature extraction from input modality through enhanced encoder-decoder architectures. The model surpasses popular approaches in predicting target observations from input, maintaining the physical consistency crucial for heliophysics research. Experiments validate that the generated images preserve essential physical characteristics, rendering them reliable tools for solar physics studies when original observations are unavailable.

The manuscript concludes with a final chapter that synthesizes the key contributions of this research. This conclusion summarizes the findings and achievements, and provides a thoughtful discussion of the current limitations in our approaches to each task and outlines promising directions for future research in this interdisciplinary field.

2

Heliophysics Background

Our Sun, born approximately 4.6 billion years ago from the gravitational collapse of a molecular cloud, represents a quintessential example of stellar formation in the Universe. The initial cloud of gas and dust underwent gravitational collapse, leading to the formation of a rotating protostellar disk. As temperatures rose and matter concentrated in the core, hydrogen fusion began through the proton-proton chain reaction, marking the birth of our G2V yellow dwarf star. The remaining material in the protostellar disk formed the planets and other bodies in our solar system, with the entire formation process taking approximately 50 million years. A schematic of the solar system formation process is shown in Figure 2.1.

In this chapter, our current understanding of the Sun [1] is presented. Section 2.1 describes the structure of the Sun and Section 2.2 depicts its internal magnetic activity and its associated features. Finally, Section 2.3 describes the observational tools and datasets that have shaped these theories and drive the current heliophysics research, with a focus on the ones used in this thesis.

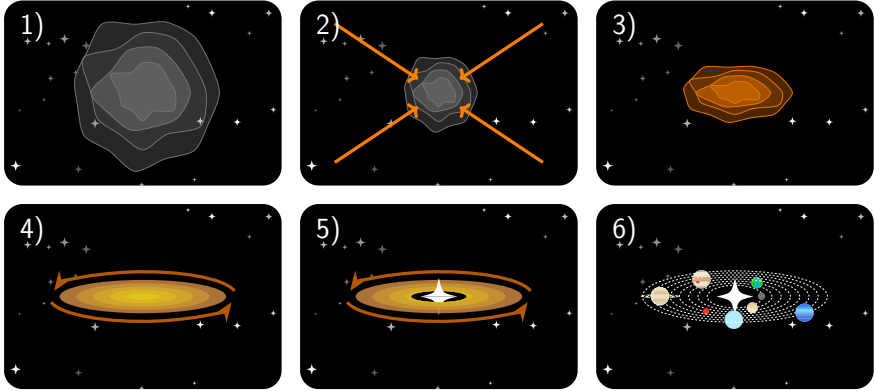


Fig. 2.1 Schematic of the Solar system formation process. 1) Initial cloud of gas. 2) Gravitational collapse of the cloud. 3) Heating of the collapsed cloud. 4) Formation of the protostellar disk. 5) Formation of the Sun core. 6) Formation of the Sun and planets of the system. Reproduced from: [2].

2.1 Structure of the Sun

Throughout its interior and external atmosphere, the Sun consists entirely of plasma at different temperatures and densities, with the photosphere marking the visible boundary between the internal and external layers. The photosphere lies at a radius of 696,340 kilometers (denoted as $R_{\odot}$) from the Sun's center. While the interior of the Sun is not directly observable, its structure has been revealed through helioseismology, a technique that uses the Sun's natural oscillations to probe its internal properties [3]. The internal layers play crucial roles in the solar dynamics, from energy generation to the observable surface phenomena.

The commonly accepted model of the Sun's interior consists of three main layers: the core (0 to $0.25R_{\odot}$) where nuclear fusion occurs, the radiative zone ($0.25R_{\odot}$ to $0.7R_{\odot}$), and the convective zone ($0.7R_{\odot}$ to $1R_{\odot}$). Above these interior regions, the solar atmosphere comprises four distinct layers: the photosphere (thickness ~ 500 km), chromosphere ($\sim 2,000$ km), transition region (~ 100 km), and corona (extending beyond 10^6 km). The complete structure is illustrated in Figure 2.2.

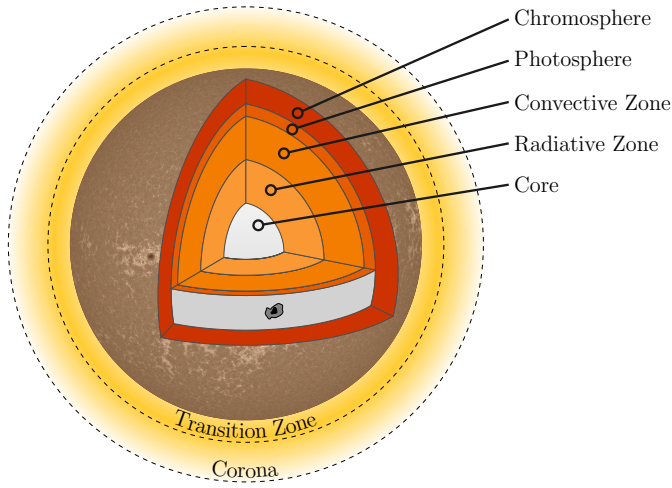


Fig. 2.2 Cross-sectional schematic of the Sun from the inner core to the corona. The photosphere, observed in white light (continuum) imaging is the solar surface visible in optical spectrum. The chromosphere is commonly revealed through calcium II K-line (CaII K) spectral observations.

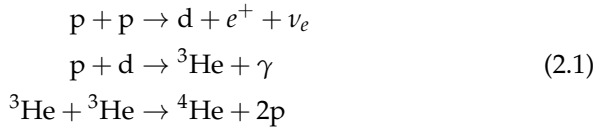
2.1.1 Solar Interior

Core

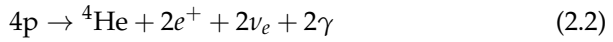
The core is the innermost layer of plasma in the Sun, extending to a radius of $0.25R_{\odot}$. The core's density reaches $1.5 \times 10^5 \text{ kg/m}^3$ and contains approximately 50% of the Sun's total mass. This plasma environment sustains extreme conditions, with temperatures reaching $1.57 \times 10^7 \text{ K}$ and pressures of $2.5 \times 10^{11} \text{ atm}$.

These conditions enable thermonuclear fusion through the proton-proton (pp) chain [4, 5], which powers the Sun by converting hydrogen into helium. The pp chain first produces deuterium d -alternatively noted ^2H - through the fusion of two protons p -or ^1H - releasing a positron (e^+) and a neutrino (ν_e) in the process. The deuterium then fuses with another proton to form helium-3 (^3He) and a gamma-ray photon (γ). Finally, two helium-3 nuclei combine to form helium-4 (^4He) and two protons.

2 | Heliophysics Background



The net reaction can be summarized as:



The emitted photons transfer energy from the core to the outer layers of the Sun, where they are eventually radiated into space.

Radiative zone

The radiative zone extends from the core boundary at $0.25R_\odot$ to approximately $0.7R_\odot$. This region is characterized by temperatures decreasing from 7×10^6 K to 2×10^6 K and densities ranging from 2×10^4 to 2×10^2 kg/m³. In this region, energy transport occurs primarily through radiative diffusion, where photons initially produced in the core undergo repeated absorption and re-emission by plasma ions. Typically, photons traverse only ~ 1 cm before interaction, resulting in energy transport timescales of approximately 1.7×10^5 years from core to radiative zone surface.

With increasing distance from the core, opacity of the plasma increases, leading to an accumulation of photons at the radiative zone's outer boundary. This energy accumulation implies steepening of the temperature gradient beyond the threshold for radiative stability (i.e., when radiation alone cannot efficiently transport energy outward). This triggers convection motion of plasma in the layer above, which replaces radiation as the main method of energy transport. The transition region, known as the tachocline ($\sim 0.04R_\odot$ thick), exhibits strong radial and latitudinal shear flows that are thought to drive the solar dynamo responsible for generating the Sun's magnetic field.

Convective zone

The convective zone extends from the tachocline at $0.7R_\odot$ to the photosphere at $R_\odot$. Right above the tachocline, temperatures of 2×10^6 K generate substantial thermal gradients. Due to its low density (under 2×10^2 kg/m³), the zone cannot sustain radiative transfer, and convection takes

over energy transport. The convective instability creates a complex network of plasma flows that transports energy outward.

The convective motions significantly influence global solar dynamics [6, 7], particularly the differential rotation pattern observed at the Sun's surface [8, 9, 10]. As plasma rises and falls in convective cells, Coriolis forces act on these flows, establishing latitude-dependent angular velocity variations. This results in a rotation profile where the equatorial regions rotate with a period of about 25 days, while regions at 30° and 60° latitude rotate more slowly with periods of approximately 27 and 30 days respectively. This differential rotation is illustrated in Figure 2.3.

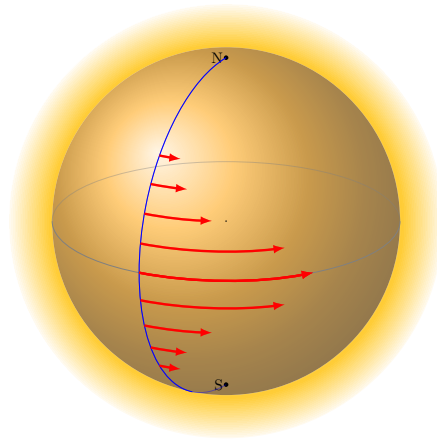


Fig. 2.3 Solar differential rotation visualized through zonal flow trajectories with latitudinal variation in angular velocity from pole to equator. Solar rotation periods range from ~25 days at the equator to ~35 days at 75° latitude. Mid-latitudes (40°) complete a revolution in ~28 days.

2.1.2 Solar Atmosphere

Photosphere

The photosphere is the Sun's visible surface layer (500 km thick). In regions of quiet Sun, which constitute the majority of the solar surface, a characteristic granulation pattern is observable. It emerges from temperature differences in convection cells (~1000 km diameter) of the underneath convective zone, where bright centers (~6600K) represent hot rising

2 | Heliophysics Background

plasma surrounded by darker, cooler downdrafts ($\sim 5800\text{K}$). Perturbations of the solar inner magnetic field cross the surface, creating visible active regions by suppressing convective heat transport. These magnetic manifestations appear in the photosphere as sunspots, which are regions with reduced temperatures ($\sim 4300\text{K}$) appearing darker than the quiet Sun, and as faculae, which are bright filamentary structures. Active region formation mechanism is further detailed in Section 2.2.

Chromosphere

The chromosphere extends approximately 2,000 kilometers above the photosphere. Its temperature rises from 4,400 K to about 20,000 K [1]. The pressure drops from ~ 10 atm at the photosphere to $\sim 10^{-3}$ atm at the upper chromosphere. Magnetic field lines emerging from the photosphere expand with height due to the sharp decrease in gas pressure. This expansion creates a horizontal canopy-like deformation that affects chromospheric features and influences the chromospheric dynamics. Magnetohydrodynamic waves in this layer contribute to mass and energy transport toward the upper solar atmosphere.

Transition Region and Corona

The transition region is a thin interface (~ 200 km) between the low temperature chromosphere (max $\sim 20,000$ K) and high temperature corona ($\sim 500,000$ K). This temperature gradient is a consequence of the heat flowing down from the corona to the chromosphere. The transition region serves as a crucial site for magnetic reconnection events and eruptive phenomena that significantly influence solar activity and space weather.

The corona, extending millions of kilometers into space, is the outermost layer of the solar atmosphere. The temperature in the corona can reach several millions degrees. The corona exhibits complex magnetic field structures, shaped by the Sun's differential rotation and internal magnetic activity. These structures include:

- *Coronal loops* : low altitude arched structures of hot plasma confined to magnetic field lines of active regions.
- *Prominences* : cool plasma suspended in magnetic fields, visible as dark filaments against the bright solar disk.
- *Streamers* : elongated structures of plasma following magnetic field lines, often visible during solar eclipses.

The corona is the source of the solar wind, a continuous stream of charged particles flowing outward through the heliosphere at speeds of ~400-800 km/s. It is also the site of powerful solar flares and coronal mass ejections (CMEs), which can expel billions of tons of magnetized plasma at speeds reaching ~2000 km/s into interplanetary space. CMEs significantly impact space weather, particularly through their interaction with Earth's magnetosphere, producing geomagnetic storms that generate auroras and can affect technological systems.

2.2 Active regions

Astronomical observations of the Sun span centuries [11], with early records originating from East Asia, the Middle East, North Africa, and Europe. These early observers had actually already documented sunspots appearing and vanishing on the solar disc, varying in both size and number. Their historical records provided the first evidence of solar variability. The invention of the telescope in the early 17th century revolutionized solar astronomy in Europe, enabling detailed sunspot observations through systematic drawings [12]. Through systematic sunspot observations in the 1840s, Schwabe discovered that solar activity follows a cyclic pattern of approximately 11 years [13]. This discovery was later confirmed by Rudolf Wolf, who introduced the index of sunspot numbers as a measure of solar activity. This index, now revised and known as the *international sunspot number* is illustrated in Figure 2.4 (top). The link between sunspots and solar magnetic activity was later established by George E. Hale in [14], leading to the identification of these regions as part of larger magnetically active structures, termed *active regions* (AR).

This section describes the emergence of active regions from the inner layers of the Sun up to their manifestations as sunspots and faculae on the photosphere, as plagues on the chromosphere and as extended magnetic field loops in the corona. While the inner layers of the Sun cannot be observed directly, several theories explaining the origin and dynamics of these active regions have been developed [15, 16, 17, 18]. We describe the most widely accepted theories about the complex processes of solar activity and their observable features.

2.2.1 Solar Dynamo

The Sun magnetic field is thought to result from a dynamo process driven by differential rotation and convection in the solar interior. Observations by Carrington in the 1860s [19] revealed what became known as Spörer's law: the progressive migration of sunspot emergence from mid-latitudes ($30^\circ \sim 40^\circ$) toward the equator. When plotted over time, this latitudinal drift creates a characteristic pattern known as the "Butterfly Diagram" [20, 21]. In the early 1900s, Hale's discovery of the magnetic nature of sunspots [14] and their systematic polarity patterns established that the complete magnetic cycle actually spans 22 years, with the global magnetic field undergoing a reversal every 11 years. The characteristic butterfly pattern and associated magnetic polarities are shown in Figure 2.4 (middle, bottom).

The cyclic aspect of this activity is thought to originate from the alternating configuration of magnetic field lines between poloidal and toroidal orientations [23]. The poloidal field, similar to Earth's dipolar field, runs along meridional lines from pole to pole, while the toroidal field runs primarily along latitude lines. This cycle is depicted in Figure 2.5. The poloidal-to-toroidal transition is widely believed to be driven by the differential rotation of the Sun and is named the Ω -effect [15]. The faster rotation near the equator extends the magnetic field lines in the toroidal direction thanks to the property of plasma to freeze in magnetic fields -i.e., magnetic field lines are unable to move freely from charged particles in motion-. As they originate from such deformation, toroidal field lines are oriented in opposite directions in each hemisphere.

The reverse process, the toroidal-to-poloidal transition, is explained by two distinct theories. The first is the α -effect [16] where toroidal lines at the base of the convective zone are deformed and twisted due to turbulence, therefore generating small-scale lines in the poloidal direction. These deformations produce on average large-scale poloidal lines. The second theory, known as the Babcock-Leighton (BL) mechanism [17, 18], is based on the decay of bipolar magnetic regions of opposite polarity appearing on north and south hemispheres. In this mechanism, bipolar magnetic regions (BMR) appear as kinks along the toroidal lines in the tachocline layer in both hemispheres and rise to the surface by buoyancy forces. While they arise in the convective zone, BMRs are tilted in the direction of their

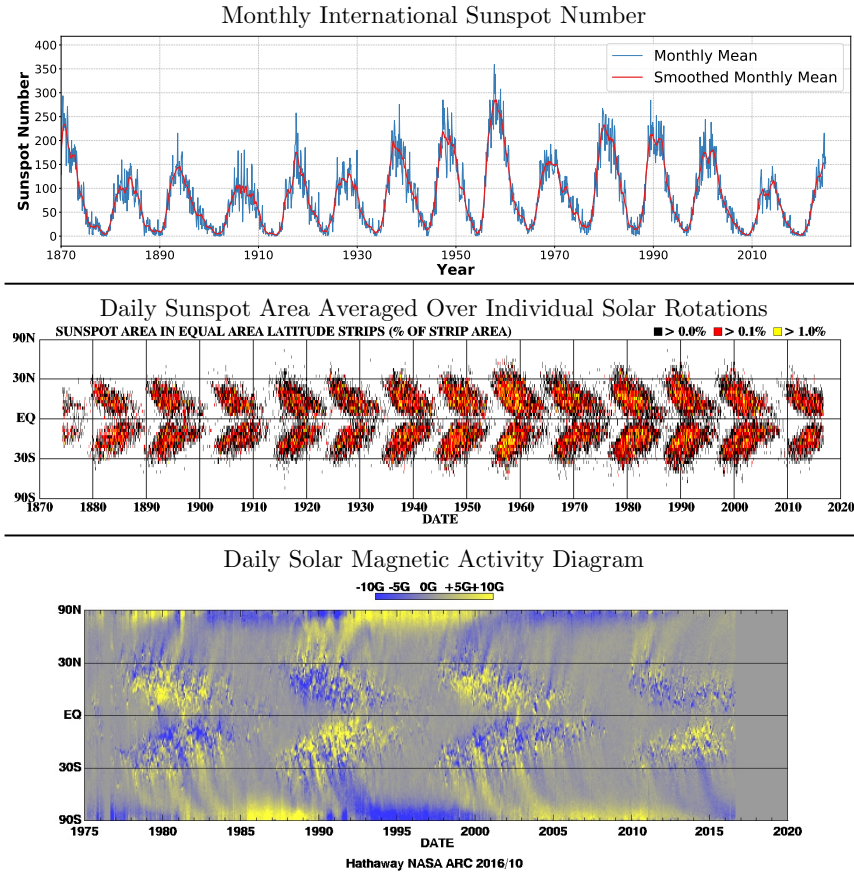


Fig. 2.4 Three-panel depiction of solar activity showing cyclic characteristics. (Top) Evolution of International Sunspot Number (Wolf) over time (credit: SILSO-WDC [22]). (Middle) Butterfly Diagram illustrating the latitudinal drift of sunspot emergence over the solar cycle. The diagram shows the progression of sunspot emergence from high latitudes towards the equator over the course of the solar cycle. (Bottom) Magnetic counterpart of the butterfly diagram showing Hale’s properties.

corresponding hemisphere by the Coriolis force, leading to a difference in latitude between the leading and trailing poles of the BMR.

When these BMRs emerge on the photosphere, they exhibit a characteristic structure relative to solar rotation. Each BMR consists of two poles

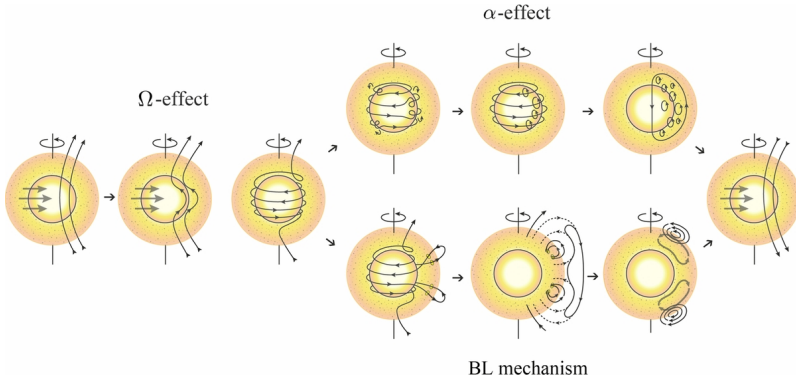


Fig. 2.5 Illustration of the magnetic cycle alternating between poloidal and toroidal configurations caused by the Ω -effect (left), the α -effect (top right) and the Babcock-Leighton (BL) mechanism (bottom right). Illustration adapted from [23].

of opposite polarity: a leading pole positioned closer to the equator that matches the polarity of its hemisphere, and a trailing pole situated farther from it with opposite polarity. The leading pole vanishes by diffusion when approaching the equator, while the trailing pole remains and is carried towards the pole by the meridional flow. The accumulation of trailing poles of BMRs at the rotational poles is thought to be responsible for the global magnetic polarity reversal (see bottom of Figure 2.4), maintaining this cyclic magnetic activity.

2.2.2 Magnetic Features

BMRs are the source of active regions on the photosphere. While rising and reaching the surface, magnetic field lines of BMRs are responsible for the perturbation of convective flow of plasma. In normal circumstances, also named *quiet Sun*, the top of convective cells can be observed on the photosphere as a honeycomb-like pattern called granules. The appearance of ARs and their features on the photosphere and chromosphere are illustrated in Figure 2.6.

The perturbation of convection cells can extend up to the breaking in the convective zone and therefore inhibit flux of hot plasma from the solar interior to the surface. This causes a lack of heat transfer to the sur-

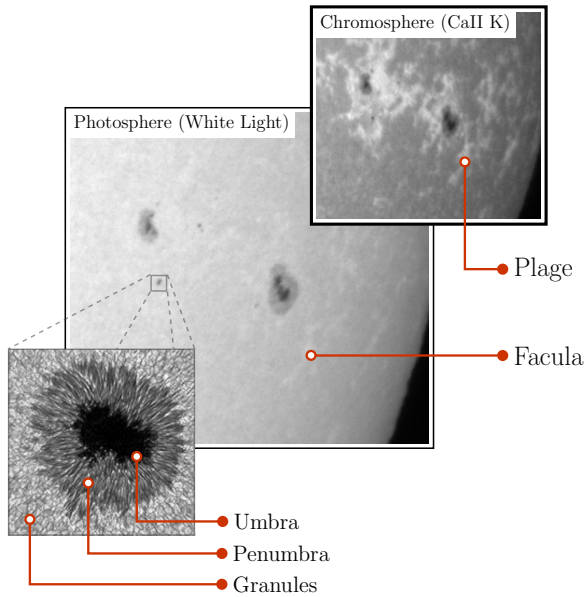


Fig. 2.6 Active region displaying photospheric and chromospheric structures. The photosphere (left) shows quiet-Sun granulation, sunspots with umbrae and penumbrae, and limb-brightened faculae. The chromospheric CaII K-line view (top right) shows plage regions corresponding to faculae in photosphere, expanded due to the magnetic canopy effect.

face, hence reducing its temperature and creating darker regions on the photosphere. These regions are called sunspots and are the most prominent features of ARs. The formation of sunspots is illustrated in Figure 2.7. ARs usually appear as pairs of sunspots with opposite polarity: one forms where magnetic field lines emerge outward through the photosphere, while the other forms where they penetrate back into the Sun. The magnetic field lines connecting these two spots are usually twisted and form a magnetic loop above the photosphere.

Sunspots are not the only photospheric features witnessing magnetic activity of the Sun, as *faculae* are also the result of magnetic field lines. Faculae appear as bright regions in the photosphere due to the concentration of magnetic field lines in bundles narrower than those in sunspots. While present across the entire solar surface and not limited to sunspot neighbor-

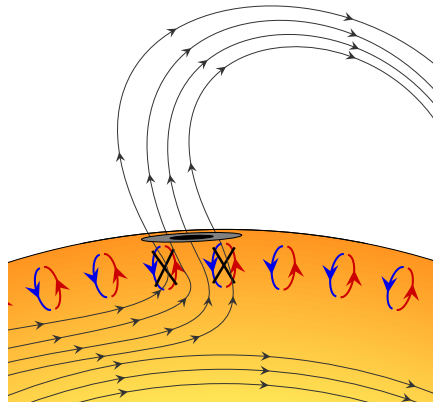


Fig. 2.7 Cross-sectional diagram of sunspot formation mechanism. Strong magnetic flux tubes emergence disrupts convection cell in convective zone. Broken cells induce thermal inhibition regions with brightness reduction on the surface above called sunspots.

hoods, faculae are most prominently visible near the limb of the solar disk, where their contrast against the darker background is enhanced compared to disk center. This is due to an observation effect known as *limb darkening* or *center-to-limb variation* (clv). Limb darkening is the result of observing deeper, hotter and brighter layers of the Sun when looking at its center, while observing shallower, cooler and dimmer layers when looking at its limb.

The magnetic structures that create these photospheric features extend upward into the solar atmosphere. ARs can thus also be seen in the chromosphere through observations in specific spectral lines (see Section 2.3). In the chromosphere, sunspots also appear as dark regions, and faculae have a counterpart named *plages* that tend to be more extended than faculae. The larger thickness results from the magnetic canopy effect: as magnetic field lines from BMRs spread out with increasing altitude, they form an umbrella-like structure in the chromosphere. This expansion of the magnetic field leads to additional heating of the overlying plasma through mechanical and magnetic energy dissipation. Above the chromosphere, ARs expand into large magnetic loops in the corona known as 'coronal loops'. These loops are filled with hot plasma and are confined by the magnetic field lines of the AR.

2.2.3 Sunspot Group Lifecycle and Classification

Sunspot groups evolve from initial photospheric pores, small dark features that mark the emergence of magnetic flux. These pores either undergo rapid dissolution within hours or develop into mature sunspot groups through sustained field loop emergence and consolidation over days. The developmental phase shows clear asymmetry, with leading spots demonstrating better coherence and longevity than their trailing counterparts.

Mature sunspots display characteristic morphological features. They are usually composed of two regions corresponding to two shade levels: the *umbra*, an intensely dark region at the center of the sunspot, and the *penumbra*, a region surrounding the umbra with brightness halfway between umbra and quiet Sun (see Figure 2.6). Penumbrae present radial flow of matter in filamentary structures known as Evershed flows, characterized by outward horizontal plasma motions from the umbra through the penumbra. The nature and cause of these flows are still open questions but are thought to be caused by magnetic field lines inclination in the penumbra.

The magnetic configuration of a sunspot group starts as a simple bipolar structure - where magnetic field lines emerge through the photosphere at one location and penetrate back at another. This evolves toward enhanced complexity through secondary flux emergence within the AR such that sunspot groups can show various aspects. The decay phase of ARs proceeds through systematic fragmentation, with trailing spots dissolving first while leading spots often maintain their integrity longer. The complete evolutionary sequence typically spans multiple days to weeks, with exceptional cases preserving coherence across several solar rotations.

Solar physicists use standardized classification schemes to categorize these stages. The Zürich system, introduced by Waldmeier in 1947, establishes a framework utilizing designations A through J to characterize developmental stages and magnetic complexity. The McIntosh scheme [24] used in this thesis builds upon this foundation by incorporating additional morphological parameters, including penumbral development characteristics and spot distribution patterns. On the other hand, the Mount Wilson scheme [25, 26] classifies ARs based on the magnetic polarity of the sunspots that compose them and their distribution.

McIntosh Classification

The McIntosh scheme is a widely used system for classifying sunspot groups based on their morphological characteristics. A visual representation of the McIntosh classification scheme is shown in Figure 2.9. This scheme provides a detailed description of sunspot group complexity and distribution using a three-letter *Zpc* code that captures key physical features.

Z-component: The first letter of the code describes the sunspot group's overall appearance and stage of development. It roughly corresponds to the *Zürich* classification scheme and ranges from class A to H¹.

- A** Single sunspot or small group not showing bipolarity similar to pore. No visible penumbra. Shows birth or late stage of decay of an AR.
- B** Bipolar group with no penumbra. Shows further emergence of a BMR loop.
- C** Bipolar group with penumbra in at least one of the spots. Shows the growth phase of an AR.
- D** Both ends of the bipolar group have penumbra. Interior has width $\leq 10^\circ$.
- E** Bipolar group with penumbra in both spots. Interior has $10^\circ < \text{width} \leq 15^\circ$.
- F** Bipolar group with penumbra in both spots. Interior has width $> 15^\circ$. This is the stage with highest level of complexity, often source of flares or CMEs.
- H** Unipolar group with penumbra. Shows decay phase of earlier bipolar group.

p-component: The second letter characterizes the *penumbral* structure of the largest sunspot in the group. It describes the presence or absence of penumbra, its shape being either symmetric or asymmetric, and the north-south penumbral diameter.

- x** No penumbra - only for groups of class A and B.
- r** Rudimentary penumbra - the developing penumbra does not present filamentary aspect.
- s** Penumbra has filamentary structure and is symmetric. The north-south diameter is $\leq 2.5^\circ$.
- k** Penumbra is well developed and is symmetric. The north-south diameter is $> 2.5^\circ$.
- a** Penumbra has filamentary structure and is asymmetric. The north-south diameter is $\leq 2.5^\circ$.
- h** Penumbra is well developed and is asymmetric. The north-south diameter is $> 2.5^\circ$.

¹Classes G and J in Zurich scheme can be recovered using appropriate combination of *Zpc* components.

c-component: The third letter describes the *compactness* of the group. It denotes the distribution of minor sunspots in the group interior -i.e space between the main spots of opposite polarity.

- x Undefined - only for unipolar groups (class A and H)
- o Open - zero to few minor spots sparsely distributed.
- i Intermediate - minor spots in group interior, none of them has a penumbra.
- c Compact - Group interior is populated with spots and at least one of them has a filamentary penumbra.

This classification system provides a comprehensive description of sunspot group morphology, enabling detailed analysis of the physical processes underlying solar magnetic activity.

Mount Wilson Classification

While the McIntosh scheme classifies sunspot groups by morphology, the Mount Wilson scheme utilizes magnetic polarity information derived from magnetograms—specialized observations revealing the strength and polarity of solar magnetic fields. In his work, Hale introduced this 4-class classification system based on magnetic configurations of increasing complexity [25]:

- α : Unipolar ARs with single magnetic polarity. Typically small and simple. Minor opposite polarity elements may exist but are disregarded if significantly smaller than primary components.
- β : Bipolar ARs exhibiting leading and trailing spots of opposite polarities with distinct separation. Groups display slight tilt relative to the solar equator. Most common Mount Wilson classification.
- $\beta - \gamma$: Complex bipolar ARs where opposite polarities appear in both leading and trailing regions. Polarity separation exhibits complex configuration (irregular inversion or multiple lines).
- γ : ARs with irregularly distributed polarities precluding clear bipolar classification.

The scheme was later augmented by Künzel with the δ suffix appended to indicate enhanced flaring probability [26, 27, 28]:

- δ : AR containing at least one mature sunspot with multiple umbrae of opposite polarities within a single penumbra, separated by less than 2° .

The complete Mount Wilson classification comprises seven classes: α , β , $\beta\delta$, $\beta - \gamma$, $\beta - \gamma\delta$, γ , and $\gamma\delta$. These classes are illustrated in Figure 2.9.

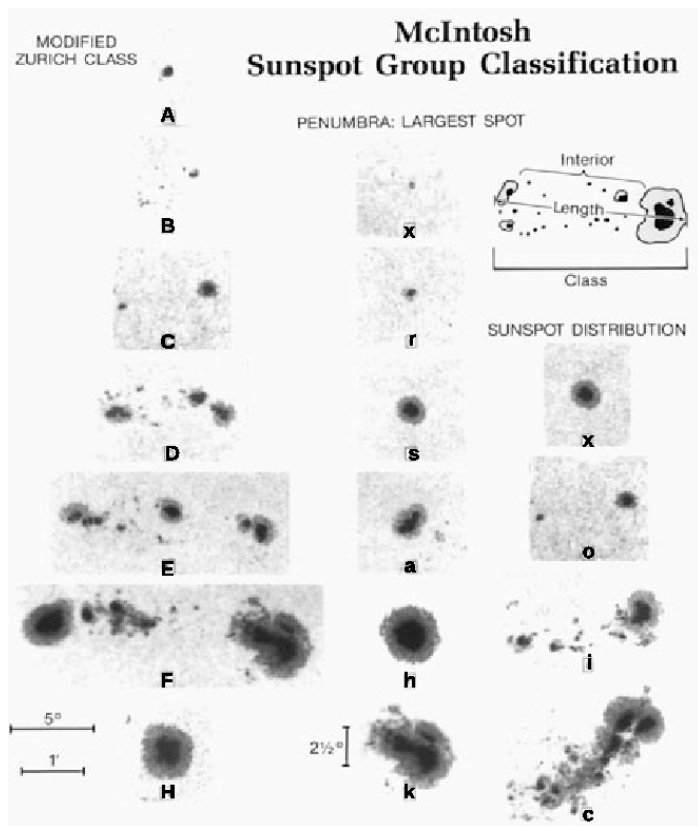


Fig. 2.8 McIntosh classification scheme for sunspot groups. The three-letter code captures key morphological features, including overall complexity based on Zurich scheme, penumbral structure of largest spot, and compactness of group interior. Source: [24]

2.3 Observation Tools and Datasets

Solar physics has evolved through distinct technological eras. Ancient astronomers relied on naked-eye observations. They documented transient phenomena including sunspots and eclipses. The development of optical telescopes in the 17th century initiated quantitative solar astronomy. Subsequently, spectroscopic innovations enabled chemical composition analysis. Each technological advance expanded our diagnostic capabilities.

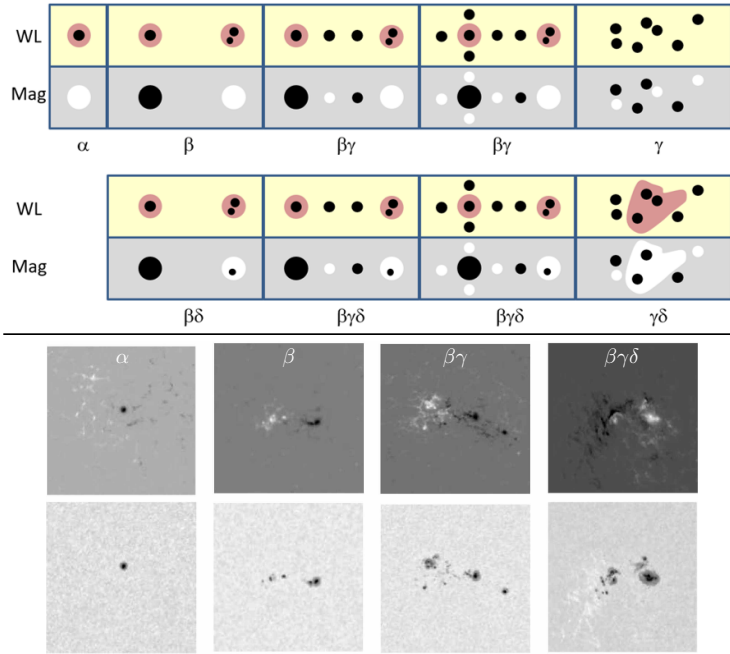


Fig. 2.9 (Top) Mount Wilson classification scheme for sunspot groups. The scheme is based on the distribution of magnetic polarity among the sunspots in the group. Image from <https://www.stce.be/educational/classification>. (Bottom) Examples of Mount Wilson classifications. Top row shows magnetograms and classes, bottom row shows the corresponding AR in WL. Image from [29].

Optical ground-based facilities are used either for synoptic observations or high-resolution imaging. In the first case, full-disk images allow to study solar cycle and monitor space weather. In the second case, the large aperture provide observations with unprecedented details. Sensor calibration allows capturing specific layers of the Sun. The photosphere is typically studied through "white light" (WL) or "continuum" observations, as this layer emits across the entire visible spectrum. The chromosphere, on the other hand, is often observed through the calcium II-potassium spectral line (CaII K, 393.4 nm) and the H α line (656.3 nm), each revealing different heights in the solar atmosphere. Observations through these lines reveal complementary aspects of the chromosphere: in addition to sunspots,

CaII K imagery shows bright plages due to magnetic field concentrations, while H α imagery displays spicules that reflect plasma-magnetic field interactions. In this manuscript, we focus on observations of the chromosphere through CaII K line imagery.

Multi-wavelength analysis now drives solar research methodology. It provides a more complete view of the Sun's atmosphere and magnetic field. An example of this is the construction of magnetograms, images that map the strength and direction of magnetic fields across the Sun's surface from several spectral lines observations. The foundation of magnetogram creation is the Zeeman effect [30], a phenomenon where spectral lines split when atoms are in a magnetic field. To measure magnetic fields, Stokes parameters, which describe the polarization state of light, are used [31]. These consist of Stokes I (total light intensity), Stokes Q and U (linear polarization in different orientations), and Stokes V (circular polarization). Circular polarization is particularly important for measuring the magnetic field component along line-of-sight (LOS), while linear polarization helps determine the field orientation in the plane of the sky. Creating magnetograms involves several key processing steps [31]:

1. A magnetically sensitive spectral line is selected, typically iron or nickel.
2. Images through different polarization filters are captured and measurements at specific wavelength positions around the chosen spectral line are taken.
3. The data undergoes careful calibration to remove instrumental effects, account for atmospheric distortion in ground-based observations, and correct for observatory motion.
4. The polarization measurements are then converted into magnetic field values through inversion of Stokes parameters [31, 32], often using the Milne-Eddington approximation which models the solar atmosphere with specific assumptions.
5. These values are used to generate visual representations, solving 180° ambiguity in the magnetic field direction.

This process results in magnetograms, where different colors or shades represent magnetic field strength and polarity.

With technological development and space programs, space-based instruments revolutionized solar physics through continuous observation outside of the Earth-Sun line. Each mission addresses distinct heliophysical objectives. The space instruments allowed the observation and study

of new aspects of the Sun such as coronal dynamics, magnetic field evolution and solar wind properties. Ground facilities excel in high-resolution imaging while space observatories access otherwise obscured wavelength regimes. This complementary approach enables comprehensive solar atmospheric modeling. Integration of these datasets advances our understanding of solar processes.

2.3.1 Ground-based Observatories

The Sun is observed from numerous ground-based facilities across all continents. Solar monitoring began in the 17th century using telescope projections and drawings, with systematic sunspot index established in the 19th century through the work of Rudolf Wolf. Modern solar observatories now employ either radio telescopes for studying solar emissions, optical telescopes with digital cameras for surface imaging, or both types of instruments depending on their research focus.

For example, a well-known network of ground-based instruments is the Global Oscillation Network Group (GONG, [33]), which comprises six identical telescopes located around the world. The GONG provides synoptic observations of the Sun as well as magnetograms targeting the nickel Ni I spectral line at 676.8 nm for space weather forecasting and helioseismology research.

The ROB, from which the data used in this thesis originates, is another leading solar physics institution. Its ground-based station USET (Uccle Solar Equatorial Table) captures optical observations in WL for the photosphere and $H\alpha$ and CaII K-line for the chromosphere [34], enabling precise solar tracking and activity monitoring². Details on observation equipment of USET are provided in Appendix A. ROB's daily WL observations date back to the 1950s through drawings, transitioning to Charge-Coupled Device (CCD) cameras in 2002, with daily $H\alpha$ and CaII K-line observations added in 2012 for comprehensive chromospheric monitoring.

A notable contribution of ROB to solar physics is the curation of the international Sunspot Number series [35, 36], which unifies historical records including Wolf's method. This metric, maintained by World Data Center

²USET Database, available at <https://vo-tap.oma.be/>

2 | Heliophysics Background

SILSO [22] at ROB, is openly available. ROB's extensive archives support long-term solar cycle studies, and its internationally recognized program continues to foster scientific collaboration worldwide [37, 38].

Atmospheric artifacts

Observing the Sun from the ground presents challenges due to artifacts introduced by Earth's atmosphere. These artifacts include seeing effects, cloud cover, and airborne objects such as dust, animals and aircrafts. Seeing effects are caused by atmospheric turbulence. They distort and blur solar images, reducing spatial resolution. Clouds and aircraft can obstruct observations, limiting data collection. These artifacts are illustrated in Figure 2.10.

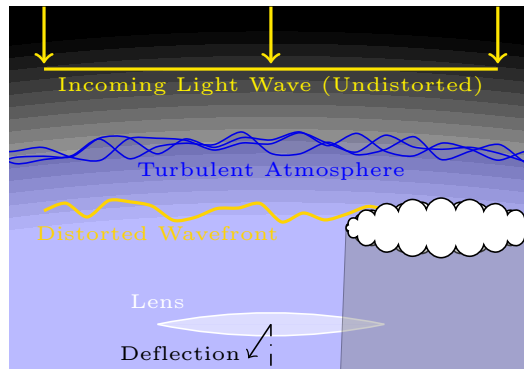


Fig. 2.10 Artifacts affecting ground-based solar observations. (a) Seeing effects due to atmospheric turbulence alters incoming light wavefront and cause distortion and blurring on image captures. (b) Masking due to flying objects (clouds, aircrafts, dust...).

2.3.2 Space-based Observatories

To overcome atmospheric limitations, solar physics research has expanded into space with the help of space programs. Space-based observatories provide continuous solar monitoring. These missions offer high-resolution imaging and spectroscopic capabilities. Among others, a few of these observatories are the Solar Dynamics Observatory (SDO, [39]), the Solar and Heliospheric Observatory (SOHO, [40]), and the Interface Region Imag-

ing Spectrograph (IRIS, [41]) and Solar TERrestrial Relations Observatory (STEREO, [42]). The fleet of space-based missions currently used in heliophysics research is shown in Figure 2.11.

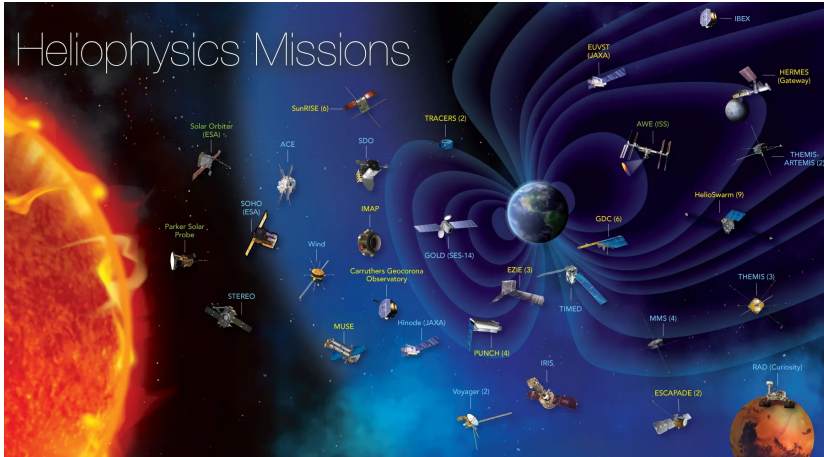


Fig. 2.11 The fleet of space missions observing the Sun as of July 2024. Image from <https://science.nasa.gov/learn/heat/missions/>

These missions embark various sensors that capture solar images across multiple wavelengths [43, 44, 45] (visible light, extreme ultra violet/EUV, ...) and viewpoints [42] that allow monitoring and reconstruction of the Sun's magnetic field and its evolution (magnetograms). For example, the SDO mission embarks the Helioseismic and Magnetic Imager (HMI), which provides provides full-disk magnetograms from multiple wavelength observations of the iron spectral line (FeI 617.3 nm), as well as the Atmospheric Imaging Assembly (AIA), which provides high resolution images of the full sun in multiple wavelengths, including seven EUV wavelength bands (centered at 9.4, 13.1, 17.1, 19.3, 21.1, 30.4, and 33.5 nm).

The datasets produced by these missions are essential for solar physics research. The data captured by space-based observatories is used worldwide for research and space weather forecasting. Large parts of the images can be obtained through the open python SunPy library [46]. Also, due to the rising interest of machine learning, dedicated datasets such as SDOML [47] are made publicly available and curated for machine learning

applications.



Summary

This chapter has explored the fundamental aspects of heliophysics, from the internal structure of the Sun to its atmosphere, active regions, and observable features like sunspots. We have examined the magnetic phenomena that drive solar activity and the various classification systems used to understand solar behavior. Understanding these complex solar phenomena relies on robust observational capabilities. The diverse observational network available composed of ground-based and space-based platforms allows for comprehensive studies of the Sun and they have significantly enhanced our ability to study it.

These complementary observation sources generated new volumes of heterogeneous data capturing solar structure, atmospheric dynamics, magnetic activity, and eruptive events across multiple scales. While carefully crafted traditional rule-based models have provided valuable insights, the scale, complexity, and high dimensionality of modern observational datasets may exceed the capabilities of predefined rules. This creates an opportunity to shift heliophysics research towards more recent data-driven techniques like deep learning, which revolutionized many domains. Deep learning capabilities are particularly well-suited to these challenges. They excel at processing high-dimensional data and images, detecting subtle features, and modeling complex phenomena.

The next chapter presents these computational approaches that could fundamentally enhance pattern recognition and feature extraction in solar physics datasets. The subsequent chapters of this thesis then specifically explore the application of these techniques to several use cases involving historical records of ARs observations, though observations of other solar phenomena such as solar flares and CMEs may similarly benefit from deep learning approaches.

3

Deep Learning Background

The advent of powerful computing resources and large storage solutions brought significant change to engineering and problem-solving applications. Initially, automated solutions to complex problems were computed algorithmically or following human-defined rules in a field that received the (overhyped) name *Artificial Intelligence*. Later, rule-based approaches have gradually given way to data-driven methods grouped under the *Machine Learning* terminology and further evolved into *Deep Learning* [48]. This field has revolutionized artificial intelligence, particularly in the domains of visual content analysis and image generation. This chapter provides a comprehensive overview of deep learning fundamentals and their applications in computer vision tasks.

Section 3.1 establishes core concepts of deep learning as well as neural network supervision fundamentals. Section 3.2 examines deep learning-based visual content analysis through its foundational network architectures, and training strategies. The considered approaches range from Convolutional Neural Networks to Vision Transformers. Section 3.3 addresses image generation applications, spanning from image restoration to cross-domain translation.

3.1 Deep Learning Basics

Machine learning algorithms are designed to automatically learn statistical rules that allow to address some task, based on data. A wide variety of algorithms was developed with the years. A few of those algorithms are k-nearest neighbors (k-NN [49, 50]), Mean-Shift [51, 52], Support-vector machines (SVM [53]) and neural networks (NN, [48]) that paved the way to deep learning. While each algorithm has its strengths, neural networks stand out for their versatility and scalability based on simple building blocks. This section introduces these building blocks and their associated terminology both intuitively and formally.

Neuron and Activation The smallest mathematical building block of NNs is the neuron, also called perceptron [54]. It computes the weighted sum of its inputs $x_1, \dots, x_N$ and some inner bias b then computes an output named *activation* denoted y :

$$y = \text{Act.}(z(x)) = \text{Act.}\left(\sum_i w_i x_i + b\right) \quad (3.1)$$

where Act. is an activation function and z is the composition of weighted sum of inputs and bias addition. Illustration of a neuron and common activation functions are shown in Figure 3.1. Functions commonly chosen as activation functions are the identity function, sigmoid, tanh, ReLU (rectified linear unit [55], expressed by $\text{ReLU}(x) := \max(0, x)$) and its variants. The task addressed by the neuron drives the choice of activation function. As an example, sigmoid or tanh fit classification tasks as the output is constrained in some interval. On the other hand, identity function and ReLU fit regression purposes. **Those logical neurons were designed to mimic the way biological neurons work as they fire a signal through their axone based on input signals received at their dendrites.**

Multi-Layer Perceptron, Neural Networks and Deep Learning By design, a neuron can model a restricted set of simple functions as it relies on a linear combination of its inputs. To approximate complex functions, neurons are combined in networks, often organized in layers. In one layer l of a neural network with L layers, each neuron takes the output of every neuron of the previous layer $l - 1$ as inputs. This configuration is called a

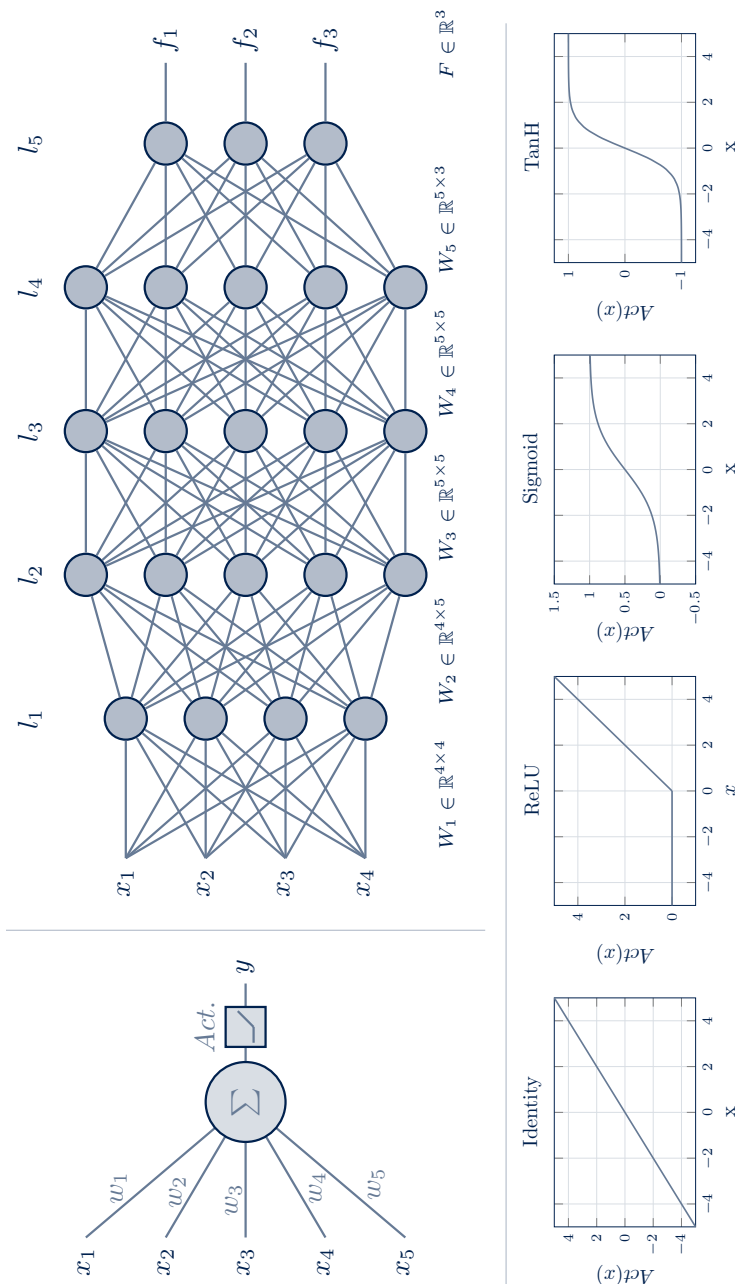


Fig. 3.1 (Left) Illustration of single neuron. (Right) Illustration of a neural network (NN) approximating 3-dimensional function F with multiple layers. Layers l_2 to l_4 are hidden layers. The dimensions of weight matrix is shown under layer connections, no biases are considered. (Bottom) Common activation functions used in neural networks.

fully connected (or dense) layer. Formally, with N_l (resp. N_{l-1}) the number of neurons in layer l (resp. $l-1$), the activation of a dense layer can be expressed using matrix representation

$$Y_l = \text{Act.}(W_l Y_{l-1} + B_l) \quad (3.2)$$

where $B_l \in \mathbb{R}^{N_l}$ and $W_l \in \mathbb{R}^{N_l \times N_{l-1}}$. Activation function Act. is applied element-wise on matrix result. In a network with L consecutive fully-connected layers, Y_0 is the input signal and Y_L is the network output.

Stackings of such layers constitute a Multi-Layer Perceptron (MLP [48, 56]). An illustration of an MLP is shown in Figure 3.1. Layers that are neither the first nor the last are called *hidden layers*. The number of layers L determines the network's expressivity —i.e. ability to approximate complex functions and capture intricate patterns in data. Deep learning leverages this principle by employing multiple hidden layers in NNs, enabling the modeling of highly non-linear relationships. With large-scale datasets and modern computing resources, MLPs demonstrated superior performance compared to previous machine learning algorithms on complex tasks. Beyond their practical success, theoretical proofs established that these models can approximate arbitrarily complex functions, provided that they are composed of enough neurons [57, 58].

Parameters and Hyperparameters Neural networks are characterized by two distinct sets of variables. Parameters, consisting of weights and biases in the neurons, are optimized during training. **Parameters correspond to the "rules" learnt from the data.** Hyperparameters shape the network's architecture and learning process, such as the number of layers and neurons. **Hyperparameters correspond to the number of rules the network is able to learn and the method to learn them.** While parameters values emerge through optimization, hyperparameters require manual tuning or automated search strategies.

Model Supervision and Training Procedure Deep neural networks can be trained under different paradigms called *supervision methods*. Supervised learning leverages input samples paired with some labels considered as the ground truth (GT). Unsupervised learning discovers patterns from unlabeled data. Semi-supervised approaches combine both labeled and

unlabeled samples. This thesis addresses only tasks in the supervised setting, which is the most common paradigm.

In supervised setting, the available data is split into three distinct sets: training, validation, and test sets. The training set is used for learning the parameters. The validation set guides model selection and hyperparameter tuning throughout the training process. The test set provides an unbiased evaluation of the final model performance. Prior to any training, hyperparameters must be selected to fix both the network architecture and training configuration. The training procedure then starts by randomly initializing the network parameters. These parameters are then iteratively adjusted through backpropagation and gradient descent optimization. At each iteration, a batch of training samples is processed, and the parameters are updated to minimize the objective -or loss- function.

Loss function and Backpropagation In supervised learning, network predictions must be compared to ground truth annotations to guide parameter updates. This measure, called the loss function (often shortened to *loss*), varies with the learning task. Classification tasks typically use cross-entropy loss to measure differences in probability distributions, while regression tasks employ mean squared error (MSE) or L1 loss. Multiple losses can be combined in a weighted sum:

$$\mathcal{L}_{Tot} = \lambda_1 \cdot Loss1(\hat{y}, y) + \lambda_2 \cdot Loss2(\hat{y}, y) + \dots \quad (3.3)$$

where y is the ground-truth label, $\hat{y}$ is the model prediction and λ_i is a hyperparameter weighting the i -th sub-loss.

Once defined, this loss guides network optimization through *backpropagation* [56], a four-step process repeated numerous times throughout training:

1. *Forward Pass*: Compute predictions (i.e., model output) from inputs.
2. *Error Measurement*: Compute the loss $\mathcal{L}$ at the output layer
3. *Backward Pass*: Compute gradients of the loss $\mathcal{L}$ for all parameters via backward error propagation
4. *Update*: Modify parameter values using gradients.

Step 3 is the cornerstone of neural network learning capability. The backward pass measures each parameter's influence on $\mathcal{L}$, enabling precise

network-wide adjustments. This is possible because neural networks are compositions f of differentiable functions: sums, product and activation functions. Therefore, given a differentiable loss $\mathcal{L}$, the chain rule allows computation of gradients with respect to each parameter, weight or bias. For example, using Stochastic Gradient Descent [59], the weights W_l of layer l are updated at each iteration t by processing mini-batches of training data:

$$W_l^{t+1} \leftarrow W_l^t - \eta \nabla_{W_l} \mathcal{L} \quad (3.4)$$

where η is the *learning rate* (LR) and $\nabla_{W_l} \mathcal{L}$ is the derivative of $\mathcal{L}$ with respect to weights W_l . Processing data in mini-batches offers a good trade-off between computational efficiency and optimization stability compared to using single samples or the full dataset. While this is the basic approach, modern deep learning often uses advanced optimizers like Adam [60] or AdaGrad [61] with adaptive learning rates. **Regardless of the optimizer, making this optimization process gradual is fundamental, as abrupt parameter changes would prevent the network from learning the underlying statistics of the data distribution.**

Evaluation Metrics Neural networks can be evaluated through various metrics. While the loss function value on the test set measures the optimization objective directly, it may not align with application-specific requirements. Classification tasks commonly use accuracy, precision, and recall, as these metrics are more interpretable than cross-entropy loss. In object detection and segmentation, the Intersection over Union (IoU) metric often quantifies the quality of spatial predictions. Image generation models pose unique evaluation challenges, requiring either human assessment or specialized metrics to measure both sample quality and diversity. The choice of evaluation metrics ultimately depends on the task's nature and the model's intended use.

Inference The forward pass process of a trained neural network is called *inference*, representing the operational phase where the model is deployed to solve real-world tasks. It transforms input data into predictions in a deterministic manner by applying the learned weights and biases across network layers that have been fixed. Unlike training where parameters are updated, inference only applies the learned representations to generalize to new, unseen data samples.

3.2 Deep Learning for Visual Content Analysis

While previous sections introduced deep learning core concepts, this section focuses on visual content analysis using deep learning models. Images are fundamentally grids of pixels and can be represented as tensors that can be 2D for grayscale images or 3D for color images (RGB channels). In this thesis, only grayscale images are considered. Visual data presents unique challenges due to high-dimensional inputs with spatial (image) and temporal (video) consistency requirements. MLPs are ill-suited for such data: they would require an impractical number of parameters and fail to exploit the local patterns that characterize visual information -e.g. edges, corners, shade, texture-. Imagine the computational burden in an MLP: one neuron per pixel in the first layer alone, multiplied across numerous layers, with additional irrelevant connections between network regions. Therefore, specialized architectures like Convolutional Neural Networks [62] and Vision Transformers [63] were developed to address these limitations by efficiently capturing spatial dependencies, enabling advances in tasks from classification to semantic segmentation.

This section reviews popular architectures for visual content analysis, alongside data augmentation techniques for improved performance.

3.2.1 Architectures for Visual Content

From their traditional computer vision origin, we first introduce the two most popular deep learning architectures used for image processing: Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). Traditional computer vision [64] utilizes various small-scale filters (e.g. Sobel filters), known as *kernels*, to capture local patterns in images. A kernel operates as a small grid of fixed weights that, when applied to a same-size pixel region in an image, highlights the presence of one specific visual pattern or *feature*. Different kernels are designed to detect distinct features such as edges, corners, or textures. The convolution operation systematically slides these kernels across the image, producing *feature maps* that highlight the detected patterns throughout the image. Figure 3.2 illustrates the convolution process and shows different feature maps obtained with different kernels.

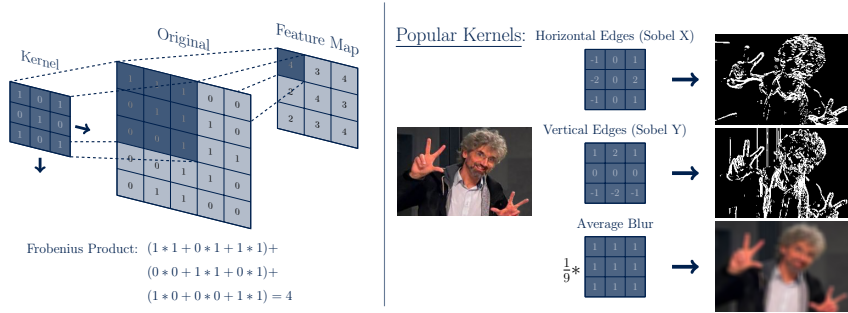


Fig. 3.2 (Left) Computation of one feature map element using a sliding kernel. (Right) Example of feature maps obtained by convolving common kernels on an input image.

CNNs build upon these traditional methods, while ViTs offer a novel attention-based approach to extract information from images. Both architectures are composed of multiple layers, each designed to capture different levels of abstraction in the input data. The hierarchical structure of these networks enables the progressive extraction of increasingly complex features, from simple edges in early layers to high-level semantic concepts in deeper layers.

Convolutional Neural Networks CNNs [62] innovate by learning convolution kernel weights directly from data, adapting to input-specific patterns more effectively than traditional hand-crafted kernels. **Learning kernels can be viewed as training small, moving MLPs that detect specific features.** These learned kernels are small and operate locally on the image. They maintain translation invariance, enabling processing of varying image sizes - solving the scalability limitation of MLPs acting on full-size images. Multiple kernels generate multiple feature maps, each capturing a low-level pattern, which are hierarchically combined in subsequent layers to detect increasingly complex, higher-level image characteristics. Formally, a convolutional layer can be expressed as:

$$Y_l = \text{Act.}(W_l \circledast Y_{l-1} + B_l) \quad (3.5)$$

where $\circledast$ denotes the convolution operation, W_l is the convolution kernel,

B_l is the bias and Y_{l-1} is the input feature map. The activation function $Act.$ is applied element-wise on the resulting feature map.

The receptive field defines the region of an input image that influences a specific feature map element. Pooling operations leverage this concept by reducing spatial dimensions and combining feature map information, progressively expanding the receptive field to enable detection of increasingly abstract patterns. Max-pooling selects regional maximum values, while average-pooling computes regional means. CNNs hierarchically stack convolutional and pooling layers to encode data into a low-dimensional embedding space, where backpropagation guides the network to learn features of growing complexity—from low-level edges in early layers to high-level, task-specific abstractions in deeper layers. Figure 3.3 illustrates a typical encoding CNN architecture.

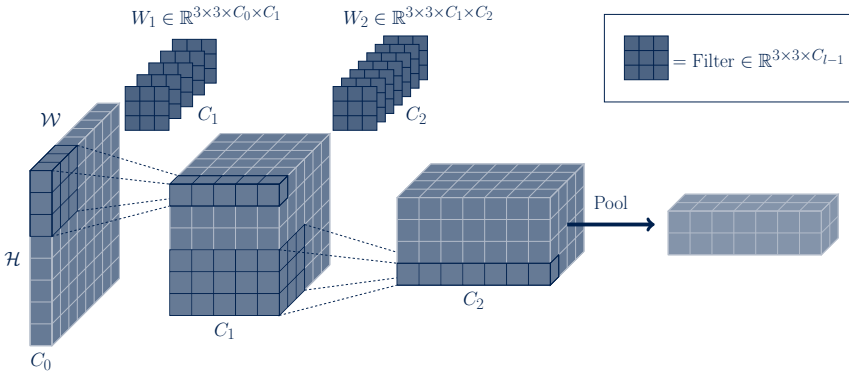


Fig. 3.3 Illustration of a CNN encoding a single-channel image into low-dimensional feature maps through two successive convolution layers followed by one pooling layer. Layers l_1 and l_2 learn C_1 and C_2 kernels respectively, with the total number of trainable parameters for each convolution shown between the feature maps.

The specific task dictates latent map processing and network architecture. Classification and regression tasks typically convert low-dimensional feature maps into vectors, then use fully connected layers for prediction. Notable networks like AlexNet [65], VGG [66], ResNet [67], and EfficientNet [68] follow this approach. Conversely, tasks producing images—such as semantic segmentation or image sample generation—use upsampling

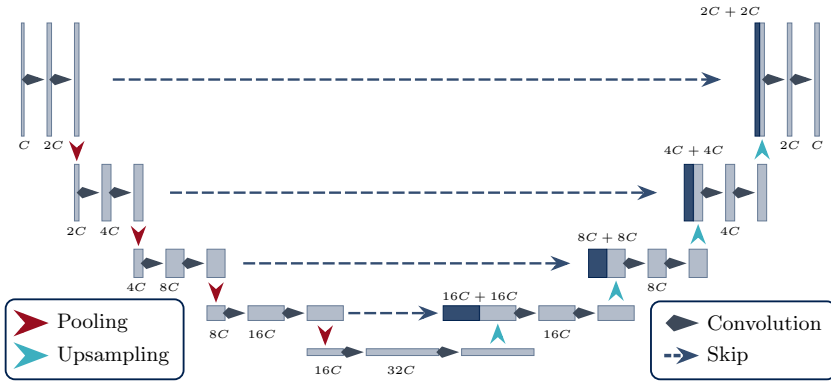


Fig. 3.4 Illustration of the U-Net architecture of depth 4. Kite-shaped arrows represent a convolutional layer followed by a ReLU activation function. The last feature map of each level is either downsampled using max-pooling (encoding part) or upsampled using transposed convolution (decoding part). At each level, one skip connection concatenates the feature map from the encoder with the feature map from the decoder.

and convolutional operations to decode the latent space into the visual output space. In these cases, the embedding space isn't necessarily low-dimensional, as it must retain all information, including spatial, needed for image generation.

Originating in medical imaging, the U-Net architecture [69] became widely used for image segmentation tasks. It employs a symmetric encoder-decoder structure with skip connections, concatenating encoder-level feature maps with corresponding decoder-level maps to preserve spatial information during downsampling. This thesis adopts the U-Net architecture for its specific segmentation task. The U-Net architecture is illustrated in Figure 3.4.

Feature Map Modulation Feature map modulation is a technique used to enhance the representational power of CNNs by dynamically adjusting the intermediate feature maps based on additional information. This approach allows the network to adaptively emphasize or suppress certain features, improving its ability to capture complex patterns in the data. One common method for feature map modulation is through the use of *Feature-*

wise Linear Modulation (FiLM) [70], which applies an affine transformation to feature maps conditioned on external inputs.

Specifically, FiLM operates at designated locations within the network architecture, referred to as *modulation points*. At each modulation point, the additional information is processed through a linear layer to generate channel-wise scaling ($\gamma_{i,c}$) and shifting ($\beta_{i,c}$) parameters. This channel-wise modulation allows for fine-grained control over individual feature channels. For a feature map $\mathbf{F}_{i,c}$, a FiLM layer applied the transformation

$$FiLM(\mathbf{F}_{i,c} | \gamma_{i,c}, \beta_{i,c}) = \gamma_{i,c} \times \mathbf{F}_{i,c} + \beta_{i,c} \quad (3.6)$$

where i indicates the instance, c represents the channel, and $\gamma_{i,c}$ and $\beta_{i,c}$ are the channel-wise scaling and shifting parameters produced by the linear layer.

This modulation mechanism enables the network to dynamically focus on features during each forward pass, providing specific adaptation at several key computational stages. By incorporating FiLM layers into CNN architectures, researchers have achieved significant improvements in performance across various visual content analysis tasks, such as object detection, image segmentation, and visual question answering.

Visual Transformers The ViT [63] adapts the Transformer architecture from natural language processing [71] to visual content analysis. Fundamentally based on the self-attention mechanism, it treats input data as a sequence, enabling dynamic capture of both short and long-range dependencies by focusing on different input sequence parts. Formally, self-attention computes attention weights through three learned linear transformations: query (Q), key (K), and value (V) matrices. The *attention* is calculated as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (3.7)$$

where d_k is the dimensionality of the key vectors and softmax is an activation function. This mechanism allows each sequence element to attend to all others, creating a dynamic, context-aware representation.

ViTs [63] process images by dividing them into fixed-size patches and transforming these into a learned vector representation acting like a code-

book of image elements. Positional encodings —i.e., learnable or fixed embeddings— are added to patch vectors to preserve spatial information, providing a unique location signature that enables the self-attention mechanism to maintain spatial relationships between patches. The Transformer encoder consists of a succession of blocks, each containing a multi-head self-attention mechanism and a feed-forward MLP. These embedding vectors are processed through this encoder structure, with the output of the last block feeding into a classification or regression head for prediction.

Compared to CNNs, ViT offers advantages including task flexibility, ability to capture local and global image dependencies, and computational efficiency through parallel processing. However, ViTs require extensive training data and significant computational resources, making them less suitable for data- or resource-constrained environments. Consequently, given the limited data available for this thesis, results of ViT networks are not presented.

3.2.2 Visual Data Augmentation

Deep learning models require substantial data to learn effectively, with data requirements varying by model complexity and task difficulty. While common RGB image tasks have abundant training data [72], specialized domains like medical or astronomical imaging struggle with data scarcity due to collection costs and expert annotation limitations [73, 74]. Several techniques have been explored to reduce annotation costs and make the most of available data. For example, *active learning* identifies and annotates the most informative unlabeled samples during model training [75, 76, 77] - typically those the model processes with lowest confidence. By targeting these challenging cases where the model is most uncertain, active learning achieves larger performance gains with fewer labeled samples. *Data augmentation*—creating new data samples from existing ones— also appears as a solution to address these data limitations. By introducing controlled variations, it not only increases dataset size but also serves as a form of implicit regularization. This approach encourages models to learn invariant features and develop more robust representations that generalize beyond specific training instances, effectively expanding the model’s understanding of the underlying data distribution.

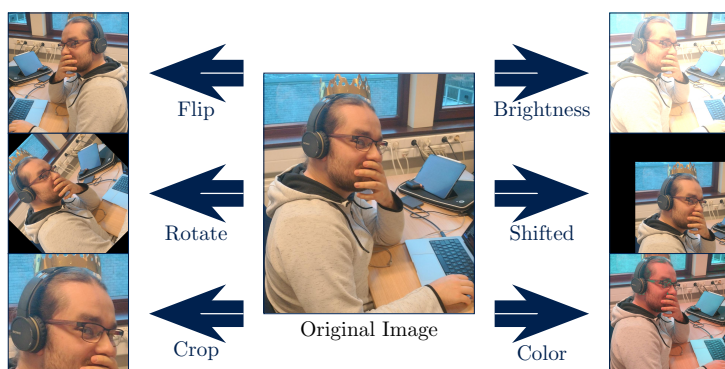


Fig. 3.5 Common data augmentation techniques for visual data.

In machine learning, data augmentation artificially expands training datasets by applying strategic transformations. These transformations introduce variability that helps models learn robust features. Non-visual tasks can augment data by adding noise, creating samples with missing values, or combining multiple samples (e.g. bagging, SMOTE [78]).

In visual tasks, data augmentation helps models learn invariance to image transformations. Augmentations can modify image nature (pixel space/distribution) or image content (semantic information). Examples of the former are color space transformations, noise addition and sample synthesis. Examples of the latter are geometric transformations, occlusions and cropping. Color space transformations include conversion to grayscale, changing the color balance, brightness, contrast, or saturation of the image. Noise addition can include Gaussian noise, salt-and-pepper noise, or quantization (uniform) noise. Several data augmentation techniques are illustrated in Figure 3.5.

3.3 Deep Learning for Image Generation

Image generation tasks are a subset of visual content analysis that focus on creating new images from existing ones or from scratch. These tasks are particularly challenging, as they require models to not only learn the

underlying data distribution but also produce visual content across potentially different data distributions. Image generation tasks can be broadly categorized into three main types: image restoration, sample generation, and image-to-image translation. Examples of these categories are shown in Figure 3.6. Each task presents unique challenges and requires distinct architectural and training approaches to achieve high-quality results.

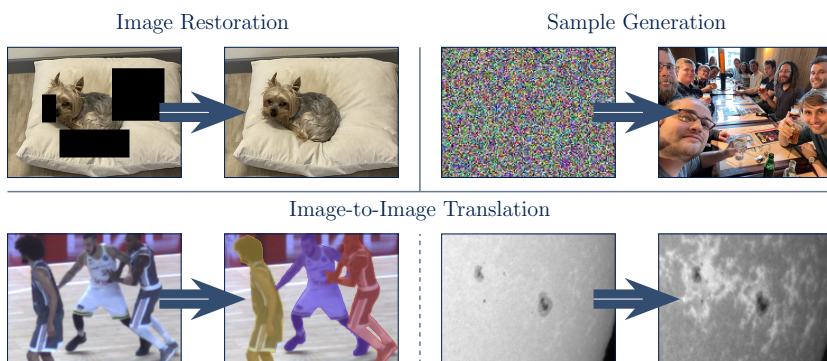


Fig. 3.6 Illustration of common visual tasks addressed with deep learning. (Top Left) Inpainting. (Top Right) Image generation from noise (Bottom Left) Image instance segmentation. (Bottom Right) Spectral band translation.

3.3.1 Use Cases

Image Restoration

Image restoration, a fundamental task in traditional computer vision [79, 64], recovers original images from corrupted or degraded versions. Common restoration tasks include denoising, super-resolution, and inpainting. Denoising [80] removes noise arising from acquisition (fog, brightness degradation), transmission (data transfer corruption), or storage (compression artifacts) processes. The challenge lies in discriminating noise patterns from image content while preserving crucial information.

Super-resolution increases image resolution by generating new pixels while maintaining visual fidelity [81]. This task demands understanding

low-level image patterns to create samples that are both visually appealing and accurate to the original image's semantic content.

Inpainting fills missing image regions or replaces content, removing unwanted objects or completing incomplete images [82]. It requires understanding of image structure and semantic context to generate coherent modifications. The task involves reasoning about texture, edge continuity, and contextual information to create seamless and realistic reconstructions.

These restoration tasks share a fundamental challenge of preserving original image content while generating semantically coherent modifications. Performance is assessed by comparing reconstructed images to the GT using metrics like Peak Signal-to-Noise Ratio (PSNR), Mean Square Error (MSE), Structural Similarity Index Measure (SSIM,[83]), and other perceptual quality measures that quantify the restoration's fidelity [84, 85].

Sample Generation

Sample generation tasks involve two key steps: first learning to represent the data distribution in a latent space, then using this learned representation to create new images. Generative models like VAEs [86], GANs [87], and other autoregressive models learn to map the original data into a low-dimensional space, capturing the underlying statistical properties of the dataset (see Section 3.3.2). The decoding network of these models then transforms randomly sampled latent vectors back into the original data space, generating new images that reflect the learned data distribution.

The latent space represents compressed, semantically meaningful data [88, 89]. By sampling and decoding this space, models generate instances capturing statistical properties of the dataset [87, 90]. This process challenges models to maintain data fidelity while avoiding mode collapse—losing original data diversity by generating only a limited sample subset. In supervised settings, *conditioning* mechanisms can even guide generation by introducing constraints like class labels, enabling sample generation with specific characteristics while preserving the distribution's overall statistical properties.

The samples generated by these models can be used for various purposes, either as is or as data augmentation for other tasks (see Section 3.2.2).

Sample quality is evaluated through statistical similarity, visual fidelity, and diversity.

Image-to-Image Translation

Image-to-image (I2I) translation tasks aim at producing visual content from a visual input space. It encompasses a wide variety of tasks, such as image segmentation, style transfer, and modality translation. **One can picture image-to-image translation as sample generation process of the target modality using images of the source modality as conditioning.**

Image segmentation transforms images into pixel-wise classification masks. It comprises two primary approaches: semantic segmentation [91, 92], which assigns pixels to abstract classes (e.g., human, car), and instance segmentation [93], which uniquely identifies individual class instances (e.g., human #1, human #2). Segmentation is a cornerstone of object detection and delineation, image editing, and comprehensive image understanding.

Style transfer modifies an image's visual aesthetic while maintaining its core content [94, 95]. It is useful for creating artistic images or transferring styles between images. **Imagine taking a picture of your pet and have a CNN applying the style of a famous painter to it.** The process requires two images: a content source and a style reference. Style transfer models extract content features from the content image and stylistic features from the style image, then generate a novel third image blending the first image's content with the second image's visual style.

Modality translation extends style transfer principles to a broader range of image transformations, potentially more serious and less artistic. It converts images between different visual representations, such as sensor types, color spaces, or imaging modalities. Typical applications include medical image synthesis (e.g., CT to PET scans [96]), remote sensing (spectral band translations [97]), and thermal-to-visible imaging [98] for surveillance. In supervised contexts, models are trained on paired images, learning to map content across modalities while maintaining fundamental image information. This technique enables critical preprocessing and cross-modal adaptations in various scientific and technological domains.

3.3.2 Frameworks/Paradigms

Deep learning paradigms for image generation have evolved significantly over time. The earliest paradigm for visual content prediction was the *Encoder-Decoder* architecture, exemplified by autoencoders and U-Net [69]. An encoder compresses input data into a compact latent representation, while a decoder reconstructs the output from this encoding. This architecture proved effective for tasks requiring deterministic input-output mapping, such as segmentation and style transfer. *Generative Adversarial Networks* [87] marked a paradigm shift by using adversarial training between generator and discriminator sub-networks. This framework, combined with random sampling of the latent space, enabled the generation of realistic yet diverse images. More recently, *Diffusion Models* [99] emerged as a powerful alternative to GANs. DMs operate by learning to reverse a gradual noise-addition process, implementing a sequence of denoising steps to transform random Gaussian noise into coherent images. Empirical studies demonstrate that DMs achieve superior sample diversity and training stability compared to GANs, while producing higher quality outputs than Encoder-Decoder architectures [100].

Encoder-Decoder Networks

Encoder-Decoder (ED) networks form the foundation of many visual content generation approaches (see Figure 3.7). **In EDs, the encoder part compresses input images into a compact representation, while the decoder part reconstructs it.** The encoding is typically a series of convolutional layers and pooling operations, while the decoding is a series of convolutional layers and upsampling operations such as transposed convolutions [101] or bilinear interpolation. This structure may be symmetric, though encoder and decoder architectures often differ based on task requirements.

Skip connections between corresponding encoder and decoder depth levels marked a key advancement in ED networks. These connections provide direct paths for high-resolution features, combining spatial details with semantic content from the bottleneck. This principle, pioneered by U-Net [69], has become fundamental in modern architectures processing visual content.

Training EDs typically relies on reconstruction losses (e.g., L1, L2) mea-

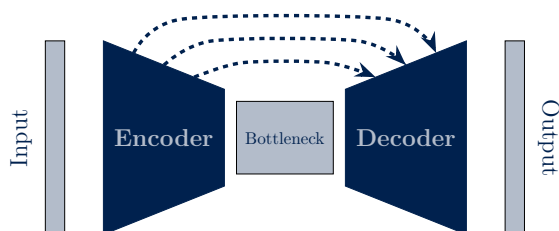


Fig. 3.7 Encoder-Decoder (ED) architecture overview. The encoder (left) progressively reduces spatial dimensions while increasing the number of features, creating a compact latent representation. The decoder (right) reverses this process to reconstruct the output. Optional skip connections (dashed lines) allow fine detail preservation.

asuring pixel-wise differences between input and output. However, these losses often lead to blurry results as they average multiple possible solutions. To address this limitation, perceptual losses operating on deep features from pretrained networks are commonly used in natural images, capturing semantic similarities rather than pixel-wise differences [84, 85]. Most implementations leverage features from VGG networks [66] that are pretrained on general-purpose dataset ImageNet [72], as these have been shown to effectively capture perceptual and semantic image characteristics.

ED networks are foundational to autoencoders [88], which learn to reconstruct their inputs while performing dimensionality reduction. Key variants include Variational Autoencoders (VAE) [86], which learn probabilistic latent representations, and Vector-Quantized VAE (VQ-VAE) [102], which discretize the latent space using learned codebooks. The resulting latent spaces enable semantic manipulation through vector arithmetic and interpolation, making autoencoders valuable for both dimensionality reduction and content editing.

These ED architectures exhibit inherent characteristics that influence their application domains. Their deterministic nature suits tasks requiring precise input-output mappings, such as segmentation [69], denoising [89, 80] and style transfer [94]. However, this determinism limits their ability to generate diverse outputs, as identical inputs produce identical outputs.

Generative Adversarial Networks

GANs [87] are based on two competing neural networks: a *generator* creating samples of a desired data distribution and a *discriminator* distinguishing real from generated samples. This creates a min-max game where an increasingly convincing generator tries to fool an increasingly stricter discriminator. The generator learns a mapping from random noise vectors to an output distribution, while the discriminator learns a binary classification task: distinguishing between real and generated samples. Figure 3.8 illustrates this training paradigm. **In simple terms, the discriminator tells if generated data looks real or fake, and the generator uses this feedback to improve its creations.**

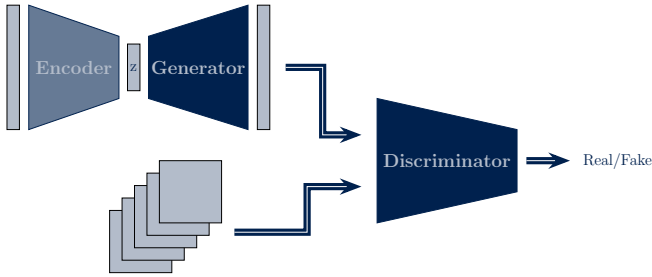


Fig. 3.8 Generative Adversarial Networks paradigm overview. The generator (left) creates realistic images, while the discriminator (right) learns to distinguish between real and generated samples. The generator input is typically random noise, or a latent representation of some input image in image-to-image tasks.

Training alternates between two steps: the discriminator updates its parameters according to its classification errors, then the generator updates its parameters according to the discriminator's classification of samples it generated. At convergence, the generator should produce samples indistinguishable from real data.

Formally, let $\mathcal{X}$ be the data domain and $\mathcal{Z}$ the space randomly sampled, with distributions p_{data} and p_z respectively. The discriminator $D : \mathcal{X} \rightarrow [0, 1]$ and generator $G : \mathcal{Z} \rightarrow \mathcal{X}$ are trained using adversarial loss:

$$\min_G \max_D \mathbb{E}_{x \sim p_{data}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z)))] \quad (3.8)$$

The discriminator maximizes this objective, while the generator minimizes it. Conditional GANs (cGAN, [103]) incorporate auxiliary information $y \in \mathcal{Y}$ to control generation. The networks become $G : \mathcal{Z} \times \mathcal{Y} \rightarrow \mathcal{X}$ and $D : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$ to account for additional information:

$$\min_G \max_D \mathbb{E}_{(x,y) \sim p_{data}} [\log D(x, y)] + \mathbb{E}_{z \sim p_z, y \sim p_{data}} [\log(1 - D(G(z, y), y))] \quad (3.9)$$

where (x, y) are paired samples from the joint data distribution and y can be class labels, style codes, or even input images. In the latter case, $\mathcal{Z}$ is not a noise space randomly sampled, but the latent space encoding the the input modality.

While GANs have demonstrated impressive results in image generation, they face significant challenges [90]. Mode collapse occurs when the generator focuses on a limited set of outputs that fool the discriminator, undermining the goal of capturing the full data distribution. Training instability can arise from poor balance between generator and discriminator capabilities. In conditional generation tasks like image-to-image translation [104], the consistency between input and output is not inherently guaranteed, leading to hallucination: generated outputs may be locally plausible yet deviate from the input conditioning, particularly due to the stochastic latent space. **Think of GANs as overeager artists: they sometimes add details that look good but aren't faithful to the requested directions.** Various architectural improvements address these challenges: Wasserstein GANs [105, 106] provide more stable training through improved distance metrics, while cycle consistency losses and self-attention mechanisms help maintain faithful and coherent generation.

The impressive results of GANs have led to their broad adoption across numerous domains from medical imaging to content creation. However, their tendency to prioritize visual appeal over faithful reconstruction raises concerns, particularly for scientific applications where accuracy is crucial and erroneous details could lead to misinterpretation.

Diffusion Models

DMs convert noise into meaningful data through an iterative denoising process based on Markov chains. Unlike GANs and VAEs, DMs [99] achieve stable training by decomposing generation into small reversible steps. The framework operates in two phases: a *forward process* that systematically adds noise to training data, and a *reverse process* that learns to recover the clean signal.

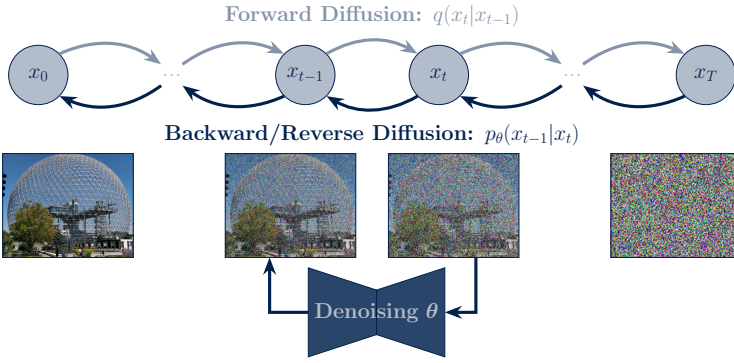


Fig. 3.9 Diffusion Model paradigm overview. *Forward* process degrades samples by adding noise. *Reverse* process uses model θ to recover original image by iteratively removing noise.

The forward process corrupts data via a Markov chain of Gaussian noise additions. Beginning from a clean sample x_0 , the process incrementally adds noise parameterized by a variance schedule $\beta_1, \dots, \beta_T$, transforming the data to pure noise after T steps. Schedules like linear, cosine, and learned approaches showed varying impacts on performance. Each transition step follows a Gaussian distribution:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I) \quad (3.10)$$

where β_t controls the noise variance. The complete forward process is therefore a cumulative product $q(x_{1:T}|x_0) = \prod_{t=1}^T q(x_t|x_{t-1})$. A key advantage of using Gaussian noise is that we can directly sample any timestep:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)I) \quad (3.11)$$

with $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$ accumulating the noise effect. This mathematical

convenience significantly improves training efficiency. It also leads to a simple sampling equation:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon \quad , \quad \epsilon \sim \mathcal{N}(0, I). \quad (3.12)$$

The reverse process is a Markov chain that iteratively removes noise using a neural network. In visual tasks, the denoising network θ typically follows a U-Net architecture. The network predicts and subtracts noise at each step, recovering cleaner data. **In simple terms, the reverse process uses a deep learning model iteratively on the noised sample to recover the original data.** This process allows reconstructing original data and generating new samples matching the data distribution, starting from pure Gaussian noise without forward process components.

Formally, the reverse process employs a neural network with weights θ that predicts the mean μ_θ of the noise distribution at each step, while considering a fixed noise variance σ_θ equal to β_t . The reverse diffusion kernel $p_\theta(x_{t-1}|x_t)$ is defined as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t)I). \quad (3.13)$$

A single denoising step using this kernel computes:

$$x_{t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sqrt{\beta_t} \epsilon \quad (3.14)$$

where $\epsilon_\theta(x_t, t)$ predicts noise at step t . This denoising step is applied repeatedly to recover the original data. At each step, training minimizes the mean squared error between the noise predicted by the network and the noise added during the forward process:

$$\mathcal{L} = \mathbb{E}_{t, x_0, \epsilon} \left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2 \right]. \quad (3.15)$$

DMs have demonstrated remarkable success across multiple domains. In computer vision, DALL-E 2 [107] and Stable Diffusion [108] excel at image generation and editing, while models like InstructPix2Pix [109] enable text-guided manipulation. Biomedical applications include tissue images sample generation [110, 111], protein design [112] and drug discovery [113]. In audio synthesis, models like AudioLDM [114] achieve text-to-speech and music generation.

However, DMs face two significant limitations. First, generation typically requires a thousand denoising steps, making real-time applications impractical. Early works like DDIM [115] reduced it to hundreds of steps with reasonable quality before approaches using principles of distillation were introduced to address this issue. Distillation approaches train student models that emulate the teacher model’s denoising capabilities in fewer steps.

Building upon a pretrained pixel-space model requiring thousands of denoising steps, [116] develops a cascade of student models where each successive model halves its predecessor’s sampling steps. This progressive compression of the sampling trajectory enables thousand-step diffusion models to generate high-quality outputs with only 4-8 steps. [117] extended this distillation framework to latent-space DMs, making it compatible with guided diffusion approaches such as classifier-free guidance. Notably, these first two methods focus primarily on simplifying the sampling procedure at the cost of longer training times. Complementary to these approaches, [118] proposed a two-phase distillation strategy where the classifier-free guidance component is first distilled while keeping the main pretrained DM frozen, followed by distillation of the sampling procedure similar to previous works. This approach not only accelerates sampling but also achieves model size reduction, addressing multiple efficiency concerns simultaneously.

Finally, taking distillation to its theoretical limit, Distribution Matching Distillation approaches [119] bypass the step-by-step process entirely by training one-step generators that directly match the final output distribution of fully-trained diffusion models rather than mimicking the intermediate denoising process. These methods typically employ direct distribution matching techniques, divergence minimization objectives, and adversarial training elements [120]. The main challenge these approaches face is maintaining the quality and diversity of generated samples without the iterative refinement that makes diffusion models so powerful.

The second significant challenge for DMs is their substantial data requirements, particularly for high-dimensional data like images. Training effective diffusion models on such data demands massive datasets of millions of samples—a fundamental limitation arising from the need to learn complete denoising trajectories across all noise levels. Unlike sampling

speed issues, this data efficiency problem remains largely unsolved and represents an active research frontier. On the other hand, simpler low-dimensional data such as tabular or vector representations remain tractable with smaller datasets.

3.3.3 Quantitative Evaluation of Generative Models

Generative models are evaluated on two key properties: quality and diversity. Quality measures how realistic or accurate the samples are, while diversity assesses how well the model covers the target distribution. Their importance varies by application. In unconditional generation, both are balanced, as generating high-quality but nearly identical samples (mode collapse) is just as problematic as producing diverse but unrealistic outputs. On the other hand, scientific applications prioritize quality for accurate details, and creative applications favor diversity for artistic expression.

Direct Comparison Metrics

One way to evaluate generative models is through direct comparison metrics, which measure the similarity between the statistical properties of generated samples and the target distribution:

- Maximum Mean Discrepancy (MMD, [121]) quantifies distribution similarity by comparing sample pairs in feature space, with lower values indicating better similarity.
- Precision and Recall metrics measure quality and diversity separately —precision shows what fraction of generated samples resemble real data, recall indicates how much of the real distribution is captured, helping diagnose mode collapse (high prec, low rec.) or poor quality (low prec., high rec.).
- Structural Similarity Index Measure (SSIM, [83]) evaluates perceptual image differences based on luminance, contrast, and structure, particularly useful for structural fidelity assessment.
- Kernel Inception Distance (KID, [122]) offers unbiased distribution comparison even with limited amount of samples, making it suitable for computationally constrained scenarios.

Deep Feature-Based Metrics

To improve alignment with human perception, pre-trained models are used in deep feature-based metrics for evaluating generative models:

- Inception Score (IS, [90]) uses a pre-trained classifier to measure quality through confident predictions and diversity through varied predicted classes. Scores range from 1-10 with higher values indicating better performance, though it doesn't directly compare against real data.
- Fréchet Inception Distance (FID, [123]) has become the standard metric, extracting features from real and generated images via an Inception network, then calculating the Wasserstein-2 distance between their multivariate Gaussian distributions. Lower scores indicate better similarity, with sensitivity to both artifacts and mode collapse.
- Learned Perceptual Image Patch Similarity (LPIPS, [84]) measures perceptual similarity by comparing deep neural network activations instead of pixel values. This better approximates human visual perception by focusing on meaningful features rather than exact pixel matches. LPIPS is especially valuable for image reconstruction and translation tasks, with lower values indicating greater perceptual similarity.

Limitations

While automated metrics provide valuable insights, certain limitations may affect their consistency and reliability in evaluation:

- **Feature Space Dependency:** Major metrics rely on ImageNet-trained networks, which may not capture domain-specific characteristics for specialized applications. These generic feature spaces often miss the most relevant quality aspects in fields like medical imaging or artistic content.
- **Insensitivity to Certain Artifacts:** Metrics like FID often miss subtle artifacts or semantic inconsistencies obvious to humans. Improvements in these metrics frequently don't correlate with human preferences, creating challenges for model selection.
- **Computational Cost:** Reliable evaluation requires thousands of samples, becoming prohibitively expensive for high-resolution images and limiting thorough assessment in resource-constrained settings.

Human Evaluation

Human evaluation remains essential for generative models despite automated metrics' advances. It captures nuanced quality aspects that algorithms miss, especially for creative applications. Assessment methods include direct quality ratings and comparative preference selection, with standardized protocols like Human eYe Perceptual Evaluation (HYPE, [124]) improving reliability by controlling factors such as exposure time and question framing. Though providing invaluable perceptual insights, human evaluation faces three major challenges: high costs, limited scalability, and subjective variability across evaluators.

Directions for Enhanced Evaluation

Research on general-purpose image datasets now employs a hybrid approach: automated metrics for development and human evaluation for validation. As generative models advance, three promising evaluation directions emerge:

- **Task-Specific Metrics:** Tailoring evaluation to the model's intended application rather than using generic measures. Scientific applications should prioritize accuracy metrics, while creative applications might emphasize aesthetic qualities, enabling more meaningful real-world assessment.
- **Perceptual Alignment:** Following the path of HYPE, developing metrics with stronger correlation to human judgments by training specialized evaluation models, incorporating vision science insights, or creating composite metrics that capture multiple perceptual dimensions.
- **Adversarial Evaluation:** Training specialized discriminators to detect specific failure modes and quality issues that current metrics miss, providing targeted assessment of model improvements and remaining challenges.

As the field matures, robust, perceptually aligned, task-appropriate evaluation will be essential for meaningful progress in generative modeling.



Summary

This chapter has covered the foundational concepts of deep learning, from basic neural network architectures to specialized frameworks for visual content analysis and image generation. We explored convolutional architectures that efficiently extract hierarchical features from images, enabling the detection of increasingly complex patterns at different scales. Additionally, we examined techniques for data augmentation that address training data limitations—a critical consideration in scientific domains where labeled data may be scarce. The chapter concluded with an overview of generative approaches that facilitate the creation of synthetic imagery, while examining the limitations of existing quality assessment approaches for generated content.

These theoretical foundations provide the necessary groundwork for the next technical chapters, which will demonstrate how these methods

can be adapted to study solar ARs through specific use cases. Implementation frameworks and quantitative evaluation methodologies will be tailored to address the unique challenges of solar physics data, including variable image quality, feature ambiguity, and the need for domain-specific performance metrics that align with solar physics objectives.

4

Study of sunspot groups in continuum imagery

Having established the theoretical foundations of both heliophysics and deep learning in the previous chapters, we now turn our attention to the practical integration of the latter in solar physics research. This chapter first examines how deep learning can serve as an assistive tool for scientists analyzing observations of active regions on the solar photosphere. We demonstrate how these computational methodologies can be effectively inserted into the scientific workflow while minimizing the integration burden and annotation workload typically associated with machine learning applications. This approach aims to complement human expertise in addressing the complex challenges of solar phenomena interpretation, without any intention of replacing scientific experts, and without imposing prohibitive data preparation requirements.

To understand the significance of this application domain, we must consider the solar phenomena under study. Solar flares and CMEs, which strongly affect Earth's space environment, originate from solar magnetic ARs, manifested as groups of dark regions on the solar photosphere named sunspot groups. The appearance and size of these regions provide important clues about their eruptive potential, drawing significant scientific attention since the late 19th century. Studying sunspot evolution on a long-

term basis therefore became keystone to several areas of Solar Physics, from helioseismology to solar irradiance modelling, which is directly linked to the evolution of the Earth. To characterize the observed ARs, different classification schemes were developed based on different properties.

As previously introduced in Section 2.2.3, the McIntosh classification system with its 3-component [Zpc] structure [125, 126, 24] provides a morphological characterization of sunspot groups and their associated active regions. Complementing this approach, the Mount Wilson scheme [25, 27] classifies sunspots according to their magnetic properties, with studies demonstrating that larger sunspot areas produce flares of greater magnitude during their evolution [28].

Statistical studies have shown that sunspot groups with higher McIntosh structural complexity classes (corresponding to larger longitudinal extent, large and asymmetric penumbrae, and more internal spots) produced higher flaring rates overall. The McIntosh scheme is therefore at the basis of several flare forecasting methods which estimate the flare occurrence rate from historical records of flares and sunspot group classes [127], possibly combining such information with observed waiting time distribution between flares [128, 129] or with the evolution of sunspot groups classification [130].

Attributing a McIntosh class to a sunspot group is thus fundamental to predicting its impact on the Earth space environment. Traditionally, such attribution was done manually, and inserted into catalogs and databases along with metadata related to the number of spots and groups, position, and area of each sunspot. Examples of such catalogs include the merged catalog from [131], the US Air Force/SOON catalog ¹, Debrecen's Photoheliographic data (DPD), Hungary [132, 133], the Spanish observatories catalogs [134, 135, 136], and the Coimbra Observatory catalog, Portugal [137].

In this chapter, we use solar observations obtained from the ground-based station USET located at ROB and we propose SunSCC, a fully automated system addressing three essential tasks regarding the observation of Sunspots on WL images: Segmentation, Clustering of individual sunspots into sunspot groups, and Classification of sunspot groups according to the

¹SOON SRS catalog, <https://www.swpc.noaa.gov/products/solar-region-summary>

McIntosh classification scheme [24]. We focus on the McIntosh classification scheme as it is based solely on WL information, unlike the Mount Wilson scheme, and as it has applications in space weather forecasting. To our knowledge, this is the first attempt to achieve these tasks using ground-based WL observations only. Most of previous works cited in Section 4.1 use space-based instruments, and possibly also magnetogram information. As in many other ground-based stations around the world, USET has been observing the Sun for decades without interruption. This allows for a consistent record and statistical studies of sunspot groups over several solar cycles, unlike most of space based instruments which observe during a smaller period of time.

SunSCC works in a cascaded sequence: the output of the segmentation task is given as input to the clustering task, and the clustering is used to locate sunspot groups for which the classification is estimated. Both the segmentation and the classification use CNNs, more precisely the segmentation task uses U-Net [138], and the classification is based on ResNet [139] followed by Multi-Layer Perceptron [48]. Interestingly, our work demonstrates that ensembles of classifiers are effective in differentiating reliable from probably incorrect classifications, based on the consistency of classifier predictions in the ensemble.

For the aggregation of sunspots into groups, we adapt the Mean-Shift algorithm [51] to the specificities of sunspot groups definition, using the results provided by the segmentation CNN.

We show that, despite the fact that our original data are prone to significant noise due to atmospheric seeing or clouds (see Section 2.3.1), and is constituted only of WL observations, our classification method achieves similar performance to methods applied on spaced-based images, possibly using both continuum and magnetogram information. As WL solar observations are extremely common, many solar observing stations around the globe can deploy the SunSCC pipeline.

This chapter is organized as follows. Section 4.1 motivates our work and provides a survey of previous related works. Section 4.2 presents the ground-based observations, and details which specific datasets were used in each task. Section 4.3 presents the SunSCC pipeline, and the methods used to achieve the three tasks, including the preparation of the necessary

training dataset. Quality metrics and a discussion of the experimental results are provided in Section 4.4. We conclude with future perspectives in Section 4.5.

4.1 Motivation and related works

4.1.1 Sunspot segmentation

Accurate detection of sunspots is fundamental for the construction of solar activity indices such as the International Sunspot Number [140], and constitutes the first step towards sunspot classification.

The delineation and classification of sunspots is however partly subjective as two distinct observers might differ in their decisions. Moreover, sunspots are irregular in shape, and have variable intensity and contrast with their surroundings. Ground-based images typically also have different levels of noise and distortion. In such a non-purely deterministic context, it is particularly useful to rely on the intrinsic statistical robustness of machine learning tools. Automated methods also allow to deal with the ever growing amount of data available and to perform statistical studies on large datasets.

Both region-based and edge-based automated detection methods were used in previous works. Region-based methods encompass on one hand histogram-based segmentation, where pixels are classified according to their intensities [141, 142, 143] possibly combining this with a Bayesian approach [144], and on the other hand region-growing procedures, where the connectivity of individual pixels is used to incorporate information about the local neighborhood [145].

Edge-based methods focus on discontinuities and thus on locating region boundaries. [146] used edge detection, followed by mathematical morphology on continuum image from the Michelson Doppler Imager (MDI, [147]) space instrument onboard the SOHO mission [40]. They compared their results with NOAA synoptic maps produced manually during one year, and found good correlation. Later on, [148] as well as [149] used mathematical morphology on WL images, before applying thresholding

to isolate the sunspots. [148] applied his algorithm on WL images from a ground-based station with the goal to reproduce the Wolf number and to compare it against value obtained manually by operators of the same station. [149] applied his method on a 1997-2003 MDI continuum dataset to investigate the effect of center to limb variation on sunspot appearance and disappearance.

In their work, [150] provided a comparison between intensity-based thresholding and mathematical morphology methods. It showed that mathematical morphology-based approaches perform better than purely intensity-based methods. They also require less pre-processing to remove atmospheric effects in ground-based observations and to correct for limb darkening.

These methods are however still very much dependent on the choice of parameters such as the size of the morphological operators. It is therefore of interest to develop a method which is robust against particular heuristics or parameter's choice. We tackle this problem by considering multiple thresholds per WL image to supervise the training of our machine learning algorithm, as described in Section 4.3.2. The resulting trained neural networks are shown to aggregate and regularize the knowledge extracted from the segmentation masks obtained with multiple thresholds (see Section 4.4.1).

4.1.2 Grouping of individual sunspots into sunspot groups

Sunspots are formed in groups that share physical properties such as belonging to the same magnetic flux loops, and are more elongated in longitude than in latitude. Considering a binary mask of sunspots (where '0' means a background pixel and '1' a sunspot foreground pixel), sunspot groups correspond to local maxima in the spatial distribution of foreground pixels inside this binary mask.

Several attempts were made in previous works to find such local maxima. In their work, [148] used a region growing procedure and predefined criteria of neighborhood to group individual sunspots into a sunspot group. [151] tested various hierarchical algorithms and a simple k-means algorithm.

Since the number of sunspot groups on a given day is not known, the number of clusters k to request from the k-means algorithm can not be defined, making this method impractical. Similarly, finding an appropriate termination condition for the hierarchical algorithms proved to be difficult. The same authors then compared in [152] the performance of several hierarchical algorithms with that of a density-based clustering algorithm known as DBSCAN (introduced in [153]). A slightly better performance for the latter was observed.

Finding the local maxima of the statistical distribution that best explains an observed set of samples in a multidimensional space is precisely the problem solved by the so-called ‘mean-shift’ algorithm [52]. Another reason to choose this algorithm in this work is that it allows for a natural integration of prior knowledge about anisotropic elongation of clusters. This makes it particularly suitable for our purpose since sunspot groups present dominant longitudinal elongation.

4.1.3 Sunspot Classification

Attributing a McIntosh class to a sunspot group is fundamental for science and space weather purposes. Most of the time, however, such classification is still being done manually by experts. This is a time-consuming process, and although classification rules are well defined, there is room for subjectivity and interpretation, and experts may disagree on sunspot group separation and classification. This in turn results in inconsistencies that stem from these human observation bias as well as non-reproducible catalogs.

To solve the inconsistency problem faced by various solar observatories and space-weather-prediction groups around the world, and to deal with the large amount of solar data available, automated classification of sunspot groups has been envisioned. A combination of thresholding and 1-layer artificial neural networks to detect, cluster, and classify sunspot groups according to the McIntosh system was proposed in [154]. Their work, still considered as a benchmark today, is using both continuum and magnetogram data from the MDI instrument. Support Vector Machines on extracted features from MDI continuum images to predict the modified Zurich classification using a one-against all method were used in [155]. The authors claimed a score of average 80% precision but the method was

applied to a restricted dataset (20 samples per class).

The advent of Deep Learning methods allowed computational models that are composed of multiple processing layers to learn representations of data with multiple and hierarchical levels of abstraction [156, 157]. In particular, CNNs have proven superior to previous methods for image classification [158]. In sunspot classification, [159] used a pre-trained deep neural network and a transfer learning approach to provide an automated McIntosh classification using MDI as well as HMI (Helioseismic and Magnetic Imager, [160]) magnetograms. They achieved an average 50% precision score, and significant confusion can be seen between the classes *D*, *E*, and *F* of the modified Zurich class. [161] address sunspot detection and classification using continuum and magnetogram images as well as the Solar Region Summary (SRS) tables of the US Air Force/SOON catalog ² as a source of information about the observed sunspot groups. These tables report the location, composition and McIntosh classification of the visible sunspot groups on a daily basis. A Faster R-CNN model is used to detect sunspots, and Inception v3 for classification. The precision score varies between 30% and 90% amongst classes, and similarly to [159], confusion between classes *D*, *E*, and *F* of component *Z* appears clearly.

4.2 Data

4.2.1 USET input dataset

Images used in this work were collected by the USET [162] station at ROB in the 2002-2019 time period. We used full sun images from the USET WL instrument, where images have been pre-processed to center the solar disk in the image at a 1 pixel level accuracy, and the images are translated to have the center in the middle of the image. This dataset is openly available for registered users through the SOLARNET Virtual Observatory [163].

The USET station is also equipped with a Visual White light telescope, that projects the image of the Sun on a paper and is used to produce sunspot drawings. Since the 1940s, human operators have made annotations on these drawings regarding the sunspot groups and total sunspot counts.

²<https://www.swpc.noaa.gov/products/solar-region-summary>

Name	Task	Time period(s)	#samples	Annotation type	Annotation origin
WL_{SA}	Seg. (Train)	2013-15	5953	Mask	Automated
WL_{SM}	Seg. (Val.)	2013-15	36	Mask	Manual
WL_C	Clust.	2002-12 $\cup$ 2016-19	3150	BBox	SGDB
WL_C	Classif.	2002-12 $\cup$ 2016-19	5017	McIntosh	DDB, SGDB

Table 4.1 Summary of the datasets used in this work. BBox stands for ‘Bounding Box’. DDB and SGDB are, respectively, the drawing data base and sunspot group database gathered from USET observations.

The information contained in these annotations, as well as some information about the drawings have been included in two linked databases: First, the Drawing Database (DDB), which stores the global metadata such as the observer’s acronym, total number of spots, sunspot groups, and Wolf number pertaining to one solar drawing, and second the Sunspot Group Database (SGDB), whose entries contain specific information about a sunspot group, such as its heliographic coordinates, area, or manually made classification in the the Zurich and McIntosh systems. Using information contained in the SGDB and the value of the solar radius retrieved from Visual WL observations, it is also possible to determine a Bounding box (BBox) for each sunspot group. Catalogs pertaining to these two DBs have been included in two Table Access Protocol (TAP) services³. The USET sunspot group catalog stemming from the SGDB was also used in [37], [38] and [131].

In this chapter, the size of the WL images were made uniform to 2048×2048 pixels. Indeed, due to a change of camera equipment in June 2008, the image sizes are inconsistent: images recorded earlier than June 2008 are of 1024×1024 pixels size while later ones are of size 2048×2048 . To avoid any issue caused by a resolution change, we upsampled earlier images to match the size of later ones using nearest neighbour interpolation.

We detail below which subset will be used for the different tasks of segmenting, clustering, and classifying the sunspots. The characteristics of the various datasets used in this work are summarized in Table 4.1.

4.2.2 Segmentation dataset

From the entire dataset, images from the 2013-2015 period were used to train and estimate the CNN-segmentation model. The chosen period cor-

³USET Database, available at <https://vo-tap.oma.be/>

responds to the peak solar activity during solar cycle 24, offering thus a decent diversity of sunspots and sunspot groups in terms of number and shape.

From the images in this period, a total of 36 images (one image per month) were extracted for segmentation performance validation. These 36 images were annotated manually by ROB experts and constitute the ground-truth for assessing the CNN segmentation's performance. This manually annotated dataset is denoted $\mathcal{WL}_{SM}$. All other images, except the ones captured the same day or the day before or after the one of a test image, form the training set $\mathcal{WL}_{SA}$, and were processed automatically to generate pseudo-labels to train our CNNs for segmentation, as explained in Section 4.3.2.

Once the CNN-segmentation model has been estimated, it is applied on the WL images within the period $2002-12 \cup 2016-19$ (outside the 2013-2015 years) and produces segmentation masks to be used for clustering and classification tasks. The set of WL images recorded during this period, and thus used for both clustering and classification tasks, is denoted $\mathcal{WL}_C$.

4.2.3 Clustering dataset

For the clustering task, we paired segmentation masks of images in $\mathcal{WL}_C$ with corresponding sunspot drawings, i.e. with entries in the DDB and SGDB. For each timestamped sunspot drawing, we extracted the chronologically closest WL image in $\mathcal{WL}_C$ to avoid most of annotation inconsistencies. Such errors may especially occur in the early stages of sunspot group development when classification may change within a few hours.

In our pairings, we set the maximum time difference between drawing annotation and image capture to 4 hours. This selection process produced 3150 pairs of (segmentation masks, sunspot drawings) for the 2002-2012 and 2016-2019 time periods and a total of around 6500 sunspot groups.

4.2.4 Classification dataset

Input to the classification task are the sunspot groups retrieved by the clustering algorithm. The corresponding number of clusters given to the classi-

fication task is of 5017 samples for the period $2002-12 \cup 2016-19$. However, sunspot groups that are close to one another deserve a specific treatment, in order to avoid their mixed patterns misleading the classification network during training and validation. Such cases are identified as groups whose bounding box image contains sunspots belonging to other group(s).

For each of these groups, before presenting the corresponding image to the classifier, we erase the sunspots of other close group(s) by setting their pixel intensity values to the mean value of the background pixels around their contour. In that way, only one group is available to the classifier at a time.

In the training dataset used in this supervised classification task, each of the 5017 samples gets an associated label that corresponds to the McIntosh class of the matched sunspot group in the SGDB.

4.3 Methods

4.3.1 Pipeline overview

The SunSCC pipeline is depicted in Figure 4.1. It takes as input images obtained from the USET White Light telescope of the Royal Observatory of Belgium and returns as output segmentation masks of individual sunspots, as well as the sunspot groups and their McIntosh classification, with a reliability score. It is divided in three independent blocks as follows.

The first block, dealing with the sunspot segmentation, is addressed by a CNN in the form of U-Net architecture [138] composed of an auto-encoder with skip connections. The encoding part of the auto-encoder has 5 levels and 4 downsampling steps, and reversely the decoding part has 5 levels and 4 upsampling steps. Each level is composed of two convolutional blocks. The numbers of kernels (or filters) estimated in each convolutional block at each level are, from highest to deepest level: 32, 64, 128, 256 and 512, respectively.

The second block receives as input the segmentation from the first block, and aggregates the detected sunspots into sunspot groups using an origi-

nal clustering method inspired by the mean-shift algorithm

Each identified sunspot group is provided along with an angular distance map to the third block. The latter constitutes a classification network designed to predict the McIntosh 3-components classification. It is composed of a ResNet34 [139] image encoder and three MLPs organized in a hierarchical way, each having 4 hidden layers and ReLU activation functions [164]. Each MLP is specialized in the classification of one component of the McIntosh system. The 4 hidden layers of each MLP contain 512, 256, 256 and 128 neurons, respectively.

An advantage of keeping the three blocks as independent as possible is that each can be used separately as a black-box, modified or enhanced with limited effect on the subsequent blocks. We detail below the methods corresponding to the segmentation, clustering, and classification of sunspot.

4.3.2 Segmentation

Our segmentation method first builds a set of unsupervised but error-prone segmentation masks, using conventional image processing techniques. Then these masks are used as pseudo-labels to train a CNN model. The following sections describe these two steps.

Training data preparation, using unsupervised image processing

The generation of unsupervised segmentation masks, also called ‘pseudo-label generation’ in the context of machine learning, is depicted in Figure 4.2. Each raw image in $\mathcal{WL}_{SA}$ is first processed to correct for limb darkening. WL images show a radial decrease in intensity from the center of the disk to the limb, which if not corrected, will hinder threshold-based segmentation methods. To correct this effect, a mask of the limb darkening is created by fitting a polynomial function of the mean intensity of the image in polar projection. The corrected image is then obtained by dividing the raw image by the limb darkening mask. Next, mathematical morphology operations, in the form of a black top-hat transform is applied, similarly to [149]. This allows to find dark areas surrounded by a bright environment corresponding to sunspots. Finally, a thresholding is applied to obtain a binary mask.

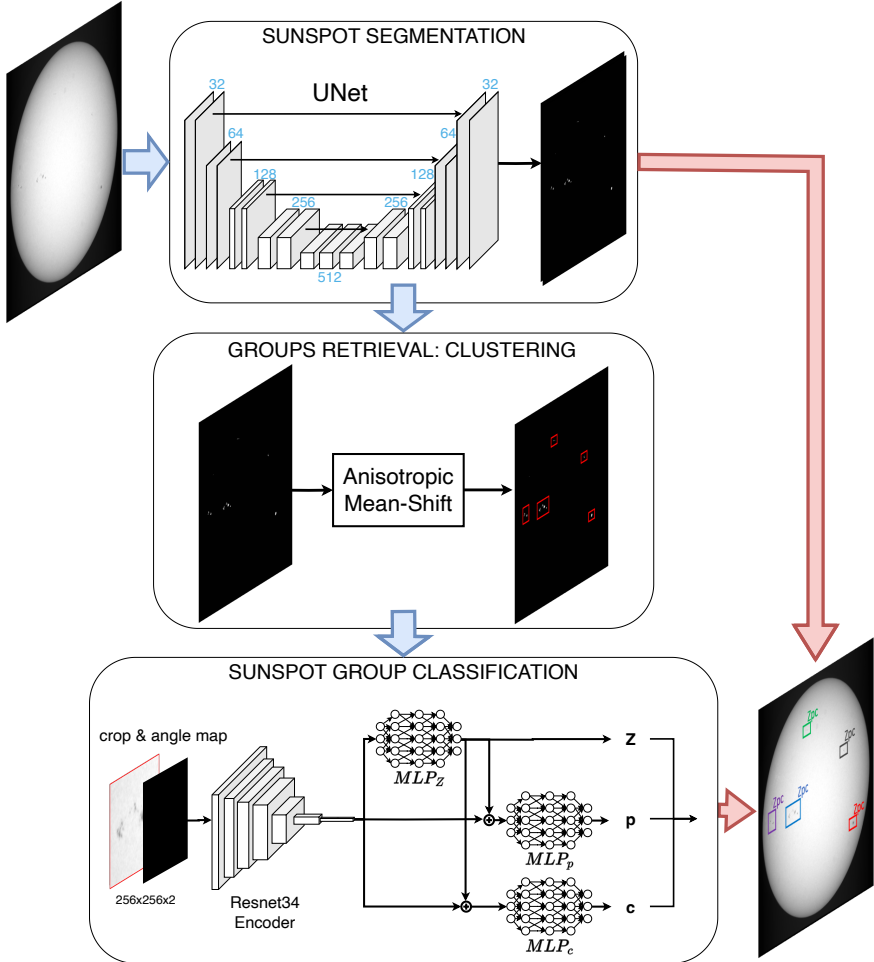


Fig. 4.1 SunSCC pipeline for sunspot segmentation, clustering, and classification. Full-disc images with a resolution of 2048×2048 are subdivided into a 4×4 grid, each containing segments of dimensions 512×512 . An U-Net segmentation network predicts masks for each segment that are reassembled into a 2048×2048 mask. The U-Net comprises 5 levels (with 32, 64, 128, 256 and 512 filters) and 4 downsampling steps in the encoding part, and reversely 5 levels (with 512, 256, 128, 64, and 32 filters) and 4 upsampling steps in the decoding part. Each level is composed of two convolutional blocks. The detected sunspots are aggregated into sunspot groups by a modified mean-shift algorithm. Each identified group is provided along with an angular distance map to a classification network composed of a Resnet34 image encoder and three MLPs with 4 hidden layers and ReLU activation function. Each MLP is specialized in the classification of one component in the McIntosh system. The 4 hidden layers of each MLP contain 512, 256, 256 and 128 neurons, respectively.

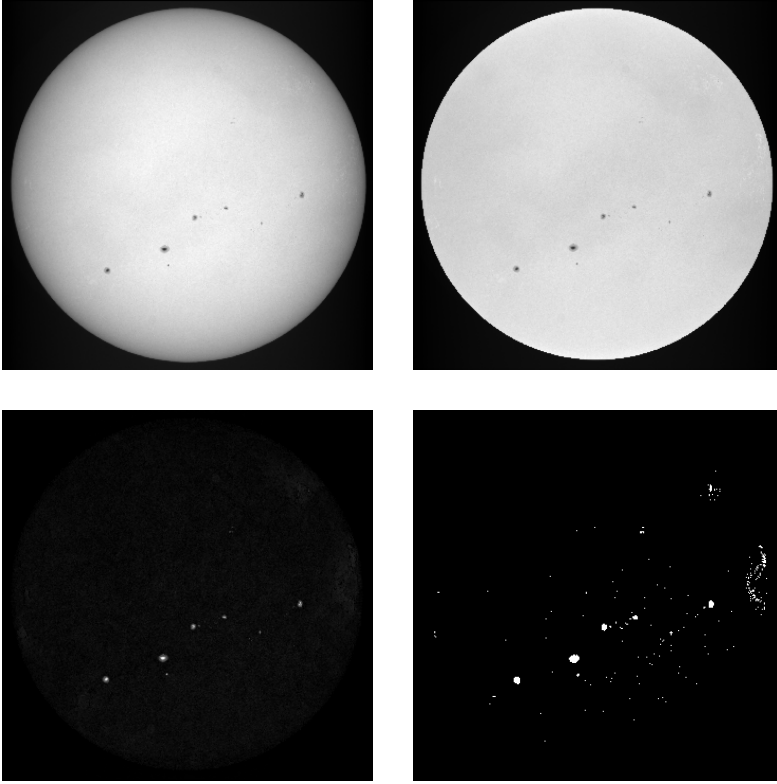


Fig. 4.2 Pseudolabel generation pipeline: (Top Left) Original image, (Top Right) Flattened center-to-limb variation, (Bottom Left) Black-TopHat transform, (Bottom Right) Binary mask obtained with thresholding.

Training the CNN with stochastic pseudo-labels

The U-Net architecture [139] is adopted for our CNN segmentation model since it has proved its accuracy in delineating high resolution segmentation masks in various fields such as medical imaging [138, 165] and natural images analysis [166, 167].

Since the smallest sunspots represent only a few pixels and thus may disappear when downsampling large images, we feed the CNN at full original resolution, but use patches of the full image (that we called ‘segment’) of size 512×512 pixels in our segmentation models. Note that unlike during the generation of the segmentation masks by thresholding methods in

previous section, the limb darkening effect is not corrected on the images given as input to the CNN. Instead, raw white light images are used. We count on the regularization properties of the CNN to be robust to limb darkening and the presence of clouds.

Finding a single threshold value during the mask generation process is a non-trivial task as the level of illumination may differ not only locally due to the presence of clouds and other artifact source such as atmospheric seeing, but it may also differs from image to image. Moreover, CNNs trained with a systematic labeling error are known to produce predictions with the same error patterns. To circumvent these issues, we consider multiple thresholds per WL images, each of them generating a specific mask or pseudo-label, and we train our model using these various pseudo-labels. This is expected to increase the error patterns variability, thereby preventing the CNN to learn them. Our experimental results demonstrate that this original strategy is effective in mitigating the bias introduced by pseudo-labels in the trained model prediction behaviour.

To describe this strategy formally, let I be an image and T a number of thresholds to consider. We denote each threshold by τ_j , with $j \in 1, \dots, T$. The set of masks (or pseudo-labels) associated to I is denoted

$$\mathcal{M}_I = \{ m_{\tau_1}, \dots, m_{\tau_T} \}, \quad (4.1)$$

where each mask m_{τ_j} , $j \in 1, \dots, T$, was created using the image processing method depicted in Figure 4.2, with a different threshold value τ_j .

When supervising the training of the CNN-segmentation, for each image I , one mask is randomly selected from $\mathcal{M}_I$ in Equation (4.1); where each mask has equal probability to be chosen. This mask is then used to compute the gradients of the binary cross-entropy loss adopted to update the model as is commonly done in segmentation of binary masks [168].

In our experiments, we used three thresholds ($T = 3$). The central threshold was chosen as the one with highest mask segmentation accuracy on the manually annotated set $\mathcal{WL}_{SM}$. The two other thresholds have been chosen to reject most false positives (at the cost of increasing missed detections) and accept most missed detections (at the cost of an increase of false positives), respectively. The segmentation network was trained us-

ing the sum of the cross-entropy loss and the log cosh Dice loss [168]. The AdamW [169] optimization algorithm was used with an initial learning rate set to 10^{-4} . The number of epochs was set to 40 with a batch size of 16. The model was trained using the images in $\mathcal{WL}_{SA}$ and their associated pseudo-labels, then validated on images in $\mathcal{WL}_{SM}$ and associated manual annotations.

4.3.3 Clustering

We propose an original method to find sunspot groups in a WL image, given the sunspot masks provided by our CNN segmentation model. Our approach adapts to the context of sunspot group retrieval the mean-shift clustering algorithm introduced in [51] and popularized in computer vision in [52]. The mean-shift clustering algorithm exploits local gradients of a density function to locate its maxima, or modes, given discrete data samples from that function. It simultaneously associates an attraction basin to each of the modes, thereby clustering the observed samples based on the basin they belong to. In the sunspot clustering context, the procedure to discover the modes, and to associate the observed sunspot to those modes, works as follows. Starting from the center of an observed sunspot, it iteratively shifts according to the gradient of the density function obtained by placing at the center of each observed sunspot a kernel shape weighted by the area of the sunspot. This shifting step is repeated until convergence to a local maximum, defining a mode. The convergence procedure is repeated for all the sunspots, which ends up in defining a discrete number of modes, to which sunspots are associated.

The rationale for choosing this particular method is as follows: it adopts a conventional linear kernel, but takes the shape of sunspot group into account, as well as the aspect of individual sunspots when defining the kernel spatial support in longitudinal and latitudinal directions. More precisely, anisotropic kernels are adopted to fit the distribution of sunspots in a group. This originates from prior knowledge that sunspots composing a group tend to be more spread along the longitudinal axis than along the latitudinal axis. Also, the longitudinal bandwidth of each kernel is defined as an increasing function of the area of the corresponding sunspot. The reason why area-dependent kernels are used comes from the observation that bigger sunspots have a larger probability to be a dipole of a large group.

Hence, a larger neighborhood should be considered when estimating the density function of the group associated to a bigger sunspot. On the other hand, the kernel bandwidth along the latitudinal axis is kept constant for all the sunspots. Mathematical details of this method are provided in Appendix B.

4.3.4 Classification

Images from the $\mathcal{WL}_C$ dataset are used to train classification models, with the sunspot group samples obtained by the clustering algorithm described above and tuned with appropriate kernel bandwidth values. The McIntosh classification scheme consists of 56 admissible classes, derived from combinations of the three components Z , p , and c (see Section 2.2.3). Some of these classes are poorly observed in our training dataset $\mathcal{WL}_C$, while others have similar morphology and are differentiated only by their sizes.

Therefore, we decided to use a reduced set of classes to tackle class imbalance and address CNN-based sunspot group classification. Additionally, we used an ensemble of classifiers to get a confidence measure for the classification prediction.

McIntosh classes aggregation

We use CNN in order to discriminate the overall aspect of the sunspot groups. In particular, classes that are differentiated only by the sunspot group size or distance between its two ends will be merged together, since a simple post-processing may separate them.

This reduction of the set of McIntosh classes operates at three levels:

1. D , E and F classes of the Z -component were merged into a single class denoted LG for 'LargeGroup'. This merger was motivated by the fact that these classes are distinguished only by the distance between leading and trailing sunspots in the group, not by any aspect of the sunspots themselves. The original classification can be recovered through post-processing by computing the longitudinal separation between group endpoints. Considering the full set of classes might actually lead to confusion of the CNN-classifier between classes D , E and F , as is precisely reported in [161] and [159].

- 2. regarding the *p*-component, the symmetric classes, i.e. classes *s* and *h*, were merged together in a *sym* class and in an analog manner the asymmetric classes, i.e. classes *a* and *k*, were joined in a new *asym* class. Distinction between these classes may easily be done by post-processing, based on a specific length measurement, as explained above.
- 3. the number of classes of the *c*-component was reduced to three by merging classes *i* and *c* used for groups where the space between dipoles is populated with smaller sunspots. We named this new class *frag*, standing for 'fragmented'. The distinction between classes *i* and *c* is qualitative in nature and hence contains a part of subjectivity. A human operator will have to make the final decision to discriminate them.

The reductions on the usual McIntosh classification scheme of are represented in Figure 4.3.

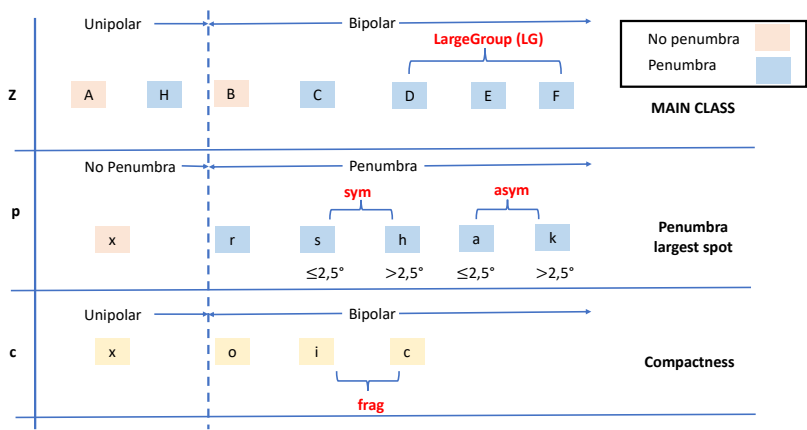


Fig. 4.3 Reduced McIntosh classification scheme used in SunSCC. Red fonts indicate how some of the initial McIntosh classes have been merged in a single class in our study (see the text for details).

CNN-based classification

As shown in Figure 4.1, our classifier relies on a ResNet34 convolutional backbone to encode the visual information associated to the sunspot group to be classified. The network is fed with standard-sized 256×256 images for all sunspot groups. Before cropping those images, the corresponding full-disk images are rotated to compensate for the sun P-angle, to make the sun rotation axis vertical. The cropped images are composed of two channels: the first channel contains the square segment of the WL image centered at center location of the sunspot group formed during clustering, and the second channel contains the corresponding segment of an angular distance map aligned with the WL image. This second channel is added to the WL input to provide information about the location of the group on the visible hemisphere to the classification network so that knowledge about distortion of sunspots shape approaching the limb can be learned.

The embedding representation of the images provided to the ResNet34 backbone is given as input to three MLPs named MLP_Z , MLP_p and MLP_c , respectively specialized in the classification of each of the Z , p and c component of the McIntosh system. The MLPs are organized hierarchically such that the output of MLP_Z is concatenated to the encoded embeddings given as the input of MLP_p and MLP_c . The objective of this architecture is to mimic the dependency of the p - and c -component classifications on Z .

Each MLP is trained separately. First the ResNet image encoder is trained along with MLP_Z using a sampling that is balanced in the representation of all the Z -classes. Then those two sub-parts (ResNet image encoder and MLP_Z) are frozen and MLP_p is trained using balanced samples of the p -classes. In the final training step, MLP_p is also frozen and MLP_c is trained with a balanced number of sample of each class of the c -component.

In order to balance the representation of the classes for each sub-training, we associate each item i in the training set to three sampling weights, one per McIntosh component. Those weights are inversely proportional to the frequency of the sample class in the training set. These weights control the sampling rate of samples during training.

The classification network was trained using the cross-entropy loss and

Adam optimization algorithm [170]. The initial learning rate was set to 5.10^{-5} . The batch size was set 16 and the number of epochs was set to 100 for each of the training phases i.e. training of each MLP.

From hard to soft predictions using ensemble of classifiers

In this work, we applied the precepts of *ensemble learning* [171, 172] to soften the prediction of a sample by a single classification network into a confidence measure of belonging to a particular class.

The founding concept of ensemble learning is that the biases of a single classifier are likely to be compensated by biases of other classifiers when combining their predictions. The overall prediction accuracy or ensemble prediction accuracy is expected to be better than the performance of a single classifier (see [171] for a review). Analyzing the predictions associated to an ensemble of models has also been successfully investigated to quantify predictive uncertainty in neural networks, i.e. to associate a level of confidence to the prediction [173, 174]. Cases for which the prediction disparity across models is large are considered to correspond to unreliable predictions. This has been studied and validated in the context of image segmentation [175, 171]. In our experiments, multiple instances of the classification network shown in Figure 4.1 are initialized with different seeds, and are trained independently, to produce multiple classifiers. Hence for each sunspot group, we obtain a pool of predictions. These predictions are then combined into one single output via a majority vote.

The results provided in Section 4.4.3 show that the class prediction inconsistency across the ensemble of classifiers is a valuable clue to identify the cases for which the majority vote prediction is likely wrong.

4.4 Results

This section gathers the discussion on the results obtained by the three tasks of the SunSCC pipeline: segmentation (Section 4.4.1), clustering (Section 4.4.2), and classification (Section 4.4.3) of sunspots.

4.4.1 Sunspot Segmentation

Evaluation metrics

We evaluated segmentation models as detection models, of which they are a pixel-wise extension. In the performance evaluation of such models, only true positives, false positives and false negatives are considered. On the other hand, true negatives object candidates are left aside as a potentially infinite number of them may lie in each image, making them irrelevant. The performance can thus be measured with the usual *precision* and *recall* that can be combined in a single metric named F1-score. Formally, using the manually annotated dataset $\mathcal{WL}_{SM}$ as ground truth, we first evaluate the quality of the segmentation with the F1-score, defined as:

$$F1 = \frac{2TP}{2TP + FN + FP} \quad (4.2)$$

where TP (True Positive) is the number of correct detections, FN (False negative) is the number of missed detections and FP (False Positive) is the number of undesired detections. The F1-score thus decreases when there are undetected sunspots (FN) or erroneous sunspot detection (FP). Second, we consider the number of multiple sunspots covered by a single detection (called ‘merged’ detection) as another quality metric. This is because the precise count of individual sunspots is an important element for the determination of the International Sunspot Number [176].

Sunspots have a wide range of possible sizes in our images, from a few pixels for the smallest spots to several thousand pixels for the largest spots. Hence, missing a few pixels in the segmentation of a small sunspot has a different meaning than missing the same number of pixels in a large sunspot. Therefore the segmentation of sunspots is analyzed as a function of the sunspot size, by grouping sunspots in five bins with increasing number of pixels. The bin boundaries were chosen so that the population of each bin is sufficient to depict a statistically relevant behaviour in order to reflect the variation of detection performance observed on sunspots of different sizes. The boundaries of these bins are given in Table 4.2.

Segmentation results

Figure 4.4 displays, as a function of the threshold value, the performance of the following three methods:

bin	0	1	2	3	4
#pixels	< 50	50–100	100–400	400–800	> 800

Table 4.2 Sunspot size bin boundaries.

- 1. unsupervised threshold-based segmentation (described in Figure 4.2),
- 2. CNN-based segmentation that is trained with one pseudo-label corresponding to one threshold value,
- 3. CNN-based segmentation trained with three pseudo-labels per WL images, each pseudo-label being obtained with a different threshold value, and being selected alternatively with equal chance during training.

The first two approaches are represented by solid coloured curves, while the third is represented by a constant dashed line as it does not depend on a single pre-defined threshold.

The first noticeable result in is that the CNN-based segmentation regularizes the flaws of small sunspot detection observed when using thresholding. This is observed as the F1 scores of CNN-based segmenters for Bin 0 and Bin 1 consistently outperform those of threshold-based segmenters. Regarding sunspot Bin 0, the highest performance reached using a thresholding method was 55%. The performance increase using CNNs goes from 7% for lower thresholds up to 24% for higher ones, reaching a new maximum of 71%. The same observation could be made on Bin 1 with a smaller increase as the thresholding method was able to achieve a F1-score of 92% F1. Bins of larger sunspots keep similar performance and reach 100% of detections.

The trade-off between the number of non-detected sunspots and the number of merged sunspots - i.e. fusions - for each threshold-based segmenter and CNN-based segmenter can be found in Figure 4.5. One can observe that CNN-based segmentation regularises sunspot detections since the number of misdetections is significantly reduced when compared with thresholding. On the other hand, CNN-based segmentation tends to slightly increase the number of fused sunspots (but not more than a few percent).

Finally, the segmentation using three alternating pseudo-labels per WL

image during training achieved performance comparable to that of the best CNN-segmenter using a single pseudo-label, both in terms of F1-score and in terms of fusion versus non-detected sunspots trade-off, as visible on Figures 4.4 and 4.5. It also achieves a smaller number of fusion than the thresholding methods with the best F1 scores, as shown on Figure 4.5. This highlights the benefit of using several pseudo-labels as a solution to train CNN-based sunspot segmentation instead of searching a unique best threshold value. This observation is especially relevant and helpful when few annotated samples are available (in the validation set) to select the unique threshold that would maximize the performance.

4.4.2 Clustering

To retrieve sunspot groups, we use the masks produced with the best performing CNN classifier in our segmentation experiments. Such masks were obtained for each image in $\mathcal{WL}_C$ (see Section 4.2).

The performance of the proposed clustering method is done by measuring the concordance between clusters found by the automatic method, and the SGDB entries. To quantify this concordance, bounding boxes computed from SGDB entries are compared with bounding boxes of groups found by clustering. We refer to SGDB bounding boxes as *target boxes* and bounding boxes of groups formed by clustering as *meanshift boxes*. The dimension of each target box is computed as a function of the McIntosh class and center position of the sunspot group, and solar radius from the corresponding visual WL observation, see Appendix D for details.

In each image of the set $\mathcal{WL}_C$, each target box is paired with the closest meanshift box. The set of matchings pairs is denoted M . If two target boxes have the same meanshift box as pair candidate, only the target box with smallest distance is paired, while the other is considered *unmatched*. If both target boxes overlap the meanshift box (and have a zero-distance), the one with largest intersection over union (IoU) is paired. We denote U_T (resp. U_M) the set of unmatched target boxes (resp. the set of unmatched meanshift boxes) along the whole image set $\mathcal{WL}_C$.

The evaluation of the performance of the proposed clustering method should directly depend on the size of U_T and U_M , denoted $\#U_T$ and $\#U_M$,

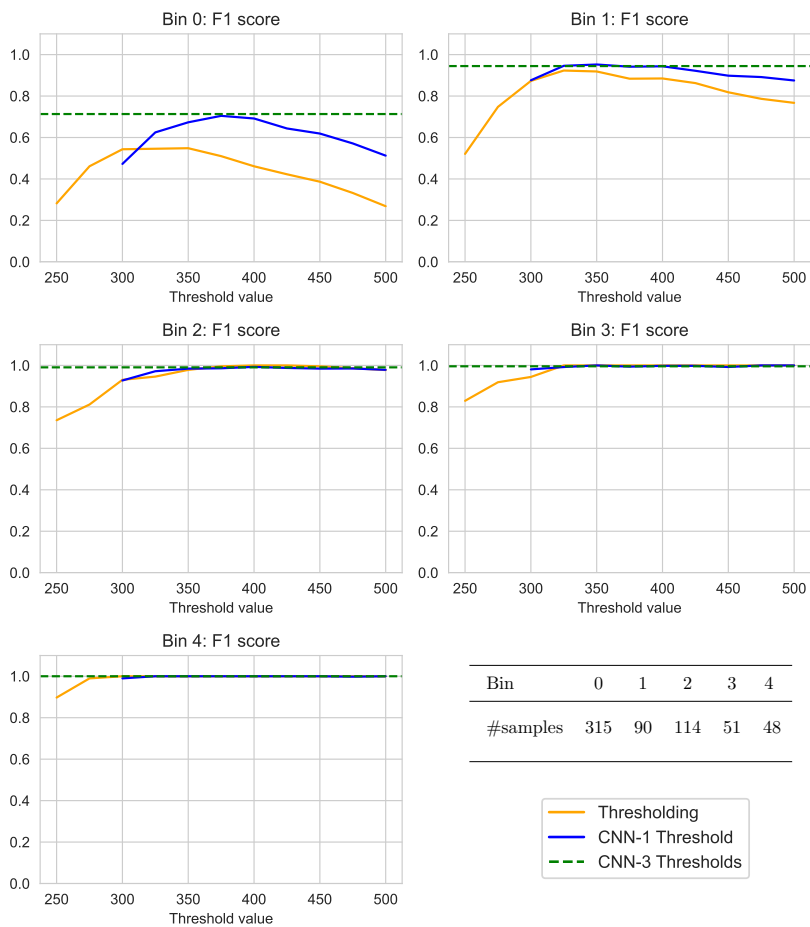


Fig. 4.4 F1-score of sunspot detection as a function of threshold value for each sunspot size bin. The table depicts the population size of each bin. The methods represented are: unsupervised segmentation using for thresholding (orange curve), CNN trained with unique pseudo-label per WL image (blue curve), CNN trained with 3 pseudo-labels per WL image (green dashed line). The last CNN was trained by selecting alternatively with equal chance the pseudo-labels generated using the following threshold values: $\tau_1 = 325$ $\tau_2 = 375$ $\tau_3 = 425$.

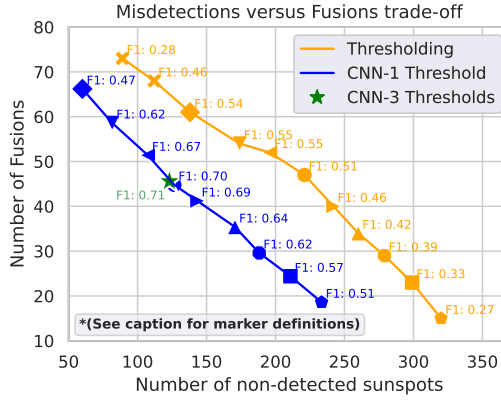


Fig. 4.5 Trade-off between sunspots remaining undetected and sunspots being merged in a single entity. The thresholding method (orange) labels as sunspot the pixels lying above a threshold value in the Black-TopHat transformed image. CNN-1 (blue) corresponds to the models whose training has been supervised with pseudo-labels obtained with a single and fixed threshold. CNN-3 is trained with 3 distinct pseudo-label images per WL image (green). Markers indicate the threshold used (diamond: $\tau = 300$, down triangle: $\tau = 325$, left triangle: $\tau = 350$, circle: $\tau = 375$, right triangle: $\tau = 400$, up triangle: $\tau = 425$, hexagon: $\tau = 450$, square: $\tau = 475$, pentagon: $\tau = 500$, star: $\tau_1 = 325$ $\tau_2 = 375$ $\tau_3 = 425$). The F1 scores indicate performance on sunspots in the first bin, which corresponds to the smallest and thus most challenging sunspots to detect.

respectively. Indeed, the unmatched target groups in U_T , and the unmatched mean-shift groups in U_M both correspond to clustering errors, except if they are isolated, i.e. far enough from all groups of the other [target/mean-shift] category. Indeed, when the sunspot segmentation is accurate, U_M corresponds to the cases for which a target group has been erroneously split into multiple meanshift groups (leading to unmatched meanshift groups), and U_T to the case for which a meanshift group has aggregated two target groups (leading to an unmatched target). When the segmentation is not correct, false positives and missed detections lead to isolated mean-shift and isolated target groups, respectively. We let $I_T \subset U_T$ and $I_M \subset U_M$ denote the unmatched bounding boxes associated to those missed or false positive sunspot detections. Since they reflect detection errors, those I_T and I_M bounding boxes should preferably be ignored when assessing the clustering performance in order for the evaluation of the clus-

tering method to be as independent as possible from the quality of the segmentation. Hence, $\#I_T$ and $\#I_M$ are removed from $\#U_T$ and $\#U_M$, respectively, when computing the clustering accuracy metric.

Formally, the accuracy of the clustering algorithm is computed using the following formula:

$$\text{Accuracy} = \frac{\#M}{\#M + (\#U_T - \#I_T) + (\#U_M - \#I_M)} \quad (4.3)$$

Appendix C presents a grid search to find appropriate values for the latitude bandwidth as well as maximum longitude bandwidth to define the kernels used for clustering. This reveals that our method achieves 80% accuracy when appropriate meanshift parameters are adopted. The values of longitudinal and latitudinal kernel bandwidths achieving this performance are .4 radians and .1 radians, respectively, see Figure C.1 in C. These values have been adopted when creating the dataset on which the classification network was trained.

4.4.3 Classification

The details about the distribution of samples among the classes of the Z, p and c components of the McIntosh system are shown in Tables 4.3, 4.4 and 4.5, respectively. In our experiments, we used 85% of the available samples for training and 15% for testing.

One can observe the presence of class imbalance in each of those tables as the representation of rarest classes can vary from half to the third of the representation of the most populated ones. Also, the respective imbalances of the classes in Z, p and c components make it difficult to train the whole network at once as balancing the training on one component does not guarantee balanced training for classes of other components. The context of this work thus requires a training supervision strategy of the classification models that takes this additional complexity into account to ensure balanced representation of each class for each McIntosh component. As explained in Section 4.3.4, we use weights associated to each item to control the sampling rate during training.

	A	B	C	LG	H
#samples	578	691	1026	1377	1345

Table 4.3 Class distribution of Z component in classification dataset.

	x	r	sym	asym
#samples	1269	533	1845	1370

Table 4.4 Class distribution of p component in classification dataset.

	x	o	frag
#samples	1923	2014	1080

Table 4.5 Class distribution of c component in classification dataset.

Mean classification accuracy as a function of the category

In our experiments, we trained 20 classification CNNs to produce ensemble predictions. Each test sample was provided to every CNN such that a majority vote could be performed on the predictions of each component Z , p and c . The performance of this ensemble learning approach was measured with the weighted F1-score on each component, and this for two sets of test samples: first, the set of sunspot group samples that lie far away from any other group on the cut-out WL image given to the CNN, and second the set of samples close to another group, either because their Bbox overlaps that of other group(s) and/or because other group(s) appear on the WL image given to the CNN. On the first set, the proposed network achieved 69% F1-score on the classification of the Z component, 62% on p and 72% on c . These results solely based on ground-based images are comparable with the solution proposed by [161] that uses magnetograms along with satellite based WL image. Their solution respectively reached 67% F1 score for the Z component, 55% for p component and 73% for c component.

On the second set of harder samples, the proposed network achieved 57% F1-score on the classification of Z component, 54% on p and 63% on c . Lower performances were expected as it is more likely that these samples have suffered errors during clustering and therefore do not present the characteristics of the McIntosh class they are labeled with. One can

	<i>Z</i>	<i>p</i>	<i>c</i>
Isolated groups	69%	62%	72%
Packed groups	57%	54%	63%

Table 4.6 F1 score of the classification with majority vote for *Z*, *p* and *c* components on test sunspot groups that are far from other group(s) (top) or close to, i.e. with overlapping bounding boxes, other group(s) (bottom).

expect that further enhancement of the clustering method allows to reach classification performances on the second set of samples similar to the first, provided that the unwanted sunspots in the input image are hidden. The results presented from now on depict the performance on the set of sunspot groups that are far from any other group.

The breakdown of our results are shown in Table 4.6, Figure 4.6 as well as Tables F.1, F.2 and F.3 from Appendix F. Beside the F1-score, these tables also present the Precision and Recall metrics, which are defined in Appendix E. Figure 4.6 shows confusion matrices presenting the performances of the classification of the *Z*, *p* and *c* components. Formally, each entry $M(i, j)$ of the confusion matrix M denotes the rate at which the ground-truth class *i* is assigned to class *j*. On these matrices, one can observe that the majority votes lead to a significant amount of correct predictions. Indeed, the diagonal entries, which represent the per-class recall, have high values and the off-diagonal entries are low or show consistency with possible confusion between ‘adjacent’ classes in the lifecycle of sunspot groups [24]. For instance, confusion between class *B* and class *C* and between *C* and *LG* for the *Z* component is expected due to the intermediate status of class *C* between class *B* and the classes grouped under *LG*. The same observation can be done regarding the confusion between class *r* and *x* for the *p* component. Such confusions could be done by a human annotator as well, depending on the meticulousness of their observation of both leading and trailing spots, or their interpretation of a ‘rudimentary’ penumbra. Therefore, due to the definition of the McIntosh classes themselves, any classification method will present minor confusion cases, up to some extent.

The next three sections presents three take-away messages that can be retrieved from our experiment: performance of CNN-classifier decreases

near the limb, even if some information about limb is provided to the algorithm, the CNN-classifier is able to find annotation errors in the SGDB, and finally the ensemble classifier allows to identify ambiguous cases in sunspot classification.

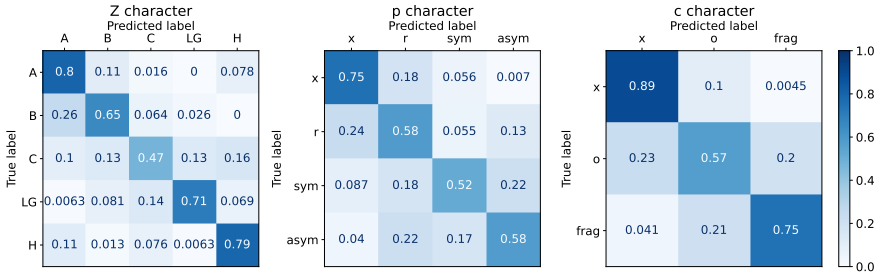


Fig. 4.6 Test set confusion matrices of classification with majority vote for the classification of the Z (left), p (center), c (right) components. In each matrix, the rows represent the true label of test samples while columns represent the label predicted by majority vote. The matrices are row-normalized such that per-class recall values are shown on the diagonals. See Appendix F for details about recall metric computation.

Classification accuracy as a function of the angular distance to the solar disc center

Figure 4.7 (top-left) depicts the classification accuracy of each Z, p and c component as a function of the angular distance to the Sun center. This distance between a sunspot group and the center of the Sun is computed as $\arcsin(r/R)$ where r is the distance between the group center and the solar disc center and R is the solar radius. We observe that the mean accuracy changes with the angular distance. This variation is not due to the stochastic nature of the training procedure [48], as shown by the small error bars presented on Figure 4.7 (top-left) to depict the standard deviation resulting from multiple stochastic trainings of the neural network. Instead, a careful analysis of the distribution of the classes in the test and training sets of each component has revealed that the drops of mean classification accuracies observed for some angular distances are primarily due to a lack of training samples for some classes that are significantly represented in the test set. In other words, they are caused by a mismatch between the distribution of classes in the test and the training sets when the number of

training sample is small. Hence, building the training set so that all classes are reasonably represented across all angular distances is expected to mitigate the erratic fluctuations of accuracy observed in the plot.

This interpretation is supported by Figure 4.7 (bottom), which presents, for the 70 – 80 degrees bin, the distributions of the labels associated to the p -component in the training and test sets. We observe that the 'sym' class is dominant in the test set, but only poorly represented in the training set (less than 20 samples). This explains why the corresponding classification accuracy sharply drops in the graph presented in the top of the figure. It is worth noting that this mismatch between the distribution of classes in the test and the training sets is especially problematic when the number of training sample is small, since it ends up in allocating a small number of training samples to a class that is largely represented in the test set, thereby hampering the overall test accuracy.

Note also that the variable number of samples available around a given angular distance is directly linked to the distribution of sunspots on the Sun over time -e.g. the butterfly diagram [21]-, combined with the position angle between the geocentric north pole and the solar rotational north pole (P). As P varies over the course of a year, the solar equator moves around the solar disk center creating a non-uniform distribution. This effect is depicted in Figure 4.7 (top-right), which presents the training sample distribution as a function of the angular distance. It peaks between 20 and 30 degrees, and reveals the small number of samples available in the bins with small or large angular distance.

A few examples of failure cases near the limb are shown in Figure 4.8. The example on the left (August 9, 2005, 11:22) shows a group on the limb that is recorded as H-asym-x in the SGDB and that is classified as C-r-o by the network. On the drawing, one can see that the observer probably classified it as A-x-x with 2 spots although a penumbra was drawn. The operator encoding the group in the catalog saw the penumbra and classified it as H. However, considering that there is a non-negligible distance between both spots and that this distance is further underestimated due to projection effects at the limb, it could also be considered as C-class. The fault in such failure cases is not to be blamed on the classification network only, nor on the operator. This illustrates complications that may occur due to the position of the group close to the limb.

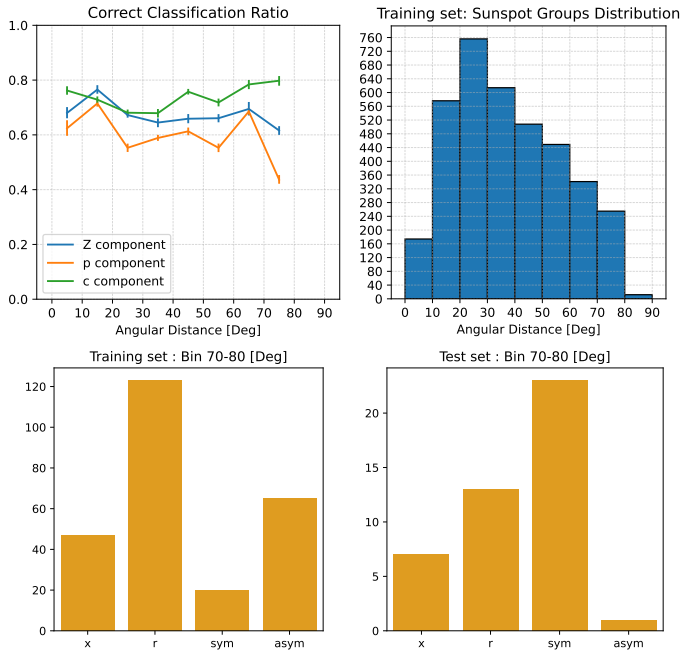


Fig. 4.7 (Top Left) Classification accuracy along angular distance with error bars shown as vertical lines. (Top Right) Distribution of test set samples along angular distance. 0 is center of the solar disk, 90 is the limb. (Bottom Left) Distribution of the train samples of bin 70-80 degrees among classes of the p component.(Bottom Right) Distribution of the test samples of bin 70-80 degrees among classes of the p component.

Finding annotation errors

Discrepancies may exist between the classification recorded in the SGDB and what can be seen on the temporally closest WL image. There are multiple explanations for this. (1) What an observer sees when annotating a drawing and what the digital camera captures are different, e.g. due to turbulence. Most of the time, discrepancies would be in the other direction, i.e. the observer has a long integration time when observing spots and groups while the WL image has routinely limited integration capabilities. (2) A human mistake may occur when saving new entries in the

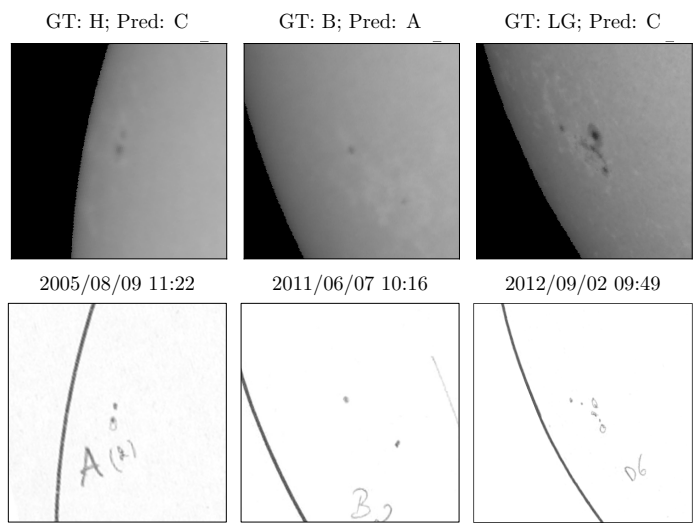


Fig. 4.8 Examples of failure cases close to the limb. For each sample, the capture date and time as well as the expected ground truth classification (GT) and the predicted classification (Pred) are given, along with the WL image presented to the classifiers (top row) and corresponding drawing with annotations about sunspot count and class written during observation (bottom row).

SGDB. An example of such discrepancies that occurred on 6/09/2004 is shown in Figure 4.9. On this figure one can see that the observer wrote “C” but did not draw the penumbra correctly. Therefore, the operator wrote “B” in the USET catalog as the drawing is historically the unique source to encode groups. However, the sunspot group was actually C-class, as clearly visible in the corresponding WL image. Still, the classification network was able to predict the classification correctly, but it was counted as an error in the performance evaluation due to this erroneous entry in the SGDB.

Identifying ambiguous cases

Using an ensemble learning approach allows to put a confidence measure on the overall classification of a sample. A noticeable result of our experiments is that correct ensemble predictions have low discrepancies between

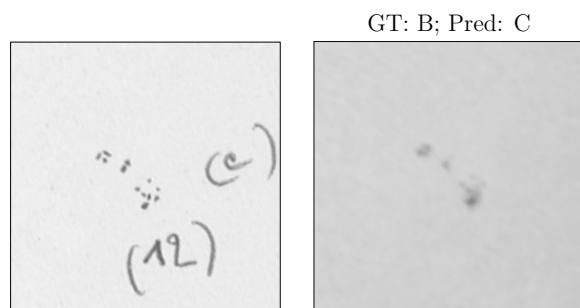


Fig. 4.9 Example of annotation error in USET catalog revealed by SunSCC. Side by side comparison of drawing annotation (Left) and corresponding WL image given as input for CNN-based classification (Right) for sunspot group observed on September 6, 2004. The observer annotated this group as C-class without penumbra delineation. When encoding in the SGDB, this group was classified as B-class for this reason. The classification network was able to retrieve the correct class.

the votes while incorrect ensemble predictions have their votes more distributed among the classes. When the majority of predictions voted for the correct class, we observe that, for each McIntosh component, more than 93% of the predictions participated to the correct majority vote, as can be seen in top row matrices of Figure 4.10. On the other hand, only a relatively low percentage of predictions that participated to an incorrect majority vote did vote for the correct class, as can be seen by the diagonal values in the confusion matrices of bottom row matrices of Figure 4.10. This behavior allows for the identification of difficult cases that may be submitted to an expert for correct classification when classifying new groups and can also help to raise a flag to signal the presence of potential annotation errors when trying to classify sunspot groups from the past. Examples of erroneous low-confidence majority vote predictions are shown in Figure 4.11.

4.5 Discussion and Perspectives

This chapter has introduced SunSCC, a comprehensive pipeline that employs a convolutional neural network for sunspot segmentation in ground-based white light images, implements an original clustering method for

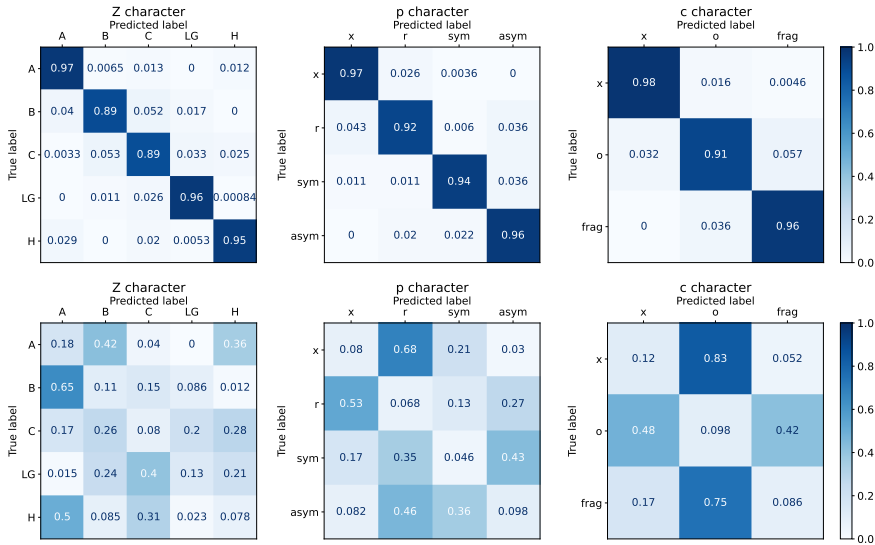


Fig. 4.10 Per-component test set confusion matrices of predictions that participated to a correct majority vote (top row matrices) or incorrect majority vote (bottom row matrices). In each matrix, a cell shows the rate of predictions that voted for its column class while the row class was expected. The more intense the blue, the higher the rate. In the top row matrices, off-diagonal values represent erroneous votes that might occur, even in case of a correct majority vote. In the bottom row matrices, some of the predictions voted for the correct class, hence the non-zero values along the matrix diagonal.

sunspot aggregation, and utilizes a second convolutional neural network for 3-components McIntosh classification. The pipeline demonstrates several notable strengths and opportunities for future development worth discussing.

In the first stage of SunSCC, the segmentation CNN effectively handles atmospheric artifacts typically present in ground-based images while producing more accurate masks compared to conventional unsupervised image processing approaches. The training methodology alternates between different automatically-generated pseudo-labels for each training sample during the learning process. This strategy alleviates the need for man-

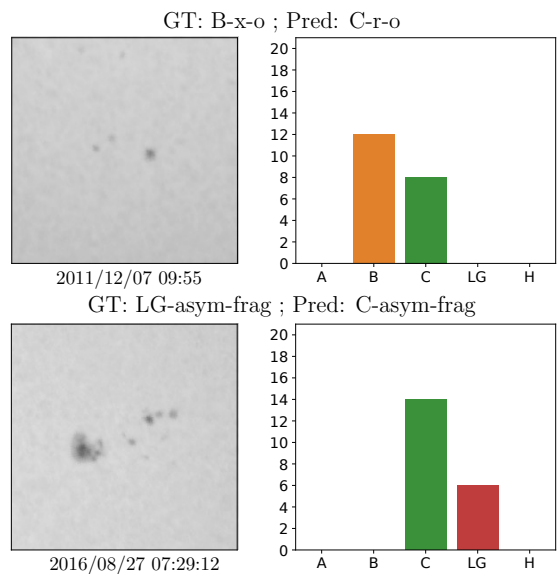


Fig. 4.11 Examples of ambiguous cases. For each sample, the expected ground truth (GT) classification and the predicted class (Pred) are shown as well as the distribution of votes taken into account in the majority prediction of the ensemble of classifiers.

ual or carefully curated training masks, as the network demonstrates an ability to retain the relevant characteristics from the diverse pseudo-labels. Despite using only automatically generated labels, this approach achieves performance comparable to training with single, carefully selected masks.

Building upon these reliable segmentation results, the clustering component of the pipeline identifies the sunspot groups. Our aggregation approach proved to successfully identify the vast majority of sunspot groups documented by experts at ROB. This effective grouping then enables the classification neural networks to achieve performance comparable to existing solutions that rely on both satellite-based images and magnetograms, despite using only ground-based imagery.

Currently deployed at the Royal Observatory of Belgium, SunSCC generates daily sunspot masks, grouping, and classification. While its classi-

fication capabilities do not yet match human expertise sufficiently to fully automate the process, it serves as a valuable auxiliary tool for operators, providing preliminary classifications and highlighting cases that may require detailed examination. Looking beyond its current implementation, the design of SunSCC allows for deployment at any ground-based station equipped with a white light telescope, with Transfer Learning techniques enabling adaptation of the CNN-based segmentation and classification to various ground-based instruments.

While the experimental results already demonstrate the effectiveness of SunSCC, its performance can be further improved through several key enhancements. The cascaded architecture of the pipeline ensures that improvements to individual components will benefit all downstream stages.

An immediate direction for improvement lies in differentiating between umbrae and penumbrae during sunspot segmentation. This refinement would enable the computation of valuable statistics that would advance research and enhance catalogs like the USET Sunspot Group Database. Incorporating such distinction into the classification model could improve classification accuracy, particularly for the Z and p components. This straightforward enhancement comes with specific challenges: smaller sunspots might be incorrectly identified as ‘penumbra only,’ which is physically impossible, and the detection of residual penumbrae may pose difficulties that even human observers struggle with. Nevertheless, given our pipeline structure, implementing this distinction in the segmentation stage would naturally enhance the classification task.

Another promising direction involves integrating temporal information across all pipeline stages. The segmentation could exploit temporal consistency by using masks from previous observations, while the clustering could maintain group identity through tracking information. The classification model would benefit significantly from considering the chronological record of a group’s observations and McIntosh classifications, as evolutionary patterns provide essential context. This ambitious extension would require linking sunspot groups across successive observations to establish their evolution sequences. While this demands extensive annotation work, the potential classification accuracy improvements could justify this effort. The approach must address practical challenges: missing observations, varying image quality, and differing observation intervals. Be-

yond enhancing the pipeline's overall performance, temporal integration would enable valuable sunspot tracking functionality in the USET Sunspot Group database.

The addition of temporal information motivates consideration of transformer architectures, which specialize in processing observation sequences and can encode both positional and temporal information. Though these architectures demonstrate remarkable performance across domains, they demand substantial training data. As the field advances, these data requirements may become less prohibitive for sunspot segmentation and classification.

The proposed improvements complement each other: penumbra/umbra distinction enhances temporal tracking through detailed features, while transformer architectures can leverage both refined segmentation and temporal data. However, these enhancements increase computational demands, particularly in training time and inference complexity. This computational cost would need to be balanced against operational requirements for daily solar monitoring.

Finally, while the proposed directions can improve classification accuracy, it's important to acknowledge the inherent limitations of the McIntosh classification system. The boundaries between certain classes are fundamentally ambiguous and subject to interpretation, making perfect classification challenging even for human experts. This ambiguity stems from the continuous nature of solar phenomena being mapped to discrete categories, creating inevitable edge cases where classification becomes subjective. It therefore creates a theoretical upper limit to performance for any automated system, including our proposed improvements.

5

Cross-modal Translation of Active Regions Observations

Now that we have established deep learning as a powerful assistive tool for experts observing active regions, we extend our approach to address another critical challenge in solar physics research: the translation between different observational wavebands. While the previous focused on extracting and categorizing features within a single observational domain, this chapter explores how deep learning can bridge gaps between different observational modalities. This transition from feature identification to cross-modal synthesis represents a natural progression in our efforts to overcome the data scarcity problem that regularly hampers progress in solar physics. This scarcity is defined as poor coverage over time, limited vantage points, or gaps in observed modalities. Many of these gaps arise from historical reasons—some modalities simply were not available in the past—raising the question of whether it would be possible to reconstruct past data from more recent datasets. By developing techniques that can translate between different wavebands, we aim to provide solutions to this persistent challenge, enabling the reconstruction of observations that were never directly captured.

For example, the USET station located at ROB has been acquiring full disk solar images in WL since 2002, but observations in the Ca II K line

started in July 2012 only [34]. The CaIIK line observations provide information from the low chromosphere about plages, which are a manifestation of solar surface magnetic fields. They allow for the reconstruction of this surface magnetic field [177], and are used in irradiance studies [178] and in comparing solar magnetic activity to other sun-like-stars [179]. WL imagery on the other hand observes the photosphere and shows sunspot groups and faculae, the two main structures at the origin of the chromospheric plages. Faculae are visible in particular near the limb, and together with the sunspot groups they physically correspond to the plages seen in CaII-K. It is thus of interest to explore how well we can recover Ca II K information from WL images, when Ca II K observations are missing.

Another issue caused by the lack of data is the ‘old magnetogram’ problem faced in space weather forecast. Indeed, for most magnetohydrodynamics (MHD) models, magnetograms are the primary input [180, 181, 182]. Currently however, most magnetograms are observing the frontside of the Sun only, either directly from the ground with the Global Oscillation Network Group (GONG,[33]) or from satellites located along the Earth-Sun line such as the Solar and Heliospheric Observatory (SOHO)/Michelson Doppler Imager (MDI, [183]) and the Solar Dynamics Observatory (SDO)/Helioseismic and Magnetic Imager (HMI,[160]). As a consequence, some of the information inputted to these models through magnetogram synoptic maps were acquired up to two weeks earlier. Since the Sun is constantly evolving, such delay causes inaccuracies in the MHD model’s output. Extreme Ultraviolet (EUV) telescopes on the contrary are less complex instruments than magnetograms, and as such are available on spacecraft which observe the Sun from other vantage points. This is the case of the Solar Terrestrial Relations Observatory (STEREO)/Extreme UltraViolet Imager (EUVI, [184]) and the Solar Orbiter (SO)/Extreme Ultraviolet Imager (EUI, [185]). Since both magnetograms and EUV imagery capture different manifestations of the same magnetic field, investigating how I2I translation techniques could derive magnetogram data from EUV observations becomes a relevant research topic.

In this chapter, we address two cross-modal image-to-image translation tasks of interest in solar physics: (1) the generation of Ca II K images from WL images, and (2) the generation of magnetograms from EUV images. The chapter is organized as follows. Section 5.1 presents the motivations of this work and reviews existing literature on image-to-image translation

in solar physics. Section 5.2 describes the datasets used for the two translation tasks we considered. Section 5.3 presents the two main strategies envisioned for image-to-image translation: adversarial strategies and non-adversarial strategies using modulation of intermediate feature maps. Section 5.4 details the experimental setup and Section 5.5 presents our results. Finally, Section 5.6 summarises our findings and discusses future work.

5.1 Motivations and related work

To mitigate the data scarcity problem, generative deep learning methods have recently become popular in solar physics, but their scientific validity has been so far largely overlooked.

As introduced in Section 3.3.2, Generative Adversarial Networks and their conditional variants have demonstrated strong capabilities for image generation and translation tasks. The Pix2Pix algorithm [104], which implements an image-to-image cGAN using U-Net architecture [69] for its generator and a classification CNN for its discriminator, has become a popular solution for image-to-image translation across numerous domains, including solar physics applications.

Next to adversarial training strategies, the diffusion models we introduced in Section 3.3.2 have emerged as another powerful approach for image generation. These models generate samples by simulating a Markov chain that gradually denoises pure noise into realistic samples [99]. DMs have demonstrated success in numerous image synthesis applications with very realistic quality, and their capabilities can be extended through conditional generation. In this approach, the backward process is guided by integrating encoded features from a reference image as input to the denoising U-Net [186, 187, 188]. These conditional DM methods work well for applications where input and output share high similarity, such as inpainting and super-resolution. However, they struggle when input and output domains significantly differ, as is the case in solar and heliophysics research.

For this reason, our work focuses on methods that directly learn the mapping between the input and the output modalities. We chose to ig-

nore the conditional DM generation process. This decision is supported by current trends in solar and heliophysics research. Indeed, to the best of our knowledge, DM have only been used for sample generation and same-modality image-to-image translation tasks in this field. For instance, [189] combines GANs and DMs to super-resolve HMI images. [190] used DM as a synthetic-to-realistic image translator. More recently, [191] have used DM to predict the next-day-same-time magnetogram of a given line-of-sight (LOS) magnetogram. In contrast to cGANs, no cross-modality image-to-image translation was performed with DM.

5.1.1 Cross-modal I2I translation in solar physics

In solar physics, cGANs were first used to generate magnetograms out of EUV images, in order to tackle the lack of solar magnetogram on the STEREO mission. [192] were the first to adopt the Pix2Pix algorithm to produce synthetic magnetograms images from STEREO/EUVI 304 Å data. Their results were limited to a maximum magnetic field strength of ± 100 Gauss in the reconstruction, and showed low correlations in solar quiet regions. [193] builds on Pix2Pix, but modifies the U-Net architecture of the generator by adding a self-attention block between the encoder and decoder parts. This allows to put more weight on active regions, and better results were reported. [194] proposed an Pix2PixHD, a modification of Pix2Pix that is capable of handling a higher dynamic range in the data via a novel adversarial loss and a multi-scale network architecture. [195] improved Pix2PixHD to handle ± 3000 Gauss dynamic range in the EUV-to-magnetogram translation, and considered multi-channel input. [196] further considered adding a correlation coefficient (CC) component to the overall objective loss function of the Pix2PixHD model, resulting in a model called Pix2PixCC, whose goal is to provide more realistic backside magnetograms in order to improve solar coronal magnetic field extrapolations. The Pix2PixCC model was utilized in [197], along with transfer learning to generate synoptic magnetograms, and in [198] for an extensive study of the role of hyperparameters in the model.

Besides the EUV-to-magnetogram cross-modality translation, cGANs were applied in other I2I translation context. We cite here the generation of white light [97] or He I 1083 nm images [199] from EUV images; of solar magnetograms and EUV images from Galileo sunspot drawings [200];

of pseudo-magnetograms from $H\alpha$ images [201] or from CaII-K images [202]. Unpaired cycle GANs may also serve in the context of instrument-to-instrument translation [203].

Our work starts by reviewing the popular Pix2PixCC model in two translation tasks, and observes that it leads to significant hallucination. This motivates the investigation of non-adversarial training strategies, and of network architectures whose behavior can adapt the input at hand, thereby offering a stronger capability to control the translation procedure and to mitigate hallucination.

5.1.2 Our contributions

Despite the large amount of works published on AI-generated solar data, few works have questioned the physical relevance of generated images. In this chapter, we scrutinize the hallucinations introduced by AI-based generation models, and raise a number of recommendations and original proposals regarding the design of I2I translation models for solar images.

The key lessons of our study can be summarized as follows.

1. We observe on two distinct representative use cases that the popular Pix2PixCC translation model is subject to significant hallucination when significantly distinct modalities are considered.
2. Our experiments demonstrate that removing the adversarial loss when training Pix2PixCC helps in mitigating the hallucinations, but also reduces the informative content of generated images.
3. We propose to enrich the architecture of the I2I translation network so as to promote the transmission of information from the input to the output, thereby resulting in more accurate reconstructions.

Specifically, our work considers the use of so-called *FiLM layers* [70] to adapt the translation process to the input. It reinforces the capability for an input modality to control the reconstruction of a corresponding output modality, thereby effectively reducing hallucination.

5.2 Data

We compare the I2I translation methods on two distinct cross-modal applications: (1) the generation of CaII-K images from WL images, and (2) the generation of magnetograms from EUV images.

5.2.1 Translation from WL to Ca II K images

We use observations recorded by the USET station located at ROB [34] and available within the SOLARNET Virtual Observatory [163]. More precisely, we consider the dataset that simultaneously monitors the photosphere in WL and the lower chromosphere in Ca II K since 2012, encompassing a temporal range comparable to a full solar cycle. We selected daily pairs of (WL, Ca II K) images, which are of size 2048×2048 , through a careful process: highest-quality daily WL images are identified, then paired with the best available Ca II K images, minimizing the time difference between them. A maximum temporal difference of four hours is considered in order to avoid discrepancies due to too large displacements. For each pair, de-rotation is applied to co-align WL and CaIIK images using the SunPy module [46]. Note that limb darkening was not corrected for either WL or Ca II K images. This step would have little effect on the contrast between faculae/plage and quiet Sun regions, which are central to this translation task.

The (WL, Ca II K) image pairs then undergo further processing as follows: sub-image of size 512×512 , called ‘segments’, are extracted from the original full disc images. These segments are centered at sunspot locations, obtained from the USET sunspot group catalog Table Access Protocol (TAP) service [204]. To reduce computational cost, the segments are resized to 256×256 pixels. Figure 5.1 shows an example of (WL, CaII-K) segment pair.

5.2.2 Translation from EUV images to magnetograms

As a second application, we aim at generating SDO/HMI magnetograms from EUV images captured by the Atmospheric Imaging Assembly (AIA)

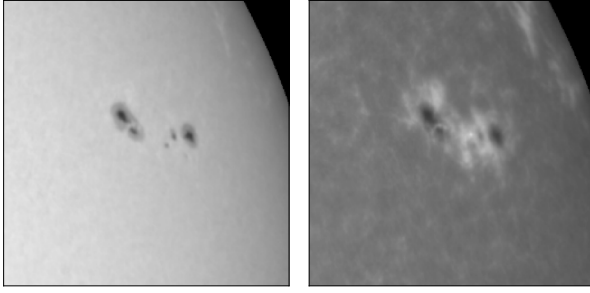


Fig. 5.1 Sample pair of WL and CaII-K segments used for **whitelight-to-calcium translation**. Segments in both modalities are spatially aligned and centered near an active region.

onboard SDO [43]. We use the SDOMLv2 dataset¹, which contains both temporally- and spatially-aligned observations of SDO/AIA and SDO/HMI. The SDOMLv2 dataset is an updated version of the original SDOML dataset, described in [205]. In this v2 revision, the creators added observations up to the year 2020, changed the resolution of the images to 512×512 pixels and updated the data to account for change in calibration of the instruments.

Our goal is to perform translation of AIA images in the $304\text{\AA}$ bandpass (further denoted AIA- $304\text{\AA}$) to LOS magnetograms provided by HMI (simply denoted HMI). To reduce sample redundancy, the original cadences of AIA- $304\text{\AA}$ (resp. HMI) images were reduced from six (resp. 12) minutes to four hours. We consider the central portion of the image, that is, a central segment of size 256×256 of the full-disk image. An example of AIA- $304\text{\AA}$ and HMI pair is shown on Figure 5.2.

As preprocessing, intensities of AIA- $304\text{\AA}$ images, measured in digital number per second (DN/s), were clipped in the $[0 \text{ DN/s}, 1000 \text{ DN/s}]$ range. This maximum value of 1000 DN/s was set to be above the 99.99th percentile of 100 randomly selected images. Analogously, the HMI LOS magnetic field measurements in Gauss (G) were clipped in the $[-1500 \text{ G}, 1500 \text{ G}]$ interval.

¹available at <https://github.com/SDOML/SDOMLv2>

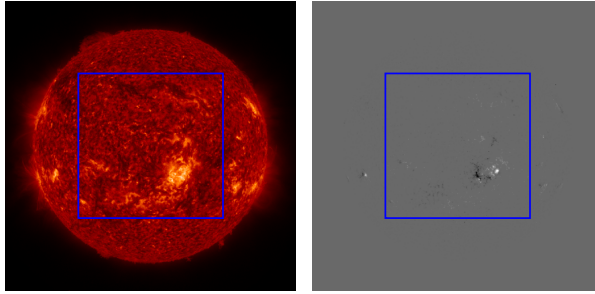


Fig. 5.2 Sample pair of AIA-304Å and HMI images used for **EUV-to-Magnetogram translation**. Only the central 256×256 pixels region is considered.

5.3 Methods

This section presents two main strategies to translate images. We first adopt the conditional adversarial network approach introduced in [104]. It employs a conventional U-Net architecture, trained using an adversarial strategy and augmented with additional losses that aim at maximizing the similarity between real and predicted images, as recommended by the Pix2PixCC model studied in [196, 198]. Our experiments reveal that, whilst popular, this model results in a significant amount of hallucinations.

Second, we explore non-adversarial strategies, beginning with a variant of the Pix2PixCC model that omit the adversarial loss during training. We then present our new method that goes one step beyond adversarial loss removal to mitigate hallucinations. It questions the network architecture, and proposes to extend the U-Net convolutional blocks to incorporate input-dependent guidance in the image-to-image translation process through input-dependent modulation of the feature maps, following the principle of FiLM [70] detailed in Section 3.2.1. For each input image, a guidance vector is extracted using a neural network and its components are linearly combined with U-Net feature maps. In practice, the parameters of the neural network in charge of extracting the guidance vector are learned during training.

Note that diffusion models were also considered as an alternative method

for our translation tasks. However, despite their emerging prominence in generative tasks, our initial experiments with DMs yielded suboptimal results for our specific applications. A comprehensive examination of the approaches we explored can be found in Appendix G.

In the rest of this section, without loss of generality, the notations are based on the WL to CaIIK translation use case. Hence, we denote a pair of corresponding input-output samples as $\mathcal{P} = (\mathcal{I}_W, \mathcal{I}_C)$, where $\mathcal{I}_W$ is the WL image, and $\mathcal{I}_C$ is the view of the corresponding Ca II K image. The reconstruction of the Ca II K image estimated by the deep learning models is denoted $\hat{\mathcal{I}}_C$.

5.3.1 Conditional adversarial network

In this work, the cGAN used is Pix2PixCC [196], a variant of the original Pix2Pix initially proposed by [104]. The Pix2PixCC model and its adversarial training are depicted in Figure 5.3. Pix2PixCC is composed of two sub-networks trained simultaneously: a generator G and a discriminator D . G is a CNN shaped according to the U-Net architecture [69]. It is given a white light view $\mathcal{I}_W$ and outputs an estimation of the corresponding calcium view $\hat{\mathcal{I}}_C = G(\mathcal{I}_W)$. The generator is trained to make it close to the original $\mathcal{I}_C$, when available. Besides, the discriminator D is given a pair of images that can either be the original pair $\mathcal{P}$ or the forged pair $\hat{\mathcal{P}} = (\mathcal{I}_W, \hat{\mathcal{I}}_C)$. The role of D is to distinguish real pairs from the forged ones. The discriminator D is implemented as a patchGAN [104] which manipulates pairs of patches (cropped in the pair of full images) instead of the image pair itself.

During training, G is updated to improve the ability to fool D using a weighted sum of three loss functions: (a) the least-squares adversarial loss function for generators [206] (b) a feature-matching (FM) loss between $\hat{\mathcal{I}}_C$ and $\mathcal{I}_C$ and (c) a correlation coefficient based loss. We denote function (a) as $\mathcal{L}_{LSGAN, G}$ and express it as

$$\mathcal{L}_{LSGAN, G}(G, D) = \frac{1}{2}(D(\mathcal{I}_W, G(\mathcal{I}_W)) - 1)^2 = \frac{1}{2}(D(\mathcal{I}_W, \hat{\mathcal{I}}_C) - 1)^2 \quad (5.1)$$

where $D(x, y)$ denotes the average of the real-or-fake results computed on patches pairs by the discriminator, considered as a probability valued 0 if

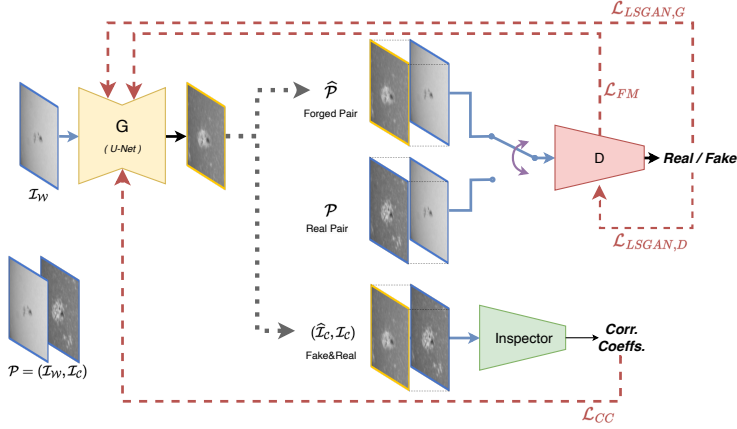


Fig. 5.3 Diagram of Pix2PixCC model and adversarial training for WL-to-CalIK translation. The generator G converts a WL image to the CalIK modality, the discriminator D distinguishes original WL-CalIK pairs from forged pairs. The *Inspector*(s) contributes to the update of the weights of G by computing correlation coefficients between original CalIK image and forged CalIK image, at multiple scales.

the input is detected as a forged pair and 1 if it is considered as a real pair by the discriminator. Loss (b), denoted $\mathcal{L}_{FM}$, takes the similarity between the reconstructed image and the target image into account to improve the reconstruction accuracy. It is given by

$$\mathcal{L}_{FM}(G, D) = \sum_{i=1}^{T_D} \frac{1}{N_i} \|D^{(i)}(I_W, I_C) - D^{(i)}(I_W, G(I_W))\| \quad (5.2)$$

with T_D the number of layers in the discriminator D , D^i the i -th layer of D , and N_i the size of the feature map at the i -th layer. Finally, loss (c) can be written as

$$\mathcal{L}_{CC}(G) = \sum_{i=0}^{T_{\downarrow}} \frac{1}{T_{\downarrow} + 1} (1 - CC_i(I_C, \hat{I}_C)) \quad (5.3)$$

where CC_i is the correlation coefficient measured at the i -th level of downsampling of $\hat{I}_C$ and I_C , and $T_{\downarrow}$ is the number of downsampling by a factor of two. The total loss minimized during the training of the generator in the Pix2PixCC framework is expressed as

$$\mathcal{L}_{tot,G}(G, D) = \lambda_1 \mathcal{L}_{LSGAN,G}(G, D) + \lambda_2 \mathcal{L}_{FM}(G, D) + \lambda_3 \mathcal{L}_{CC}(G). \quad (5.4)$$

with λ_1, λ_2 and λ_3 being hyperparameters defining the relative importance of the sub-losses. In our experiments, we used $\lambda_1 = 2$, $\lambda_2 = 5$ and $\lambda_3 = 10$. On the other hand, D is modified to improve its ability to distinguish between real and fake pairs using the *least-square* GAN loss function for discriminators:

$$\mathcal{L}_{LSGAN,D}(G, D) = \frac{1}{2}(D(\mathcal{I}_W, \mathcal{I}_C) - 1)^2 + \frac{1}{2}(D(\mathcal{I}_W, G(\mathcal{I}_W)))^2. \quad (5.5)$$

5.3.2 Non-adversarial strategies

The adversarial setup of GAN models often generates outputs with hallucinations, or unwanted artifacts. To mitigate this issue in our physics-sensitive application, we tested several approaches, described below.

Pix2PixCC without adversarial loss

We denote by $\text{Pix2Pix}_{w/o-Adv}$, a variant of Pix2PixCC trained without adversarial loss or feature modulation and utilizing an L1 loss between the reconstruction and the target:

$$\mathcal{L}_{w/o-Adv,G} = \left\| \mathcal{I}_C - \hat{\mathcal{I}}_C \right\|_1 = \frac{1}{N_p} \sum_i^{N_p} \left| \hat{\mathcal{I}}_{C,i} - \mathcal{I}_{C,i} \right| \quad (5.6)$$

where N_p is the number of pixels in images.

Modulation of internal feature maps: l2lwFiLM

To mitigate unwanted artifacts in our physics-sensitive application, we propose an alternative approach that utilizes an image-to-image model with enhanced capacity to adapt the output reconstruction to the input content. This algorithm modulates the model's internal feature maps based on Feature-wise Linear Modulation (FiLM, [70]), thereby offering the network the ability to inject input-related information through the image translation process. FiLM conditioning modulates deep neural feature maps via learned affine transformations. When the modulation factors are derived from the input signal this enables flexible task adaptation to this input.

FiLM has been used across various tasks, from visual reasoning [70] to image translation [207, 208, 95]. To our knowledge, the principles of FiLM have rarely been considered in solar physics research. In [208], FiLM layers have been used to generate surface temperature maps of stars, conditioned on the available observational phases. In the context of ground-based solar observations, which are often disturbed by Earth’s atmospheric effects, [209] has proposed a CNN to emulate a spatially variant convolution with a Point Spread Function (PSF), and uses FiLM layers to condition this convolution on the ground based instrument’s characteristics.

Formally, image-to-image translation with feature-wise linear modulation (I2IwFiLM) is implemented as depicted in Figure 5.4. To increase the

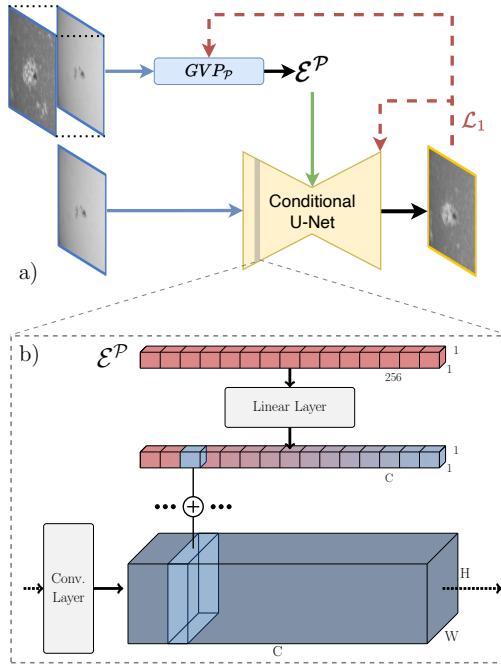


Fig. 5.4 First training stage of I2IwFiLM: a) Main U-Net model predicts $\hat{\mathcal{I}}_C$ from input $\hat{\mathcal{I}}_W$ and $\mathcal{E}^{\mathcal{P}}$, the guidance extracted from the pairs $\mathcal{P} := (\mathcal{I}_W, \mathcal{I}_C)$ using a Guidance Vector Prediction network, denoted $GVP_{\mathcal{P}}$. b) Modulation of an intermediate feature map of the U-Net model by $\mathcal{E}^{\mathcal{P}}$ using a FiLM layer.

chance that the guidance vector captures information that is relevant to the output, in a preliminary training stage, the guidance vector $\mathcal{E}^{\mathcal{P}}$ is defined based on the complete WL and CaIIK pair $\mathcal{P}$. Specifically, an encoding convolutional neural network denoted $GVP_{\mathcal{P}}$, and named *Guidance Vector Prediction* network is trained to extract a 256-dimensional guidance vector $\mathcal{E}^{\mathcal{P}}$ from the concatenated input and output signals. The guidance vector $\mathcal{E}^{\mathcal{P}}$ modulates the output of several convolutional layers in the main U-Net through dedicated linear layers. Each modulated convolutional layer is paired with its own linear layer which takes $\mathcal{E}^{\mathcal{P}}$ as input and produces an additive modulation signal that is combined with the convolutional layer's output (see Figure 5.4b). Then, in a second step, another CNN is trained to predict the guidance vector $\mathcal{E}^{\mathcal{P}}$ from the input signal only. We denote $\widehat{\mathcal{E}^{\mathcal{P}}}$ the predicted guidance vector. The second training step is depicted in Figure 5.5. Analogously to the first training stage, the network trained to predict $\mathcal{E}^{\mathcal{P}}$ from the sole WL image is denoted $GVP_{\mathcal{W}}$.

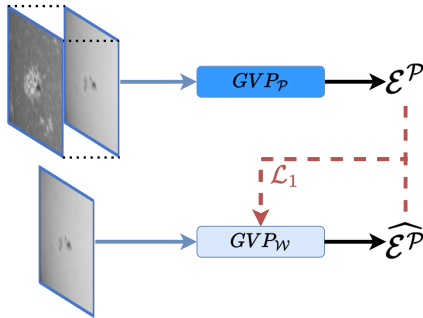


Fig. 5.5 Second training stage of I2IwFiLM. $GVP_{\mathcal{W}}$, composed of convolutional layers followed by a Multi-Layer Perceptron, learns to recover $\mathcal{E}^{\mathcal{P}}$ vectors directly from $\mathcal{I}_{\mathcal{W}}$. $GVP_{\mathcal{P}}$ (dark blue) is frozen during this stage while the main U-Net is left unconstrained to adapt to discrepancies due to guidance modification.

The idea of training a FiLM guidance vector generator in two steps has initially been introduced in [210, 211] to deal with image-to-image tasks corresponding to super-resolution, deblurring and inpainting (and not cross-modal image translation). However, in contrast to our approach, as depicted in Appendix H, DiffI2I [211] uses a diffusion model to generate the guidance vector from a noise sample through a conditional denoising

process. We recommend a straightforward prediction of the guidance vector using a feed-forward MLP, without resorting to a diffusion process. In that way, we avoid introducing randomness (through the diffusion process random seed) when computing the guidance vector, and make the image translation deterministic. Our experiments demonstrate the relevance of our approach.

5.4 Experiments

Our experiments investigate deep learning models trained with or without adversarial loss, using input-dependent feature modulation or not, for two I2I translation tasks. To evaluate how adversarial training and feature modulation impact physics-sensitive translation, we compare three models:

1. Pix2PixCC as proposed by [198], serving as the baseline model.
2. Pix2Pix_{w/o-Adv}, a variant of Pix2PixCC trained without adversarial loss, utilizing an L1 loss between the reconstruction and the target.
3. I2IwFiLM, implementing feature modulation through input-dependent guidance $\hat{\mathcal{E}}^{\mathcal{P}}$, directly predicted from WL only. The training of this model minimizes L1 loss for both translation (reconstruction vs. target) and $\hat{\mathcal{E}}^{\mathcal{P}}$ prediction.

Section 5.4.1 details the train-test split adopted for both translation tasks. Model evaluation metrics are detailed in Section 5.4.2.

5.4.1 Train-test data splitting

To create the test sets of our two translation tasks, two months were selected from each year for both datasets: one winter month and one summer month, to ensure sample diversity. All samples corresponding to these selected months were allocated to the respective test sets. All other samples were selected for training, except the ones falling within a 12-day window on either side of the test samples. The window size was chosen as half of the Sun's rotation period, as observed by [10].

This results in a WL-to-CalIK translation dataset including 6,387 samples for training and 1,410 for testing. For the AIA-304Å-to-HMI translation, the number of samples is 17,072 for training and 3,220 for testing.

5.4.2 Performance metrics

Our image-to-image translation models were evaluated at two distinct levels: generic image reconstruction level and application-specific level. Generic evaluation was performed using metrics detailed in Section 5.4.2, while application-specific performance assessment was performed using metrics presented in Section 5.4.2 for AIA304Å-to-HMI translation and in Section 5.4.2 for WL-to-CalIK.

General metrics

The reconstruction performances of the deep learning models were compared using metrics that are common in the context of image reconstruction. We used the *root mean squared error* (RMSE), *peak signal-to-noise ratio* (PSNR), *structural similarity index measure* (SSIM) and its extension, the *multi-scale structural similarity index measure* (MS-SSIM). The detailed descriptions of these metrics can be found in Appendix I.

Metrics specific to WL-to-CalIK application

In the case of WL-to-CalIK, a physically relevant metric consists in measuring the capacity of the models to reconstruct the plages and extended network of the target calcium image. To do so, the plage segmentation masks computed from the original CalIK and from the calcium images generated from WL are compared.

The segmentation masks are defined on full-disc images where the center-to-limb brightness variation was corrected. They include pixels brighter than a threshold defined as:

$$\mu_{fd} + m * \sigma_{fd} \quad (5.7)$$

where μ_{fd} is the mean brightness value of the full-disc, σ_{fd} its standard deviation, and m a user-defined multiplicative factor, set to 4 in our experiments.

We employed two metrics to compare the segmentation masks. First,

the *Plage Pixel Ratio* (PPR) is defined as the ratio of plage pixels in generated and original images. Second, the *Intersection over Union* (IoU) is computed as the ratio between the number of pixels that belong to both masks and the number of pixels that belong to at least one mask.

Metrics specific to the AIA-to-HMI application

In the context of AIA-to-HMI translation, HMI magnetograms exhibit bipolar pixel distributions centered at zero, with features of interest occupying a relatively small spatial extent. This characteristic poses challenges for conventional evaluation metrics such as SSIM, which may yield artificially elevated values due to large regions of zeros in reconstructed images, as observed by [198]. To mitigate this limitation, we introduce ΔSSIM , defined as the difference between two measures: (1) the SSIM between the generated and target magnetograms ($\text{SSIM}_{\text{model}}$) and (2) a baseline SSIM computed between the target magnetogram and a reference image filled with zero values ($\text{SSIM}_{\text{flat}}$). Formally, ΔSSIM is computed as

$$\Delta\text{SSIM} = \text{SSIM}_{\text{model}} - \text{SSIM}_{\text{flat}} \quad (5.8)$$

This differential metric accounts for the inherent sparsity of magnetic features in HMI observations and allow performance comparison focused on active regions.

5.5 Results

5.5.1 Generic reconstruction quality metrics

An overview of the reconstruction performance can be seen in Table 5.1 for WL-to-CaIIK and AIA304Å-to-HMI translations. On this table, one can observe that removing adversarial training improves reconstruction performance: Pix2PixCC_{w/o-Adv} outperforms Pix2PixCC across all generic metrics, suggesting non-adversarial training provides a stronger foundation for physics-sensitive I2I tasks. Furthermore, incorporating modulation based on FiLM principles further enhances translation quality, as evidenced by the superior metrics of I2IwFiLM over Pix2PixCC_{w/o-Adv}.

We interpret these results as a consequence of the input lacking complete information for perfect target reconstruction, as evidenced by spa-

Task	Model	RMSE ↓	PSNR [dB] ↑	SSIM ↑	MS-SSIM ↑
WL-to-CaIIK	Pix2PixCC	16.77 ± 8.01	24.51 ± 3.83	0.72 ± 0.09	0.75 ± 0.12
	Pix2PixCC _{w/o-Adv}	15.07 ± 7.74	25.56 ± 4.10	0.75 ± 0.07	0.80 ± 0.10
	I2IwFiLM (Ours)	14.44 ± 7.07	25.90 ± 4.12	0.80 ± 0.09	0.80 ± 0.11
AIA304Å-to-HMII	Pix2PixCC	$15.61 \times 10^{-3} \pm 5.56 \times 10^{-3}$	36.62 ± 2.88	0.96 ± 0.03	0.96 ± 0.03
	Pix2PixCC _{w/o-Adv}	$14.61 \times 10^{-3} \pm 5.00 \times 10^{-3}$	37.16 ± 2.76	0.94 ± 0.02	0.96 ± 0.02
	I2IwFiLM (Ours)	$13.23 \times 10^{-3} \pm 5.06 \times 10^{-3}$	38.10 ± 2.98	0.96 ± 0.03	0.96 ± 0.03

Table 5.1 Quantitative results of generic image reconstruction metrics for **whitelight-to-calcium** and **EUV-to-Magnetogram** translation tasks. GAN-based model Pix2PixCC is compared with non-adversarial models: Pix2PixCC_{w/o-Adv} and our proposed I2IwFiLM including input-dependent conditioning.

tial mismatches between WL faculae and CaIIK plages. While adversarial models like Pix2PixCC produce visually compelling results by over-interpreting available data to match target distributions, this leads to hallucinations and lower reconstruction metrics.

In the remainder of this section, we analyze the performance of Pix2PixCC and I2IwFiLM across both image-to-image translation tasks. Section 5.5.2 presents a comparison of the models for the WL-to-CaIIK translation task, while Section 5.5.3 focuses on the AIA-304Å-to-HMI translation.

5.5.2 WL-to-CaIIK

Figure 5.7 and Figure 5.8 illustrate examples of white light to calcium modality translations at different locations on the solar disk. The former shows high-quality reconstructions while the latter demonstrates subpar examples with additional physically inaccurate plage regions.

These figures demonstrate typical model behaviors: Pix2PixCC, trained in adversarial setup, produces visually appealing results with high-frequency features but exhibits a tendency to generate plage regions where there are none, a behaviour termed as ‘hallucination’. In contrast, I2IwFiLM generates outputs that appear blurrier but achieve more physically accurate reconstructions without hallucinations. Additionally, Figure 5.7 and Figure 5.8 reveal the variations in translation quality with respect to the angular distance $\sin(\theta)$.

Analysis of reconstruction metrics binned by $\sin(\theta)$, presented in Figure 5.9, quantifies these variations. Two distinct regimes emerge in the translation quality: regions with $\sin(\theta) < 0.6$ show poor translation performance, while regions with $\sin(\theta) > 0.6$ exhibit increasingly accurate reconstructions as they approach the limb. Those two regimes are explained by Figure 5.6, which shows that in both input and output images, the contrast between AR features increases with $\sin(\theta)$. Alongside sunspots, visible AR features are faculae in WL images, while plage and extended network are visible in Ca II K images.

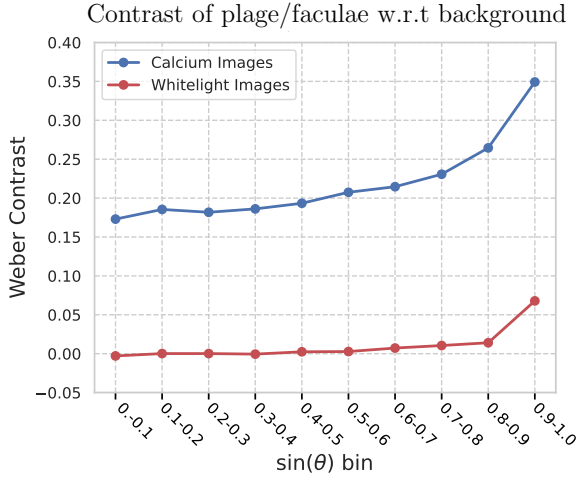


Fig. 5.6 Weber contrast between plage pixels intensity and background pixels intensity along angular distance bins in WL and CaIIK images. The Weber contrast is defined as the ratio of luminance difference between foreground and background over the background luminance (see Appendix J). Quiet Sun is considered background while foreground is the combination of plage and extended network in Ca II K images, and faculae in WL images.

Beyond the impact of $\sin(\theta)$, Figure 5.9 also shows that I2IwFiLM consistently outperforms Pix2PixCC with respect to every metric across all angular distances bins.

Application-specific performance

Figure 5.10 evaluates the performance of our models using application-specific metrics. It presents histograms of PPR and IoU metrics on test set samples for 3 ranges of $\sin(\theta)$: (a) the center of the Sun ($0 \leq \sin(\theta) \leq .5$) where faculae are indistinguishable from the background in WL images, (b) the transitional region ($.5 \leq \sin(\theta) \leq .75$), and (c) the region nearing the limb ($.75 \leq \sin(\theta) \leq 1.$) with high contrast between faculae and background.

Both Pix2PixCC and I2IwFiLM models demonstrate improved reconstruction performance as $\sin(\theta)$ increases, with higher mean PPR and IoU values and the reduction in their standard deviations. However, both mod-

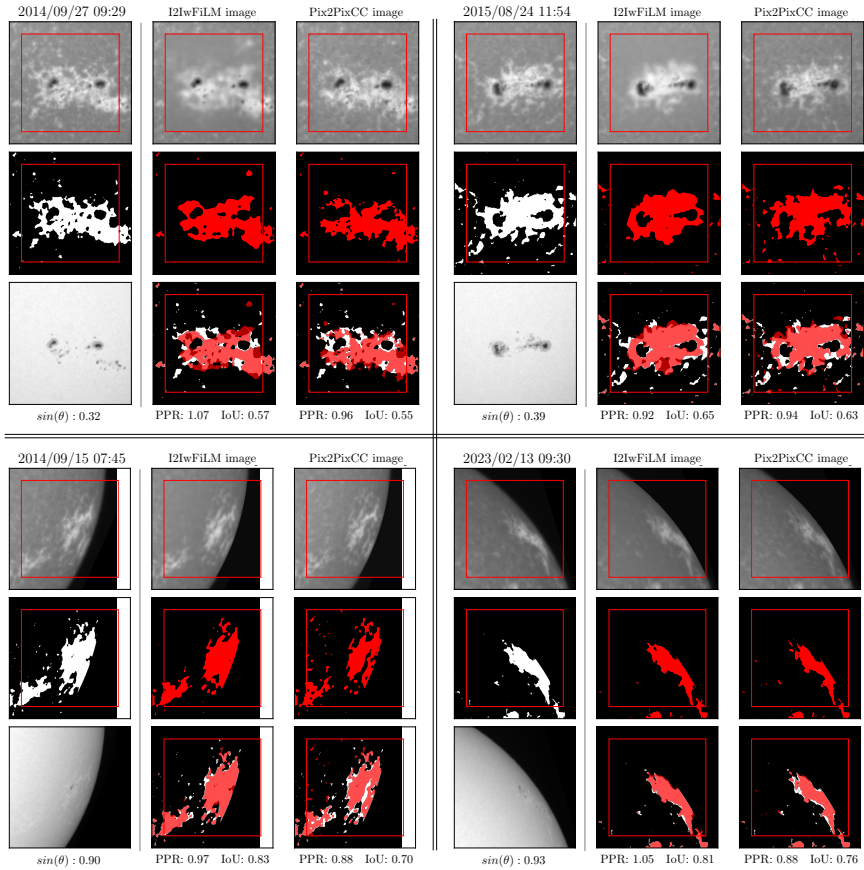


Fig. 5.7 Highest-quality CaIIK reconstructions of ARs near the center of the solar disc (Top grids) and near the limb (Bottom grids). In each 3x3 grid: the top row shows the original CaIIK image (left) and its reconstructions by I2IwFiLM (middle) and Pix2PixCC (right). The middle row shows segmentation masks of plages visible in the original CaIIK image (left) and the corresponding reconstructions by I2IwFiLM (middle) and Pix2PixCC (right). Bottom row shows input WL image (left) and overlays of the target plage mask with the model-generated plage masks from I2IwFiLM (middle) and Pix2PixCC (right), facilitating a direct comparison.

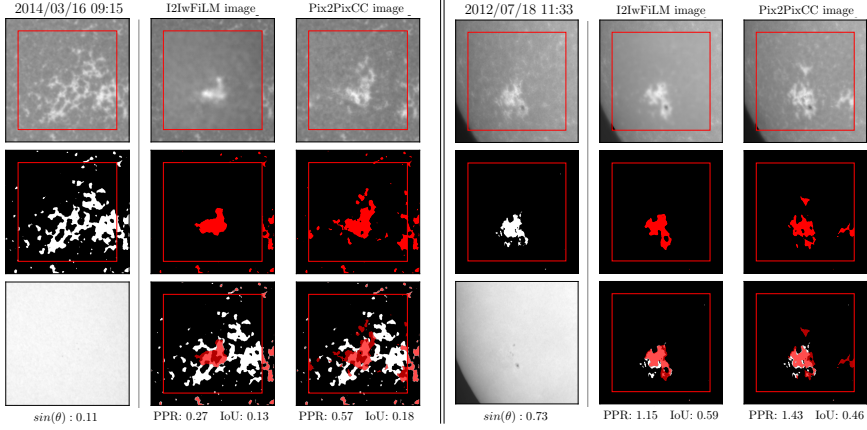


Fig. 5.8 Subpar CaIIK reconstructions of ARs near the center of the solar disc (left grid) and near the limb (right grid). In each 3x3 grid: the top row shows the original CaIIK image (left) and its reconstructions by I2IwFiLM (middle) and Pix2PixCC (right). The middle row shows segmentation masks of plages visible in the original CaIIK image (left) and the corresponding reconstructions by I2IwFiLM (middle) and Pix2PixCC (right). Bottom row shows input WL image (left) and overlays of the target plage mask with the model-generated plage masks from I2IwFiLM (middle) and Pix2PixCC (right), facilitating a direct comparison.

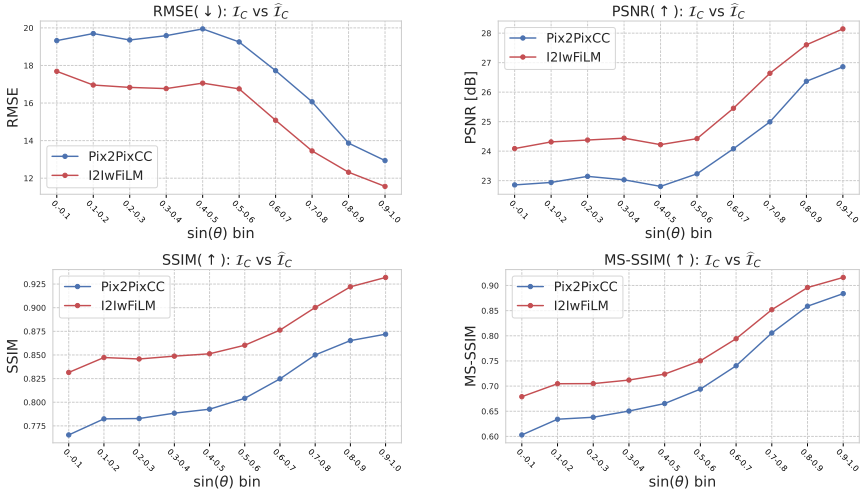


Fig. 5.9 Comparison of Pix2PixCC and I2IwFiLM performances using generic image reconstruction metrics along the angular distance bins to the center of the Sun. The center of the Sun corresponds to $\sin(\theta) = 0$, while the limb corresponds to $\sin(\theta) = 1$.

els appear to be limited by the visibility of faculae, resulting in the production of plage regions smaller than the actual ones in the targets. This limitation is revealed by both PPR and IoU achieving their best values in the regions where faculae are most easily discernible.

Examining the models individually, I2IwFiLM exhibits low average PPR and IoU with large variance for $\sin(\theta) \leq .5$, indicating that only a fraction of plages in these active regions are consistently reconstructed, as observed in Figure 5.7 (top row) and Figure 5.8 (left). In contrast, Pix2PixCC achieves a respectable PPR but poor IoU for these regions, suggesting consistent incorrect reconstruction of plages due to hallucinations. When comparing the relative performance of the two models, it is noteworthy that the mean PPR of I2IwFiLM initially lags behind that of Pix2PixCC for $\sin(\theta) \leq .5$ but surpasses it for $\sin(\theta) > .75$, accompanied by a substantial reduction in its standard deviation.

Given these results, I2IwFiLM appears to be more suitable for white light-to-calcium translation compared to Pix2PixCC. These results also illustrate the possibility of reconstructing CaIIK plages from faculae observed in WL when these are located at an angular distances of about 40 degrees or more from the disk center. In future work, these generated images could be compared to observations from data archive, and help to remove inconsistencies these historical archive might have [212]. To reconstruct solar irradiance at CaIIK line though, one would need to obtain information also at smaller angular distances. This would requires to process not only a single image, but a time series of observations, to allow transferring of information from one timestamp to another.

5.5.3 AIA-304Å-to-HMI

An example of AIA-304Å-to-HMI translation is depicted in Figure 5.11. As observed for WL-to-CaIIK translation, Pix2PixCC generates visually more appealing images than I2IwFiLM, but hallucinations degrade quantitative results.

The example illustrates that flat HMI images inherently share high metric values when compared with target images, making a fair comparison of models challenging. Both models accurately locate the active regions and

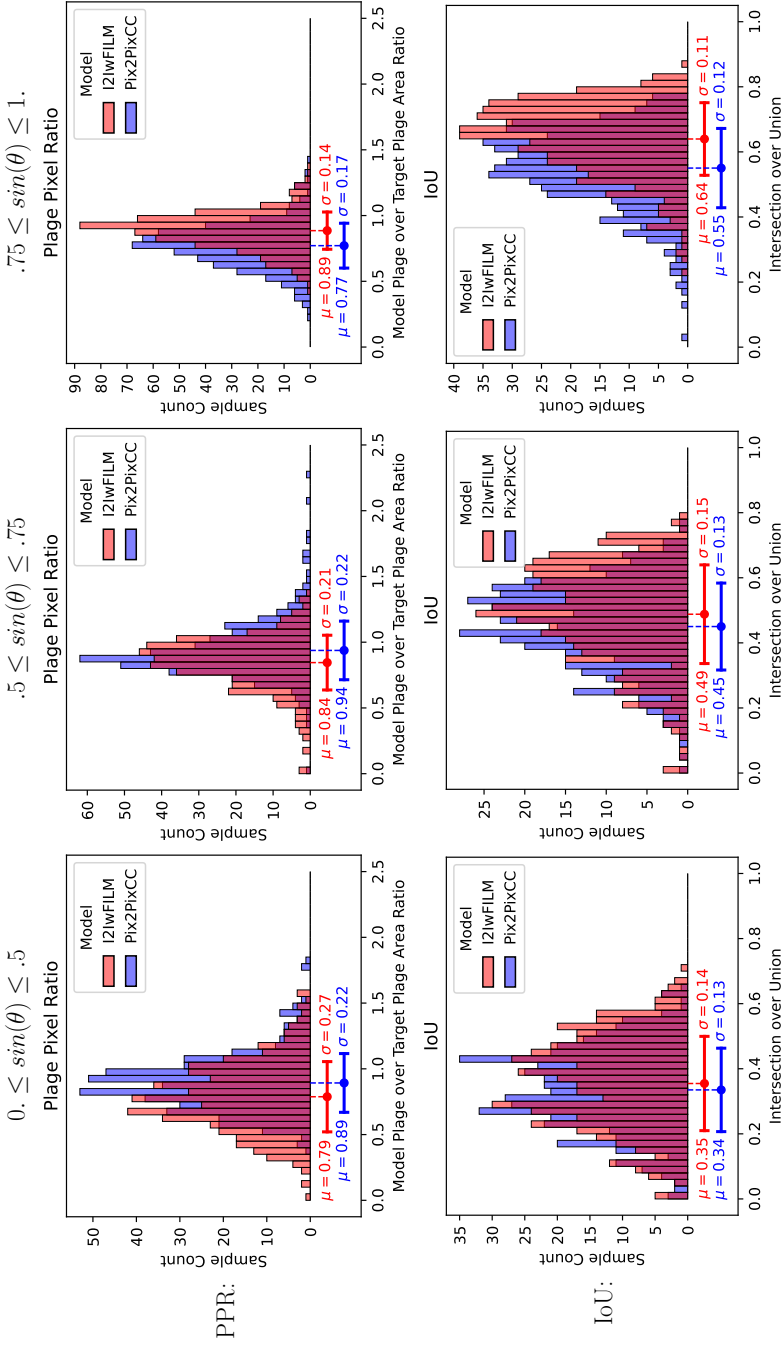


Fig. 5.10 Plage reconstruction performance of Pix2PixCC and I2IwFiLM models for active regions located near the center of the Sun (left column), at intermediate distance to the center (middle column) and near the limb (right column). Top row shows PPR metric histograms, bottom row shows intersection over union (IoU) histograms.

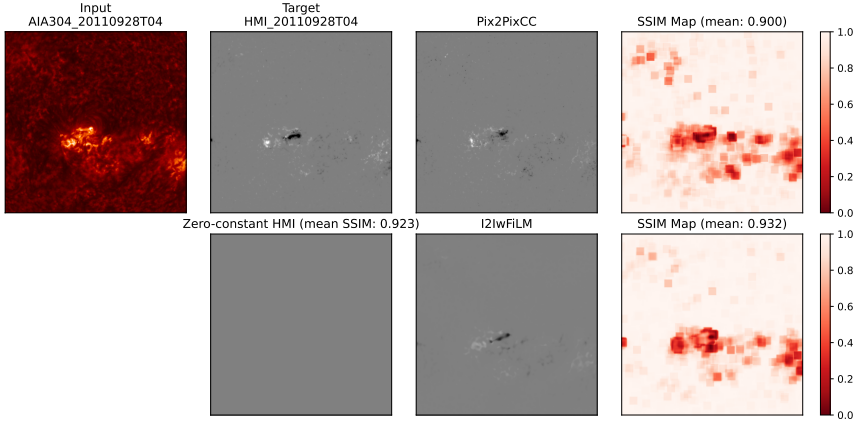


Fig. 5.11 Example of reconstructed HMI Bz using Pix2PixCC model (GAN-based) and I2IwFiLM (ours). From left to right: Input AIA-304Å image, target HMI image compared with flat HMI, model-reconstructed HMI images, structure similarity maps between model-reconstructed HMI and target.

their polarities in their output, but fail at outlining details of active regions from AIA-304Å images. Indeed, the active regions in the generated images only faintly correspond to the target, as indicated by the SSIM maps.

Application-specific performance

To assess whether a translation model does a better job than a trivial solution adopting a zero image as a reconstructed image, we have introduced the Δ SSIM metric, computed on the entire image (see Section 5.4.2). The left part of Figure 5.12 shows histograms of Δ SSIM metrics of all test samples for both Pix2PixCC and I2IwFiLM models. The histogram for Pix2PixCC ($\mu = -0.019$, $\sigma = 0.0096$) lies entirely on the left of the vertical line at Δ SSIM = 0, indicating that Pix2PixCC’s reconstructions correspond to the target even less than a flat HMI image due to the hallucination it generates. In contrast, the histogram of I2IwFiLM ($\mu = 0.0053$, $\sigma = 0.0052$) is positioned on the right of the line, demonstrating that I2IwFiLM performs better than a zero-constant reconstruction of the target.

To further investigate the reconstruction quality of active regions, we computed the same Δ SSIM histograms at the level of active region level.

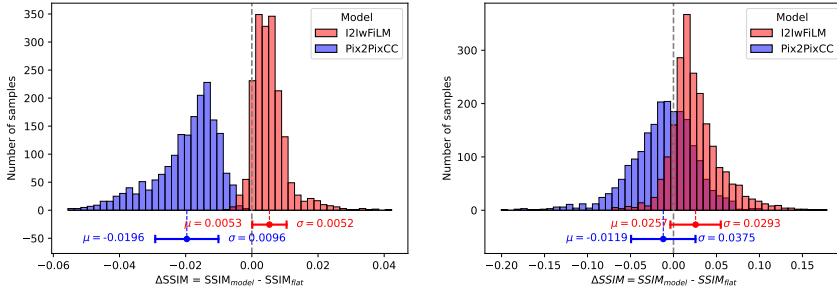


Fig. 5.12 Δ SSIM histograms for I2IwFiLM and Pix2PixCC models computed on images in AIA-304Å-to-HMI task test set. (Left): Entire images. (Right): Crops centered on large ARs only (McIntosh class C and larger).

To do so, crops of ARs were extracted from test samples and Δ SSIM metrics were computed over those crops. Only active regions of classes C and larger according to the modified Zurich classification scheme [24] were considered in this analysis, while small-sized active regions were excluded. The classes of active regions were obtained from the USET sunspot groups catalog [204]. The resulting Δ SSIM histograms for active regions are shown in the right part of Figure 5.12. By focusing solely on the sparse ARs, the histograms become more widely distributed compared to the full-image analysis. This widening occurs because the results in ARs are less diluted by the overall good performance in the large zero-valued regions, which typically narrow the histograms obtained on entire images, and shift them closer to zero. The observations made for the full-image analysis still hold: the positive average of I2IwFiLM's histogram ($\mu = 0.0293$) with narrow standard deviation ($\sigma = 0.0293$) and the negative average of Pix2PixCC's histogram ($\mu = -0.0119$) with higher standard deviation ($\sigma = 0.0375$) shows that the hallucinations produced by Pix2PixCC make it less suitable for physically accurate image translation than I2IwFiLM.

5.6 Discussion and Perspectives

Solar physics benefits from continuously evolving observational capabilities that provide new measurement modalities. However, these modern

observations were not available in the past, creating gaps in historical data. We leverage deep learning methods to reconstruct these missing historical modalities from physically related measurements that were available at the time of interest. GANs have become the standard approach for such reconstructions.

We address two translation tasks in solar physics: WL-to-CaIIK and EUV-to-magnetogram. Unlike traditional image reconstruction tasks—e.g., superresolution—where input contains most of the target information, the modalities involved in our cross-modal translation tasks generally share only a fraction of their respective visual cues. This lack of guidance makes GANs prone to hallucinations, driven by the need to match target modality statistics, a critical flaw for scientific applications. Therefore, we propose a novel deep learning approach named I2IwFiLM that replaces adversarial training by enhanced guidance from input modalities. I2IwFiLM relies on a side model to extract information from the input, to modulate the feature maps of the main translation model. This design preserves information content while benefiting from non-adversarial training to reduce hallucinations. Our study shows that, while methods based on GANs produce more visually attractive results, I2IwFiLM generates images closer to the target images, with better reconstruction metrics.

Following the directions for accurate evaluation of generative models presented in Section 3.3.3, we introduced domain-specific metrics for both WL-to-CaIIK and EUV-to-magnetogram tasks. These metrics allow us to rigorously assess and compare the physical accuracy of the different methods. The evaluations demonstrate superior performance for our I2IwFiLM method, across these metrics. The approach established in this chapter allows for physically accurate reconstruction of missing solar observations, providing exploitable data for downstream space weather and solar physics research. For example, I2IwFiLM can immediately generate chromospheric images from reliable historical photospheric data, serving as a benchmark to assess archival chromospheric observations from potentially miscalibrated instruments.

It is important to note the variability in performance between our two translation tasks. The WL-to-CaIIK translation task yielded solid results with high physical accuracy, demonstrating that our I2IwFiLM approach can effectively reconstruct missing historical observations in this context.

However, the EUV-to-Magnetogram translation was less successful, suggesting that this particular cross-modal translation presents additional challenges. This variability highlights the task-specific nature of cross-modal translations in solar physics. While both magnetograms and EUV imagery capture different manifestations of the same underlying magnetic field, their relationship appears more complex than initially anticipated, potentially requiring additional physical constraints or alternative architectural approaches in future work.

Looking forward, our work on WL-to-CaIIK and EUV-to-Magnetogram translation tasks could be, in a first step, extended in two main directions: improvement of the methods and further scientific validation. Regarding methodological improvements to benefit more challenging translation like EUV-to-Magnetogram, performance may be enhanced by leveraging multidimensional input, either from time-series data to exploit temporal continuity or from multiple channel observations to incorporate additional physical information. This would be an approach similar to [196], where some reference data pairs were used besides the EUV information as input of models generating magnetograms from EUV images. During training, EUV information consists of three SDO/AIA channels at time of interest, and the reference data pair is composed of three SDO/AIA images and an SDO/HMI magnetogram, which were observed one solar rotation prior to the time of interest. At inference, the algorithm inputs the three STEREO-EUVI channels (for which the corresponding magnetogram must be generated) and the reference data pairs are built using the latest corresponding SDO data, i.e. the frontside data selected by considering the separation angle between STEREO and SDO spacecrafts. In other words, it is expected that the frontside magnetograms will constrain the model with information about the overall magnetic field distribution.

Regarding scientific validation, further work could involve using our generated images as boundary conditions in physical simulations, as explored by [197] in the context of solar wind prediction. This would allow us to compare the quality of resulting simulation outputs against those obtained with original observations. While our task-specific metrics complement traditional image comparison approaches, incorporating the generated images into physical simulations would represent the ultimate test of their scientific fidelity.

Then, in a second step, beyond the translation tasks addressed in this work, the cross-modal translation approach developed in this chapter opens exciting possibilities for solar physics data analysis. A particularly intriguing application would be generating White Light images from human-made drawings at the ROB, especially valuable for the USET catalog's extensive sunspot drawing collection. This new translation task differs fundamentally from those addressed this chapter, making I2IwFiLM potentially unsuitable. In our WL to CaIIK translation, both modalities contained rich but different information, with limited features in common. Here, drawings and WL images represent the same modality, but drawings capture only a fraction of the information a CCD camera provides, lacking faculae and center-to-limb variation. Given this need to generate substantial missing information within the same modality, generative models like GANs and Diffusion models might be more appropriate, as their inherent randomness could produce plausible completions of the drawings.

Ultimately, after thorough scientific validation (i.e. through physical simulations), the true value of such translations would be realized when the generated images enable new scientific discoveries in solar physics research. A particularly promising direction would be studying the Canopy effect in the chromosphere by analyzing the expansion factor between faculae in WL images (or plages in our generated CaIIK images) and plages in the original CaIIK observations.

6

Conclusion

Throughout this thesis, we have explored the transformative potential of deep learning techniques in solar physics research. Our investigations have demonstrated the dual value of these approaches: as powerful assistive tools for ground-based sunspot monitoring through automatic segmentation and classification tasks, and as effective methods to fill observational gaps by translating imagery between different wavebands while preserving scientific usability. By bridging computational science and solar physics, our work demonstrates how cross-disciplinary approaches can yield innovations that benefit multiple fields of study. On one hand, this integration has enabled us to develop deep learning solutions that serve as assistive tools for human researchers in their observational tasks. On the other hand, it has allowed us to pursue innovative applications that, with further development and integration into downstream research, may contribute to enhancing our understanding of solar phenomena.

The intersection of these disciplines has proven fertile ground for scientific advancement. As we conclude this thesis journey, we've both answered our initial research questions and uncovered important new inquiries. This final chapter synthesizes our key contributions and findings, while also exploring promising directions to tackle emerging challenges in solar physics.

6.1 Research Questions recap and Contributions

Sunspot group identification and characterization Our first research direction focused on developing an automated assistance tool for the experts at the Royal Observatory of Belgium in their critical task of daily solar activity monitoring. This investigation was guided by one fundamental research question:

Q1 Can a deep learning model reliably detect and characterize sunspot groups from ground-based images despite atmospheric artifacts and weak annotation?

In response to this question, Chapter 4 presented SunSCC, an integrated framework for analyzing ground-based white light observations. SunSCC uses deep learning models for sunspot segmentation and group classification, connected by a dedicated algorithm for clustering individual sunspots into groups.

The framework employs an innovative supervision strategy that alternates between multiple automatically-generated pseudolabels per training sample. This approach shows robustness against atmospheric artifacts, detects typically-missed small sunspots, and minimizes manual annotation to the test set only.

The developed clustering methodology leverages sunspot group extents and individual sizes to successfully replicate USET catalog patterns. The proposed CNN architecture for McIntosh classification outperforms existing solutions relying on satellite-based imagery, despite using ground-based observations and the inherent limitations of the subjective McIntosh scheme.

Finally, we proposed an ensemble learning approach to identify ambiguous cases and detect pre-existing USET catalog annotation errors. By highlighting challenging cases to human experts, it establishes a robust quality control mechanism that enhances the reliability of solar activity monitoring.

Cross-modality image translation Our second research direction focused on developing a model capable of predicting missing observations in one modality from available observations in another modality. This approach aims to fill gaps in solar observation records where certain historical observation modes are unavailable. This investigation was guided by two fundamental research questions:

Q2 Is it possible to fill gaps in historical records by exploiting observations of another correlated modality?

Q3 Do generative models produce physically reliable predictions that can be used in heliophysics research?

In Chapter 5, we explored two image translation tasks: predicting CaIIK images from white light observations and magnetograms from EUV data. Following common practice in solar physics research, we first investigated GAN-based solutions for their ability to generate realistic details. Yet, we found that previous works overlooked how this property can introduce physically unfounded artifacts, especially when source and target modalities display substantially distinct characteristics.

Motivated by this insight, we proposed I2IwFILM, a novel model architecture for solar physics. By combining advanced feature extraction with feature map modulation, this approach enables enriched reconstruction of the target modality through explicitly extracted source features. Evaluations demonstrated that I2IwFILM generates physically consistent predictions with fewer hallucinations compared to GAN-based approaches.

A detailed comparison of I2IwFILM and adversarial methods employed standard reconstruction metrics and domain-specific criteria assessing physical feature correspondence. The systematic evaluation established the approach's superior reconstruction quality and reliability for cross-modality image translation tasks requiring high input-related guidance. These results suggest potential use of generated images for applications in future solar physics models.

6.2 Perspectives

While this thesis demonstrated clear benefits in automating observation tasks and bridging data gaps, Sections 4.5 and 5.6 highlighted important limitations and improvement paths that shape future research directions. In this final perspectives section, rather than reiterating the detailed improvements and challenges already presented in previous chapters, we step back to consider overarching principles that should guide future work at the intersection of deep learning and solar physics. As this interdisciplinary domain evolves, we must critically evaluate how technical advancements can best serve scientific understanding.

As research progresses, we must first recognize that although further work may increase numerical performance metrics, these advances should not be pursued without critical evaluation of their scientific value. The McIntosh sunspot classification scheme perfectly illustrates this methodological challenge: the inherent ambiguity in its class definitions establishes a theoretical ceiling on achievable performance regardless of model sophistication. Pursuing performance improvements beyond this natural ceiling risks overcomplicating models without meaningful scientific gains.

Instead, development should prioritize interpretable models that preserve domain knowledge and acknowledge underlying uncertainties. Keeping humans in the loop remains essential, with deep learning serving as an assistive tool rather than an autonomous replacement for expert analysis. The interdisciplinary collaboration between solar physicists and computer scientists creates natural checks and balances, ensuring algorithms remain scientifically meaningful rather than merely statistically impressive.

This collaborative approach becomes particularly important when considering that deep learning as a discipline remains in its relative infancy within specialized scientific domains like solar physics. As with many emerging technologies, deep learning methods often experience a period of inflated expectations before their capabilities and limitations are realistically understood. Our work contributes to this maturation process by cautioning against adopting generative methods like GANs simply because they produce visually appealing results. As we demonstrated in our cross-

modal translation experiments, scientific usability must take precedence over aesthetic quality when evaluating model outputs.

Building on this principle of scientific usability, generated solar imagery must preserve the physical characteristics and quantifiable parameters that make them valuable for subsequent analysis. Our evaluation of cross-modal translation models demonstrated that while visual resemblance is important, maintaining key physical properties is essential for scientific applicability. By balancing skepticism with exploration, researchers can adopt approaches complementing methodological foundations rather than disrupting them. If this critical spirit is maintained, new image translation tasks can be envisioned for solar physics applications, from historical data exploitation to building more accurate physical models or simulations.

Looking toward the future, the methodologies developed in this thesis will continue to evolve alongside both computational and solar physics advances. We emphasize that the most successful approaches will emerge from sustained dialogue between these fields. Each innovation should be evaluated based on its capacity to address long-standing scientific challenges rather than following technological trends. As deep learning tools mature in the solar physics domain, they have the potential to accelerate discovery by revealing patterns in our vast solar observation archives that might otherwise remain hidden to human analysis alone.

The integration of deep learning into solar physics research workflows represents not merely a technical advancement but a fundamental shift in how we approach observation, analysis, and theoretical development in heliophysics. By maintaining rigorous scientific standards while embracing computational innovation, we can ensure that this integration yields meaningful contributions to our understanding of solar phenomena. Ultimately, these advances may enhance our ability to predict solar activity and its impacts on Earth's technological systems, space exploration capabilities, and the broader heliospheric environment.

6.3 Final Thoughts

As a cooperation between UCLouvain and the Royal Observatory of Belgium, this thesis exemplifies the value of interdisciplinary research in developing tools that enhance rather than replace human expertise. This thesis demonstrated that several tasks in solar physics research can be addressed using deep learning tools, from historical data curation, exploitation and enhancement. Throughout this work, our cooperation maintained a critical perspective on deep learning's achievements and limitations, both in performance and methodology, ensuring solutions that prioritize scientific precision while remaining practically usable. The promising results and methods developed open exciting perspectives for future research using such tools. It has been an honor to contribute to this field through this thesis work and the fruitful collaboration between our institutions.



Appendices

A USET observation equipment

The Uccle Solar Equatorial Table (USET) facility operates at the Royal Observatory of Belgium (ROB) in Uccle, Brussels. Since the 1940s, this station has maintained continuous solar observations. Its instruments are housed in a 10m high tower to minimize surface-level atmospheric turbulence. The observation equipment rests on a tilted table to align with the Earth rotation axis. A motorized mount enable precise tracking of the Sun's apparent motion. The equipment is shown in Figure A.1.

The station's original instrument is a 160mm visual refractor. This refractor has been dedicated to sunspot monitoring using drawings since the 1940s, preserving a continuous database of solar activity. A 150mm white light (WL) refractor was added to capture solar photospheric images in support to the historical refractor. To this day, the historical refractor and th WL refractor continue to provide data for the long-term solar activity database, with drawing still being used as the reference method for sunspot counting. Chromospheric observations come from two additional instruments: an 80mm H-alpha refractor and a 132mm CaII K refractor. The WL and H α refractors began digital operations in 2002 with CCD integration. The CaII K refractor joined the digital suite in 2012. Each dig-

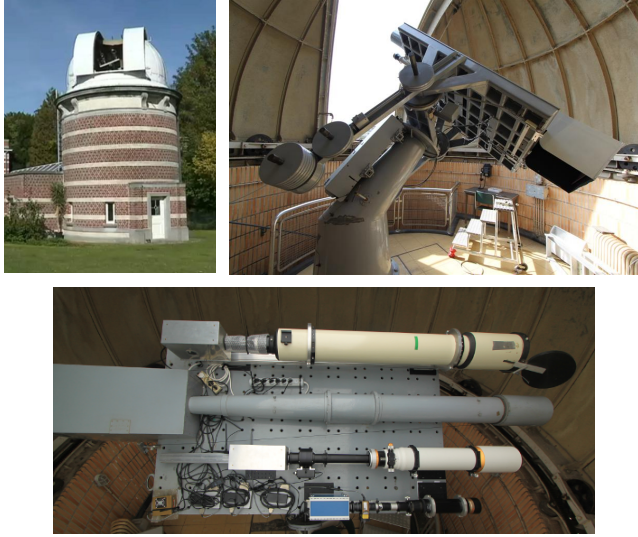


Fig. A.1 Uccle Solar Equatorial Table (USET) facility at the Royal Observatory of Belgium.(Top Left) observation tower, (Top Right) table with mounted instruments, (Bottom) close-up of the instruments.

ital instrument features a high-resolution 2048×2048 CCD camera. These cameras capture images at 15-minute intervals during regular operations. When solar flares or eruptions occur, the imaging rate can accelerate to 4 frames per second to track these dynamic events.

B Clustering method details

Here we provide details on the clustering method. Let S_I be the set of sunspots in an image I , each sunspot $s \in S_I$ is defined by its latitude coordinate l_s , its longitude coordinate L_s and its area a_s in *millionth of hemisphere* [μHem]. The probability density function associated to the observed sunspots in an image I is defined as a weighted sum of kernels

$$pdf(l, L) = \sum_{s \in S_I} a_s * k \left(\left(\frac{l - l_s}{h} \right)^2 + \left(\frac{L - L_s}{w(a_s)} \right)^2 \right) \quad (\text{B.1})$$

with h the kernel bandwidth along the latitude axis, $w(a_s)$ the bandwidth along the longitude axis and $k(\cdot)$ the linear kernel function defined as

$$k(d) = \begin{cases} 1 - d, & \text{if } ||d|| \leq 1. \\ 0, & \text{otherwise.} \end{cases} \quad (\text{B.2})$$

The value of $w(a_s)$ is defined in the $[h, w_{\max}]$ interval according to the following formula:

$$w(a_s) = h + \alpha_s(w_{\max} - h), \quad (\text{B.3})$$

where α_s is the longitudinal expansion factor, defined as

$$\alpha_s = \min\left(\frac{a_s}{a_{\max}}, 1\right). \quad (\text{B.4})$$

and $w_{\max}$ is a parameter defining the maximal kernel bandwidth along the longitude axis to associate with sunspots with area over $a_{\max}$. In our experiments, the parameters h and $w_{\max}$ were set to $h = .1$ [rad] and $w_{\max} = .4$ [rad] (see C for details). We set the value of $a_{\max}$ to 200 [μHem], which is close to the 75th percentile value of the area of a C class sunspot group as surveyed by the ROB.

As a result, during the shifting step, the latitude coordinate is updated according to the following formula

$$l \leftarrow \frac{\sum_{s \in N(l, L)} a_s * l_s}{\sum_{s \in N(l, L)} a_s} \quad (\text{B.5})$$

where $N(l, L)$ is the set of samples lying in the elliptic neighborhood where the kernel derivative is non zero. Given Eq.(B.2), this corresponds to locations (l_s, L_s) for which $\left(\frac{l-l_s}{h}\right)^2 + \left(\frac{L-L_s}{w(a_s)}\right)^2 \leq 1$, where the derivative of the kernel function is equal to 1.

The longitude coordinate is updated in an analogous fashion. This process converges in only a few iterations.

C

Optimal Clustering parameters grid search

The accuracy of the proposed clustering method using each combination of latitudinal kernel bandwidth and maximal longitude bandwidth is shown

on Figure C.1. We considered the $[.02, .12]$ radian range for the latitudinal bandwidth with a step of $.02$ and the $[.1, .75]$ radian range with a step of $.05$ for maximal longitudinal bandwidth. For each combination, we applied clustering on images in $\mathcal{WL}_C$ and measured the accuracy of the sunspot group recovery. Our method achieves 80% accuracy when the latitudinal and longitudinal kernel bandwidths are set to $.1$ radians and $.4$ radians, respectively.

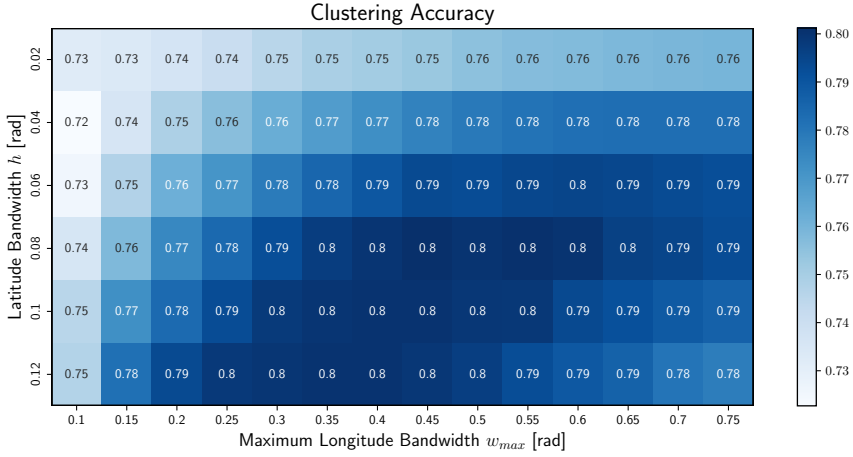


Fig. C.1 Search for optimal kernel longitudinal and latitudinal bandwidths values for the proposed clustering method applied on images in $\mathcal{WL}_C$. Each cell shows the accuracy of the clustering tuned with kernel dimensions of the corresponding row and column.

D Sunspot Group Database entry bounding box computation

The dimension of each target box is computed as a function of the McIntosh class, latitude l and longitude L of the sunspot group center, and solar radius R from the corresponding visual WL observation. Target boxes dimensions along longitudinal axis B_L and latitudinal axis B_l are computed as:

$$B_L = s_Z \cdot |\cos(L)| \quad ; \quad B_l = s_Z \cdot |\cos(l)| \quad (\text{D.1})$$

where s_Z is the length of a side of the sunspot group squared bounding box

as a function of the class of the Z component. The values of s_Z at Sun center are given in Table D.1 for each class of the Z-component in the McIntosh classification system. These values were obtained by expanding the typical size of each Zurich class described in [24] to take into account the tilt and P-angle and are expressed as a function of the radius of the solar disc in the processed image.

	A	B	C	D	E	F	H
Box side	$R/30$	$R/9$	$R/9$	$R/9$	$R/6$	$R/4$	$R/18$

Table D.1 Length of one side of a squared bounding box for sunspot group located at Sun center and following a particular class according to Z-component classification in the McIntosh system. R denotes the sun radius on the corresponding visual WL observation.

E Classification precision, recall, support and F1 score

The precision, recall and F1 score metrics for classification are computed using the following formulae

$$precision = \frac{TP}{TP + FP} \tag{E.1}$$

$$recall = \frac{TP}{TP + FN} \tag{E.2}$$

$$F1 = \frac{2TP}{2TP + FN + FP} = \frac{2 * precision * recall}{precision + recall} \tag{E.3}$$

where TP are true positives, FP are false positives and FN are false negatives.

The weighted average F1 score for classification with multiple classes is calculated by taking the mean of all per-class F1 scores while considering each class's support i.e. the number of occurrences of the class in the set.

F **Classification results details**

Classification metrics are shown for components Z, p, and c of the McIntosh system in Tables F.1, F.2, and F.3, respectively.

Class	Precision	Recall	F1-score	Support
A	0.5	0.8	0.61	64
B	0.57	0.65	0.61	78
C	0.59	0.47	0.53	120
LG	0.86	0.71	0.77	160
H	0.78	0.79	0.79	158
Weighted Avg	0.7	0.68	0.69	580

Table F.1 Ensemble approach classification performance on classification of Z-component. See E for details about computation of columns.

Class	Precision	Recall	F1-score	Support
x	0.74	0.75	0.75	142
r	0.24	0.58	0.34	55
sym	0.73	0.52	0.61	208
asym	0.66	0.58	0.61	175
Weighted Avg	0.66	0.6	0.62	580

Table F.2 Ensemble approach classification performance on p-component. See E for details about computation of columns.

Class	Precision	Recall	F1-score	Support
x	0.77	0.89	0.82	222
o	0.73	0.57	0.64	235
frag	0.66	0.75	0.7	123
Weighted Avg	0.73	0.73	0.72	580

Table F.3 Ensemble approach classification performance on c-component. See E for details about computation of columns.

G Attempts at I2I translation using diffusion-based models

In addition to the methods presented in the main text, we experimented with several diffusion-based models for the WL-to-CalIK image-to-image translation task. All models were trained on the identical dataset used for the I2IwFiLM and Pix2PixCC methods described in Chapter 5. We explored three different diffusion model architectures:

We first implemented *Palette* [188], a conditional diffusion model utilizing a U-Net architecture that operates in pixel space. As shown in Figure G.1, this model generates samples by progressively denoising from pure noise while conditioning on the source modality. Though *Palette* produced CalIK images with fine details, it created cloudy artifacts near the solar limb and failed to capture plages with the same precision as I2IwFiLM. Additionally, *Palette* required significantly more computational resources, memory, and training time.

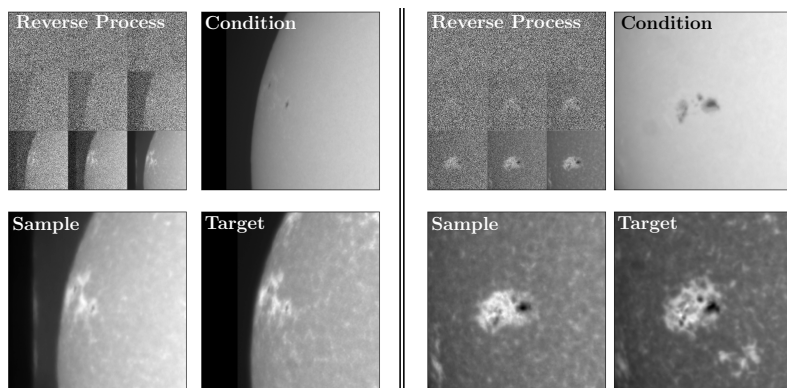


Fig. G.1 Examples generated using *Palette* models showing sharp images with cloudy effect near the limb and weak reconstruction compared to I2IwFiLM.

Next, we explored *Latent Brownian Bridge Diffusion Model* (LBBDM [213]), which operates in latent space using two VQ-VAEs to encode and decode images, see Figure G.2. Unlike *Palette*'s linear noise progression, LBBDM starts from the source modality and follows an increasing-then-decreasing

noise trajectory—creating a brownian bridge between latent spaces. This approach resulted in severe artifacts, including distorted solar limbs and poor correlation between reconstructed and original plages.

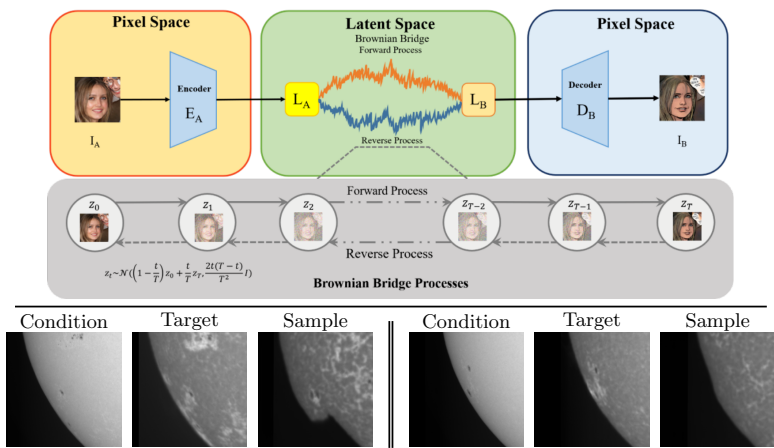


Fig. G.2 (Top) LBBDM architecture (credits: [213]). (Bottom) Examples generated using LBBDM showing distorted limb.

Finally, we explored *Diffusion Transformer* (DiT [214]) models, testing variants with either cross-attention or self-attention for conditioning (with various methods of incorporating conditioning information). As illustrated in Figure G.3, results showed visible patch boundaries—a known behavior of transformers when trained on limited datasets. We also observed weak correlation between reconstructed and original plages, which may show fundamental limitations of the approach. Despite multiple attempts, these issues persisted, likely due to our modest dataset size.

Given these challenges—particularly the limited training data for such large models and the lack of fine details in generated images—, and assuming correct implementations of these approaches, we focused our efforts on GANs and encoder-decoder architectures like I2IwFiLM, which produced higher quality results with fewer computational demands.

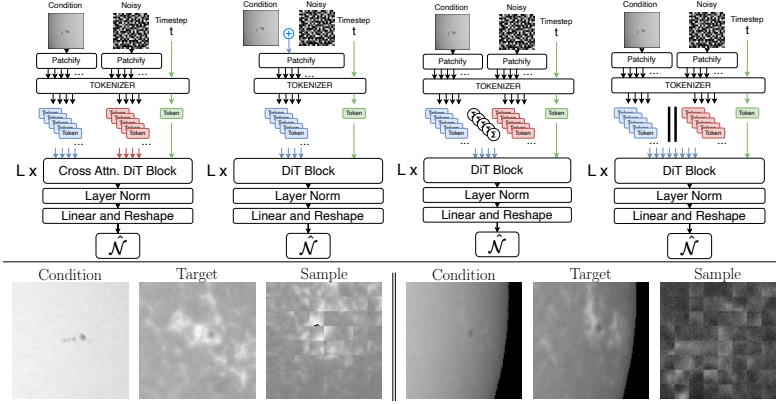


Fig. G.3 (Top) Variants of DiT architectures explored. (Bottom) Examples generated using DiT models showing patch patterns.

H I2IwFiLM versus DiffI2I

In contrast to I2IwFiLM which uses a deterministic CNN to predict guidance vectors $\widehat{\mathcal{E}}^{\mathcal{P}}$ (see Figure 5.5), DiffI2I [211] adopts a diffusion-based process to generate the guidance vectors from random seeds. The training of this diffusion process is depicted in Figure H.1. $CPEN_{\mathcal{P}}$ is a network with similar role as $GVP_{\mathcal{P}}$ in Figure 5.4 and $CPEN_{\mathcal{W}}$ learns an intermediate embedding space that influences the generation of guidance vectors using a diffusion model DM . During this training stage, the weights of both DM and $CPEN_{\mathcal{W}}$ are updated using the L1 loss between the clean $\epsilon^{\mathcal{P}}$ (prior to forward diffusion) and its reconstruction $\widehat{\epsilon}^{\mathcal{P}}$.

Table H.1 compares translation performance of our proposed I2IwFiLM with DiffI2I using generic reconstruction metrics. These results demonstrate that the simplified network architecture of I2IwFiLM, which omits diffusion-based guidance vector generation, performs as effectively as the more complex design of DiffI2I for cross-domain image translation.

Task	Model	RMSE ↓	PSNR [dB] ↑	SSIM ↑	MS-SSIM ↑
WL-to-CaIIK	DiffI2I	14.53 ± 7.20	25.88 ± 4.18	0.80 ± 0.09	0.80 ± 0.11
	I2IwFILM (Ours)	14.44 ± 7.07	25.90 ± 4.12	0.80 ± 0.09	0.80 ± 0.11
EUV-to-Mag	DiffI2I	$13.09 \times 10^{-3} \pm 4.67 \times 10^{-3}$	38.19 ± 3.03	0.97 ± 0.02	0.97 ± 0.02
	I2IwFILM (Ours)	$13.23 \times 10^{-3} \pm 5.06 \times 10^{-3}$	38.10 ± 2.98	0.96 ± 0.03	0.96 ± 0.03

Table H.1 Comparison of I2IwFILM and DiffI2I performance using generic image reconstruction metrics for **whitelight-to-calcium translation** and **EUV-to-Magnetogram translation** tasks. Both models implement FILM feature modulation, with DiffI2I generating guidance vectors $\hat{\mathcal{E}}^{\mathcal{P}}$ using a diffusion process with random seeds, while I2IwFILM uses direct mapping from input image.

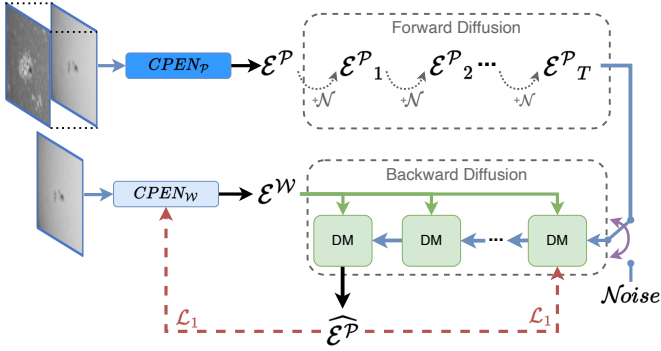


Fig. H.1 Second training stage of DiffI2I model learning to recover $\mathcal{E}^P$ vectors from embedding space $\mathcal{E}^W$. DiffI2I model uses forward and backward diffusion process for recovery. In contrast to our deterministic I2IwFiLM approach, involving noise introduces diversity in $\mathcal{E}^P$ recovery, potentially affecting physical accuracy of $\hat{\mathcal{I}}_C$ prediction.

I Quality metrics

PSNR

The Peak Signal to Noise Ratio (PSNR) is defined as

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\max(I_C)^2}{\text{MSE}(\mathcal{I}_C, \hat{\mathcal{I}}_C)} \right) \quad (\text{I.1})$$

where $\max(I_C)$ is the maximum possible pixel value in I_C , and MSE is the mean square error.

SSIM and MS-SSIM

Structure SIMilarity (SSIM) index measure [83] is a perceptual metric that evaluates image similarity by comparing structural information between an image and a reference, which captures luminance l , contrast c and the spatial dependencies s of pixels. The complete SSIMs metric is expressed

as follows:

$$\begin{aligned} \text{SSIM}(\mathcal{I}_C, \hat{\mathcal{I}}_C) &= l(\mathcal{I}_C, \hat{\mathcal{I}}_C) \cdot c(\mathcal{I}_C, \hat{\mathcal{I}}_C) \cdot s(\mathcal{I}_C, \hat{\mathcal{I}}_C) \\ &= \frac{(2\mu_{\mathcal{I}_C}\mu_{\hat{\mathcal{I}}_C} + k_1L)(2\sigma_{\hat{\mathcal{I}}_C\mathcal{I}_C} + k_2L)}{(\mu_{\mathcal{I}_C}^2 + \mu_{\hat{\mathcal{I}}_C}^2 + k_1L)(\sigma_{\mathcal{I}_C}^2 + \sigma_{\hat{\mathcal{I}}_C}^2 + k_2L)} \end{aligned} \quad (\text{I.2})$$

where $\mu_{\mathcal{I}_C}, \sigma_{\mathcal{I}_C}$ (resp. $\mu_{\hat{\mathcal{I}}_C}, \sigma_{\hat{\mathcal{I}}_C}$) are the mean and standard deviation of the target (resp. generated) calcium image, $\sigma_{\hat{\mathcal{I}}_C\mathcal{I}_C}$ denotes the covariance between the covariance between $\hat{\mathcal{I}}_C$ and $\mathcal{I}_C$. Finally, L denotes the dynamic range of the pixel values and k_1, k_2 are constants ensuring non-zero denominators.

An extension of SSIM metric, named multi-scale SSIM (MS-SSIM [215]), was introduced to enhance its quality by taking the multi-scale nature of the human vision into account in its computation. Both reference and generated images are downsampled to lower resolution such that the luminance, contrast and structure elements of the SSIM can be computed at every scale level. The final MS-SSIM is obtained by multiplying the luminance comparison of the lowest downsampling level with the cumulated product of the contrast and structure comparisons of every down sampling level. Formally, the MS-SSIM is computed as follows:

$$\text{MS-SSIM}(\mathcal{I}_C, \hat{\mathcal{I}}_C) = l_m(\mathcal{I}_C, \hat{\mathcal{I}}_C) \cdot \prod_{j=1}^m [c_j(\mathcal{I}_C, \hat{\mathcal{I}}_C) \cdot s_j(\mathcal{I}_C, \hat{\mathcal{I}}_C)] \quad (\text{I.3})$$

where m denotes the maximum downsampling level and c_j, s_j, l_j denote the contrast, luminance and structure comparisons of the inputs scaled down to the j th level.

J

Weber contrast

The Weber contrast (C) is defined as the ratio of luminance difference between foreground and background over the background luminance. In this work, the quiet Sun is considered as background, while the foreground is defined as the combination of plage in active regions and extended net-

work in CaIIK images and faculae in WL images. Weber contrast (C) is expressed as.

$$C = \frac{\mu_{AR} - \mu_{bg}}{\mu_{bg}} \quad (J.1)$$

where μ_{AR} is the average luminance of pixels belonging to an active region, i.e. plage or extended network in Ca II K and faculae in WL, and μ_{bg} is the average luminance of quiet Sun pixels.

Bibliography

- [1] M. Y. Marov, *The Sun and Heliosphere*, pp. 157–175. New York, NY: Springer New York, 2015.
- [2] “Lesson explainer: The formation of the solar system | nagwa.” [Online; accessed 2024-11-13].
- [3] J. Christensen-Dalsgaard, “Helioseismology,” *Rev. Mod. Phys.*, vol. 74, pp. 1073–1129, Nov 2002.
- [4] H. A. Bethe, “Energy Production in Stars,” *Physical Review*, vol. 55, pp. 434–456, Mar. 1939.
- [5] D. D. Clayton, *Principles of stellar evolution and nucleosynthesis*. 1968.
- [6] M. S. Miesch, “Large-Scale Dynamics of the Convection Zone and Tachocline,” *Living Reviews in Solar Physics*, vol. 2, p. 1, Dec. 2005.
- [7] A. S. Brun and J. Toomre, “Turbulent Convection under the Influence of Rotation: Sustaining a Strong Differential Rotation,” *ApJ*, vol. 570, pp. 865–885, May 2002.
- [8] R. Howard and J. Harvey, “Spectroscopic Determinations of Solar Rotation,” *Sol. Phys.*, vol. 12, pp. 23–51, Apr. 1970.
- [9] M. Stix, “The Sun’s Differential Rotation,” *Reviews in Modern Astronomy*, vol. 2, pp. 248–266, Jan. 1989.
- [10] B. K. Jha, A. Priyadarshi, S. Mandal, S. Chatterjee, and D. Banerjee, “Measurements of Solar Differential Rotation Using the Century Long Kodaikanal Sunspot Data,” *Sol. Phys.*, vol. 296, p. 25, Jan. 2021.

- [11] J. M. Vaquero and M. Vázquez, *The Sun Recorded Through History: Scientific Data Extracted from Historical Documents*, vol. 361. 2009.
- [12] J. Casanovas, "Early Observations of Sunspots: Scheiner and Galileo," in *1st Advances in Solar Physics Euroconference. Advances in Physics of Sunspots* (B. Schmieder, J. C. del Toro Iniesta, and M. Vazquez, eds.), vol. 118 of *Astronomical Society of the Pacific Conference Series*, p. 3, Jan. 1997.
- [13] H. Schwabe, "Sonnenbeobachtungen im Jahre 1843. Von Herrn Hofrath Schwabe in Dessau," *Astronomische Nachrichten*, vol. 21, p. 233, Feb. 1844.
- [14] G. E. Hale, "On the Probable Existence of a Magnetic Field in Sun-Spots," *ApJ*, vol. 28, p. 315, Nov. 1908.
- [15] E. N. Parker, "Hydromagnetic Dynamo Models.," *ApJ*, vol. 122, p. 293, Sept. 1955.
- [16] M. Steenbeck, F. Krause, and K. H. Rädler, "Berechnung der mittleren LORENTZ-Feldstärke für ein elektrisch leitendes Medium in turbulenter, durch CORIOLIS-Kräfte beeinflusster Bewegung," *Zeitschrift Naturforschung Teil A*, vol. 21, p. 369, Apr. 1966.
- [17] H. W. Babcock, "The Topology of the Sun's Magnetic Field and the 22-Year Cycle.," *ApJ*, vol. 133, p. 572, Mar. 1961.
- [18] R. B. Leighton, "A Magneto-Kinematic Model of the Solar Cycle," *ApJ*, vol. 156, p. 1, Apr. 1969.
- [19] R. C. Carrington, *Observations of the spots on the Sun: from November 9, 1853, to March 24, 1861, made at Redhill*. 1863.
- [20] E. W. Maunder, "Note on the Distribution of Sun-spots in Helio-graphic Latitude, 1874-1902," *MNRAS*, vol. 64, pp. 747-761, June 1904.
- [21] D. H. Hathaway, "The Solar Cycle," *Living Reviews in Solar Physics*, vol. 12, p. 4, Dec. 2015.
- [22] SILSO World Data Center, "The International Sunspot Number," *International Sunspot Number Monthly Bulletin and online catalogue*, 2002-2025.

- [23] S. Sanchez, A. Fournier, and J. Aubert, "The Predictability of Advection-dominated Flux-transport Solar Dynamo Models," *ApJ*, vol. 781, p. 8, Jan. 2014.
- [24] P. S. McIntosh, "The Classification of Sunspot Groups," *Sol. Phys.*, vol. 125, pp. 251–267, Sept. 1990.
- [25] G. E. Hale, F. Ellerman, S. B. Nicholson, and A. H. Joy, "The magnetic polarity of sun-spots," *Astrophysical Journal*, vol. 49, p. 153, vol. 49, p. 153, 1919.
- [26] H. Künzel, "Zur Klassifikation von Sonnenfleckengruppen," *Astronomische Nachrichten*, vol. 288, p. 177, Dec. 1965.
- [27] H. Künzel, "Die Flare-Häufigkeit in Fleckengruppen unterschiedlicher Klasse und magnetischer Struktur," *Astronomische Nachrichten*, vol. 285, p. 271, Aug. 1960.
- [28] I. Sammis, F. Tang, and H. Zirin, "The Dependence of Large Flare Occurrence on the Magnetic Structure of Sunspots," *ApJ*, vol. 540, pp. 583–587, Sept. 2000.
- [29] D. C. Stenning, T. C. M. Lee, D. A. van Dyk, V. Kashyap, J. Sandell, and C. A. Young, "Morphological feature extraction for statistical learning with applications to solar image data," *Statistical Analysis and Data Mining: An ASA Data Science Journal*, vol. 6, no. 4, pp. 329–345, 2013.
- [30] P. Zeeman, "The Effect of Magnetisation on the Nature of Light Emitted by a Substance," *Nature*, vol. 55, p. 347, Feb. 1897.
- [31] J. M. Borrero and K. Ichimoto, "Magnetic Structure of Sunspots," *Living Reviews in Solar Physics*, vol. 8, p. 4, Sept. 2011.
- [32] J. C. del Toro Iniesta and B. Ruiz Cobo, "Inversion of the radiative transfer equation for polarized light," *Living Reviews in Solar Physics*, vol. 13, p. 4, Nov. 2016.
- [33] J. W. Harvey, F. Hill, R. P. Hubbard, J. R. Kennedy, J. W. Leibacher, J. A. Pinter, P. A. Gilman, R. W. Noyes, A. M. Title, J. Toomre, R. K. Ulrich, A. Bhatnagar, J. A. Kennewell, W. Marquette, J. Patrón, O. Saá, and E. Yasukawa, "The global oscillation network group (GONG) project," *Science*, vol. 272, no. 5266, pp. 1284–1286, 1996.

- [34] S. Bechet and F. Clette, "USET images / l1-centered." <https://doi.org/10.24414/nc7j-b391>, april 2002.
- [35] F. Clette, D. Berghmans, P. Vanlommel, R. A. M. Van der Linden, A. Koeckelenbergh, and L. Wauters, "From the Wolf number to the International Sunspot Index: 25 years of SIDC," *Advances in Space Research*, vol. 40, pp. 919–928, Jan. 2007.
- [36] F. Clette and L. Lefèvre, "The New Sunspot Number: Assembling All Corrections," *Sol. Phys.*, vol. 291, pp. 2629–2651, Nov. 2016.
- [37] A. J. P. Aparicio, L. Lefèvre, M. C. Gallego, J. M. Vaquero, F. Clette, N. Bravo-Paredes, P. Galaviz, and M. L. Bautista, "A Sunspot Catalog for the Period 1952 - 1986 from Observations Made at the Madrid Astronomical Observatory," *Sol. Phys.*, vol. 293, p. 164, Dec. 2018.
- [38] V. M. S. Carrasco, J. V. Álvarez, and J. M. Vaquero, "Sunspots During the Maunder Minimum from Machina Coelestis by Hevelius," *Sol. Phys.*, vol. 290, pp. 2719–2732, Oct. 2015.
- [39] W. D. Pesnell, B. J. Thompson, and P. C. Chamberlin, "The Solar Dynamics Observatory (SDO)," *Sol. Phys.*, vol. 275, pp. 3–15, Jan. 2012.
- [40] V. Domingo, B. Fleck, and A. I. Poland, "The SOHO Mission: an Overview," *Sol. Phys.*, vol. 162, pp. 1–37, Dec. 1995.
- [41] B. De Pontieu, A. M. Title, J. R. Lemen, G. D. Kushner, D. J. Akin, B. Allard, T. Berger, P. Boerner, M. Cheung, C. Chou, J. F. Drake, D. W. Duncan, S. Freeland, G. F. Heyman, C. Hoffman, N. E. Hurlburt, R. W. Lindgren, D. Mathur, R. Rehse, D. Sabolish, R. Seguin, C. J. Schrijver, T. D. Tarbell, J. P. Wülser, C. J. Wolfson, C. Yanari, J. Mudge, N. Nguyen-Phuc, R. Timmons, R. van Bezooijen, I. Wengro, R. Brookner, G. Butcher, B. Dougherty, J. Eder, V. Knagenhjelm, S. Larsen, D. Mansir, L. Phan, P. Boyle, P. N. Cheimets, E. E. DeLuca, L. Golub, R. Gates, E. Hertz, S. McKillop, S. Park, T. Perry, W. A. Podgorski, K. Reeves, S. Saar, P. Testa, H. Tian, M. Weber, C. Dunn, S. Eccles, S. A. Jaeggli, C. C. Kankelborg, K. Mashburn, N. Pust, L. Springer, R. Carvalho, L. Kleint, J. Marmie, E. Mazmanian, T. M. D. Pereira, S. Sawyer, J. Strong, S. P. Worden, M. Carlsson, V. H. Hansteen, J. Leenaarts, M. Wiesmann, J. Aloise, K. C. Chu, R. I. Bush, P. H. Scherrer, P. Brekke, J. Martinez-Sykora, B. W. Lites, S. W. McIntosh, H. Uitenbroek, T. J. Okamoto, M. A. Gummin, G. Auker,

- P. Jerram, P. Pool, and N. Waltham, "The Interface Region Imaging Spectrograph (IRIS)," *Sol. Phys.*, vol. 289, pp. 2733–2779, July 2014.
- [42] M. L. Kaiser, T. A. Kucera, J. M. Davila, O. C. St. Cyr, M. Guhathakurta, and E. Christian, "The STEREO Mission: An Introduction," *Space Sci. Rev.*, vol. 136, pp. 5–16, Apr. 2008.
- [43] J. R. Lemen, A. M. Title, 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, N. L. Katz, G. D. Kushner, M. Levay, R. W. Lindgren, D. P. Mathur, E. L. McFeaters, S. Mitchell, R. A. Rehse, C. J. Schrijver, L. A. Springer, R. A. Stern, T. D. Tarbell, J.-P. Wuelser, C. J. Wolfson, C. Yanari, J. A. Bookbinder, P. N. Cheimets, D. Caldwell, E. E. Deluca, R. Gates, L. Golub, S. Park, W. A. Podgorski, R. I. Bush, P. H. Scherrer, M. A. Gummin, P. Smith, G. Auker, P. Jerram, P. Pool, R. Soufli, D. L. Windt, S. Beardsley, M. Clapp, J. Lang, and N. Waltham, "The Atmospheric Imaging Assembly (AIA) on the Solar Dynamics Observatory (SDO)," *Sol. Phys.*, vol. 275, pp. 17–40, Jan. 2012.
- [44] J. Schou, P. H. Scherrer, R. I. Bush, R. Wachter, S. Couvidat, M. C. Rabello-Soares, R. S. Bogart, J. T. Hoeksema, Y. Liu, T. L. Duvall, D. J. Akin, B. A. Allard, J. W. Miles, R. Rairden, R. A. Shine, T. D. Tarbell, A. M. Title, C. J. Wolfson, D. F. Elmore, A. A. Norton, and S. Tomczyk, "Design and Ground Calibration of the Helioseismic and Magnetic Imager (HMI) Instrument on the Solar Dynamics Observatory (SDO)," *Sol. Phys.*, vol. 275, pp. 229–259, Jan. 2012.
- [45] T. N. Woods, F. G. Eparvier, R. Hock, A. R. Jones, D. Woodraska, D. Judge, L. Didkovsky, J. Lean, J. Mariska, H. Warren, D. McMullin, P. Chamberlin, G. Berthiaume, S. Bailey, T. Fuller-Rowell, J. Sojka, W. K. Tobiska, and R. Viereck, "Extreme Ultraviolet Variability Experiment (EVE) on the Solar Dynamics Observatory (SDO): Overview of Science Objectives, Instrument Design, Data Products, and Model Developments," *Sol. Phys.*, vol. 275, pp. 115–143, Jan. 2012.
- [46] The SunPy Community, W. T. Barnes, M. G. Bobra, S. D. Christe, N. Freij, L. A. Hayes, J. Ireland, S. Mumford, D. Perez-Suarez, D. F. Ryan, A. Y. Shih, P. Chanda, K. Glogowski, R. Hewett, V. K. Hughitt, A. Hill, K. Hiware, A. Inglis, M. S. F. Kirk, S. Konge, J. P. Mason, S. A.

- Maloney, S. A. Murray, A. Panda, J. Park, T. M. D. Pereira, K. Reardon, S. Savage, B. M. Sipőcz, D. Stansby, Y. Jain, G. Taylor, T. Yadav, Rajul, and T. K. Dang, "The sunpy project: Open source development and status of the version 1.0 core package," *The Astrophysical Journal*, vol. 890, pp. 68–, 2020.
- [47] R. Galvez, D. F. Fouhey, M. Jin, A. Szenicer, A. Muñoz-Jaramillo, M. C. M. Cheung, P. J. Wright, M. G. Bobra, Y. Liu, J. Mason, and R. Thomas, "A Machine-learning Data Set Prepared from the NASA Solar Dynamics Observatory Mission," *ApJS*, vol. 242, p. 7, May 2019.
- [48] E. Alpaydin, *Multilayer Perceptrons*, pp. 267–316. 2014.
- [49] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," *International Statistical Review / Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.
- [50] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.
- [51] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Transactions on Information Theory*, vol. 21, no. 1, pp. 32–40, 1975.
- [52] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.
- [53] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, Sept. 1995.
- [54] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.," *Psychological review*, vol. 65, no. 6, p. 386, 1958.
- [55] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (J. Fürnkranz and T. Joachims, eds.), (Haifa, Israel), pp. 807–814, Omnipress, June 2010.

- [56] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, Oct. 1986.
- [57] K. Hornik, M. Stinchcombe, and H. White, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks," *Neural networks*, vol. 3, no. 5, pp. 551–560, 1990.
- [58] B. C. Csáji *et al.*, "Approximation with artificial neural networks," *Faculty of Sciences, Eötvös Loránd University, Hungary*, vol. 24, no. 48, p. 7, 2001.
- [59] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.
- [60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2015.
- [61] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. 61, pp. 2121–2159, 2011.
- [62] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [63] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net*, 2021.
- [64] R. Gonzalez and R. Woods, *Digital Image Processing Global Edition*. Pearson Deutschland, 2017.
- [65] 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. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

- [66] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," pp. 1–14, Computational and Biological Learning Society, 2015.
- [67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
- [68] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.
- [69] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), (Cham), pp. 234–241, Springer International Publishing, 2015.
- [70] E. Perez, F. Strub, H. Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 09 2017.
- [71] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
- [72] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.
- [73] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, p. 60, July 2019.
- [74] S. Dieleman, K. W. Willett, and J. Dambre, "Rotation-invariant convolutional neural networks for galaxy morphology prediction," *MNRAS*, vol. 450, pp. 1441–1459, June 2015.
- [75] B. Settles, "Active learning literature survey," 2009.

- [76] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net, 2018.
- [77] N. Sayez and C. D. Vleeschouwer, "Accelerating the creation of instance segmentation training sets through bounding box annotation," in *26th International Conference on Pattern Recognition, ICPR 2022, Montreal, QC, Canada, August 21-25, 2022*, pp. 252–258, IEEE, 2022.
- [78] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [79] M. Bertero, P. Boccacci, and C. De Mol, *Introduction to inverse problems in imaging*. CRC press, 2021.
- [80] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.
- [81] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365–3387, 2021.
- [82] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4470–4479, 2019.
- [83] 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.
- [84] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.
- [85] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision – ECCV 2016*

- (B. Leibe, J. Matas, N. Sebe, and M. Welling, eds.), (Cham), pp. 694–711, Springer International Publishing, 2016.
- [86] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.
- [87] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds.), vol. 27, Curran Associates, Inc., 2014.
- [88] G. E. Hinton and R. R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Networks,” *Science*, vol. 313, pp. 504–507, July 2006.
- [89] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” *Journal of Machine Learning Research*, vol. 11, no. 110, pp. 3371–3408, 2010.
- [90] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, “Improved techniques for training gans,” in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.
- [91] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.
- [92] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
- [93] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

- [94] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016.
- [95] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, (Los Alamitos, CA, USA), pp. 1510–1519, IEEE Computer Society, Oct 2017.
- [96] A. Ben-Cohen, E. Klang, S. P. Raskin, S. Soffer, S. Ben-Haim, E. Konen, M. M. Amitai, and H. Greenspan, "Cross-modality synthesis from ct to pet using fcn and gan networks for improved automated lesion detection," *Engineering Applications of Artificial Intelligence*, vol. 78, pp. 186–194, 2019.
- [97] B. Lawrance, H. Lee, E. Park, I.-H. Cho, Y.-J. Moon, J.-Y. Lee, A. Shanmugaraju, and S. Rahman, "Generation of Solar Coronal White-light Images from SDO/AIA EUV Images by Deep Learning," *ApJ*, vol. 937, p. 111, Oct. 2022.
- [98] N. G. Nair and V. M. Patel, "T2v-ddpm: Thermal to visible face translation using denoising diffusion probabilistic models," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–7, 2023.
- [99] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 6840–6851, Curran Associates, Inc., 2020.
- [100] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 8780–8794, Curran Associates, Inc., 2021.
- [101] R. Fergus, M. D. Zeiler, G. W. Taylor, and D. Krishnan, "Deconvolutional networks," in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2528–2535, IEEE Computer Society, 2010.
- [102] A. van den Oord, O. Vinyals, and k. kavukcuoglu, "Neural discrete representation learning," in *Advances in Neural Information Processing*

- Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
- [103] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *ArXiv*, vol. abs/1411.1784, 2014.
- [104] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 5967–5976, IEEE Computer Society, July 2017.
- [105] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 214–223, PMLR, 06–11 Aug 2017.
- [106] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of wasserstein gans,” in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
- [107] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical Text-Conditional Image Generation with CLIP Latents,” *arXiv e-prints*, p. arXiv:2204.06125, Apr. 2022.
- [108] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 10674–10685, IEEE Computer Society, June 2022.
- [109] T. Brooks, A. Holynski, and A. A. Efros, “InstructPix2Pix: Learning to Follow Image Editing Instructions,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 18392–18402, IEEE Computer Society, June 2023.
- [110] W. H. L. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. Da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M. J. Cardoso, “Brain imaging generation with latent diffusion models,” in *Deep Generative Models*

- (A. Mukhopadhyay, I. Oksuz, S. Engelhardt, D. Zhu, and Y. Yuan, eds.), (Cham), pp. 117–126, Springer Nature Switzerland, 2022.
- [111] G. Müller-Franzes, J. M. Niehues, F. Khader, S. T. Arasteh, C. Haarbarger, C. Kuhl, T. Wang, T. Han, T. Nolte, S. Nebelung, J. N. Kather, and D. Truhn, “A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis,” *Scientific Reports*, vol. 13, p. 12098, July 2023.
 - [112] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, and D. Baker, “De novo design of protein structure and function with RFdiffusion,” *Nature*, vol. 620, pp. 1089–1100, Aug. 2023.
 - [113] G. Corso, H. Stärk, B. Jing, R. Barzilay, and T. S. Jaakkola, “Diffdock: Diffusion steps, twists, and turns for molecular docking,” in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, OpenReview.net, 2023.
 - [114] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, “AudioLDM: Text-to-audio generation with latent diffusion models,” in *Proceedings of the 40th International Conference on Machine Learning* (A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, eds.), vol. 202 of *Proceedings of Machine Learning Research*, pp. 21450–21474, PMLR, 23–29 Jul 2023.
 - [115] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
 - [116] T. Salimans and J. Ho, “Progressive distillation for fast sampling of diffusion models,” in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net, 2022.
 - [117] C. Meng, R. Rombach, R. Gao, D. Kingma, S. Ermon, J. Ho, and T. Salimans, “On distillation of guided diffusion models,” in *Proceedings of*

- the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14297–14306, June 2023.
- [118] Y.-T. Hsiao, S. Khodadadeh, K. Duarte, W.-A. Lin, H. Qu, M. Kwon, and R. Kalarot, “Plug-and-play diffusion distillation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13743–13752, June 2024.
 - [119] T. Yin, M. Gharbi, R. Zhang, E. Shechtman, F. Durand, W. T. Freeman, and T. Park, “One-step diffusion with distribution matching distillation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623, June 2024.
 - [120] T. Yin, M. Gharbi, T. Park, R. Zhang, E. Shechtman, F. Durand, and W. T. Freeman, “Improved distribution matching distillation for fast image synthesis,” in *Advances in Neural Information Processing Systems* (A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, eds.), vol. 37, pp. 47455–47487, Curran Associates, Inc., 2024.
 - [121] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, “A kernel two-sample test,” *Journal of Machine Learning Research*, vol. 13, no. 25, pp. 723–773, 2012.
 - [122] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton, “Demystifying MMD gans,” in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net, 2018.
 - [123] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
 - [124] S. Zhou, M. Gordon, R. Krishna, A. Narcomey, L. F. Fei-Fei, and M. Bernstein, “Hype: A benchmark for human eye perceptual evaluation of generative models,” in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

- [125] A. L. Cortie, "On the Types of Sun-Spot Disturbances," *ApJ*, vol. 13, p. 260, May 1901.
- [126] K. O. Kiepenheuer, "Solar Activity," in *The Sun* (G. P. Kuiper, ed.), p. 322, 1953.
- [127] P. L. Bornmann and D. Shaw, "Flare Rates and the McIntosh Active Region Classifications," *Sol. Phys.*, vol. 150, pp. 127–146, Mar. 1994.
- [128] P. T. Gallagher, Y. J. Moon, and H. Wang, "Active-Region Monitoring and Flare Forecasting I. Data Processing and First Results," *Sol. Phys.*, vol. 209, pp. 171–183, Sept. 2002.
- [129] D. S. Bloomfield, P. A. Higgins, R. T. J. McAteer, and P. T. Gallagher, "Toward Reliable Benchmarking of Solar Flare Forecasting Methods," *ApJ*, vol. 747, p. L41, Mar. 2012.
- [130] A. E. McCloskey, P. T. Gallagher, and D. S. Bloomfield, "Flare forecasting using the evolution of McIntosh sunspot classifications," *Journal of Space Weather and Space Climate*, vol. 8, p. A34, June 2018.
- [131] L. Lefevre and F. Clette, "Survey and Merging of Sunspot Catalogs," *Sol. Phys.*, vol. 289, pp. 545–561, Feb. 2014.
- [132] T. Baranyi, L. Győri, and A. Ludmány, "On-line Tools for Solar Data Compiled at the Debrecen Observatory and Their Extensions with the Greenwich Sunspot Data," *Sol. Phys.*, vol. 291, pp. 3081–3102, Nov. 2016.
- [133] L. Győri, A. Ludmány, and T. Baranyi, "Comparative analysis of debrecen sunspot catalogues," *MNRAS*, vol. 465, pp. 1259–1273, Feb. 2017.
- [134] A. J. P. Aparicio, V. M. S. Carrasco, M. C. Gallego, and J. M. Vaquero, "A Sunspot Catalog by Rafael Carrasco at the Madrid Astronomical Observatory for the Period 1931 - 1933," *Sol. Phys.*, vol. 297, p. 58, May 2022.
- [135] A. J. P. Aparicio, J. M. Vaquero, V. M. S. Carrasco, and M. C. Gallego, "Sunspot Numbers and Areas from the Madrid Astronomical Observatory (1876 - 1986)," *Sol. Phys.*, vol. 289, pp. 4335–4349, Nov. 2014.

- [136] J. J. Curto, J. G. Solé, M. Genescà, M. J. Blanca, and J. M. Vaquero, "Historical Heliophysical Series of the Ebro Observatory," *Sol. Phys.*, vol. 291, pp. 2587–2607, Nov. 2016.
- [137] V. M. S. Carrasco, J. M. Vaquero, M. C. Gallego, A. Lourenço, T. Barata, and J. M. Fernandes, "Sunspot Catalogue of the Observatory of the University of Coimbra (1929 - 1941)," *Sol. Phys.*, vol. 293, p. 153, Nov. 2018.
- [138] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241, Springer, 2015.
- [139] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
- [140] F. Clette, D. Berghmans, P. Vanlommel, R. A. Van der Linden, A. Koeckelenbergh, and L. Wauters, "From the wolf number to the international sunspot index: 25 years of sidc," *Advances in Space Research*, vol. 40, no. 7, pp. 919–928, 2007.
- [141] T. Pettauer and P. N. Brandt, "On Novel Methods to Determine Areas of Sunspots from Photoheliograms," *Sol. Phys.*, vol. 175, pp. 197–203, Sept. 1997.
- [142] M. Steinegger, J. A. Bonet, M. Vázquez, and A. Jiménez, "On the Intensity Thresholds of the Network and Plage Regions," *Sol. Phys.*, vol. 177, pp. 279–286, Jan. 1998.
- [143] V. V. Zharkova, J. Aboudarham, S. Zharkov, S. S. Ipson, A. K. Benkhalil, and N. Fuller, "Solar Feature Catalogues In EGSO," *Sol. Phys.*, vol. 228, pp. 361–375, May 2005.
- [144] M. Turmon, J. M. Pap, and S. Mukhtar, "Statistical Pattern Recognition for Labeling Solar Active Regions: Application to SOHO/MDI Imagery," *ApJ*, vol. 568, pp. 396–407, Mar. 2002.
- [145] J. Muraközy, "Variations of the Internal Asymmetries of Sunspot Groups during Their Decay," *ApJ*, vol. 925, p. 87, Jan. 2022.

- [146] S. Zharkov, V. Zharkova, S. Ipson, and A. Benkhalil, "Technique for Automated Recognition of Sunspots on Full-Disk Solar Images," *EURASIP Journal on Applied Signal Processing*, vol. 2005, p. 318462, Dec. 2005.
- [147] P. H. Scherrer, R. S. Bogart, R. I. Bush, J. T. Hoeksema, A. G. Kosovichev, J. Schou, W. Rosenberg, L. Springer, T. D. Tarbell, A. Title, C. J. Wolfson, I. Zayer, and MDI Engineering Team, "The Solar Oscillations Investigation - Michelson Doppler Imager," *Sol. Phys.*, vol. 162, pp. 129–188, Dec. 1995.
- [148] J. J. Curto, M. Blanca, and E. Martínez, "Automatic Sunspots Detection on Full-Disk Solar Images using Mathematical Morphology," *Sol. Phys.*, vol. 250, pp. 411–429, Aug. 2008.
- [149] F. Watson, L. Fletcher, S. Dalla, and S. Marshall, "Modelling the Longitudinal Asymmetry in Sunspot Emergence: The Role of the Wilson Depression," *Sol. Phys.*, vol. 260, pp. 5–19, Nov. 2009.
- [150] S. Carvalho, S. Gomes, T. Barata, A. Lourenço, and N. Peixinho, "Comparison of automatic methods to detect sunspots in the Coimbra Observatory spectroheliograms," *Astronomy and Computing*, vol. 32, p. 100385, July 2020.
- [151] T. T. Nguyen, C. P. Willis, D. J. Paddon, and H. S. Nguyen, "On learning of sunspot classification," in *Intelligent Information Processing and Web Mining* (M. A. Kłopotek, S. T. Wierzchoń, and K. Trojanowski, eds.), (Berlin, Heidelberg), pp. 59–68, Springer Berlin Heidelberg, 2004.
- [152] T. T. Nguyen, C. P. Willis, D. J. Paddon, and H. S. Nguyen, "A hybrid system for learning sunspot recognition and classification," in *2006 International Conference on Hybrid Information Technology*, vol. 2, pp. 257–264, 2006.
- [153] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96*, p. 226–231, AAAI Press, 1996.
- [154] T. Colak and R. Qahwaji, "Automated McIntosh-Based Classification of Sunspot Groups Using MDI Images," *Sol. Phys.*, vol. 248, pp. 277–296, Apr. 2008.

- [155] M. A. Abd, S. F. Majed, and V. Zharkova, "Automated classification of sunspot groups with support vector machines," in *Technological Developments in Networking, Education and Automation* (K. Elleithy, T. Sobh, M. Iskander, V. Kapila, M. A. Karim, and A. Mahmood, eds.), (Dordrecht), pp. 321–325, Springer Netherlands, 2010.
- [156] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, p. 436, 521 2015.
- [157] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. <http://www.deeplearningbook.org>.
- [158] 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. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.
- [159] I. Knyazeva, A. Rybintsev, T. Ohinko, and N. Makarenko, "Deep-Learning Approach for McIntosh-Based Classification Of Solar Active Regions Using HMI and MDI Images," in *Advances in Neural Computation, Machine Learning, and Cognitive Research III*, pp. 239–245, Jan. 2020.
- [160] P. H. Scherrer, J. Schou, R. I. Bush, A. G. Kosovichev, R. S. Bogart, J. T. Hoeksema, Y. Liu, T. L. Duvall, J. Zhao, A. M. Title, C. J. Schrijver, T. D. Tarbell, and S. Tomczyk, "The Helioseismic and Magnetic Imager (HMI) Investigation for the Solar Dynamics Observatory (SDO)," *Sol. Phys.*, vol. 275, pp. 207–227, Jan. 2012.
- [161] L. Palladino, E. Ntagiou, J. Klug, J. Palacios, and R. Keil, "Sunspot Groups Detection and Classification on SDO/HMI Images using Deep Learning Techniques," in *2022 IEEE Aerospace Conference (AERO)*, pp. 1–10, Mar. 2022. ISSN: 1095-323X.
- [162] "Uccle Solar Equatorial Table." <https://www.sidc.be/uset/>, 1956.
- [163] B. Mampaey, V. Delouille, and R. Vansintjan, "The SOLARNET Virtual Observatory: An Effective Way to Search Solar Datasets and Cross-Search with Solar Events," *Solar Physics*, vol. 300, p. 16, Feb. 2025.

- [164] A. F. Agarap, "Deep Learning using Rectified Linear Units (ReLU)," 2019.
- [165] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, Feb. 2021. Number: 2 Publisher: Nature Publishing Group.
- [166] M. Siam, M. Gamal, M. Abdel-Razek, S. Yogamani, M. Jagersand, and H. Zhang, "A comparative study of real-time semantic segmentation for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.
- [167] T. Sugirtha and M. Sridevi, "Semantic segmentation using modified u-net for autonomous driving," in *2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pp. 1–7, 2022.
- [168] S. Jadon, "A survey of loss functions for semantic segmentation," in *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp. 1–7, IEEE, 2020.
- [169] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019.
- [170] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," 2019.
- [171] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *Multiple Classifier Systems*, Lecture Notes in Computer Science, (Berlin, Heidelberg), pp. 1–15, Springer, 2000.
- [172] E. J. Peake, R. Chevasson, S. Pszczolkowski, D. P. Auer, and C. Arthofer, "Ensemble learning for robust knee cartilage segmentation: data from the osteoarthritis initiative," *BioRxiv*, pp. 2020–09, 2020.
- [173] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of

- Proceedings of Machine Learning Research*, pp. 1321–1330, PMLR, 06–11 Aug 2017.
- [174] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” *Advances in neural information processing systems*, vol. 30, 2017.
 - [175] J. Léger, L. Leyssens, G. Kerckhofs, and C. De Vleeschouwer, “Ensemble learning and test-time augmentation for the segmentation of mineralized cartilage versus bone in high-resolution microCT images,” *Computers in Biology and Medicine*, vol. 148, p. 105932, Sept. 2022.
 - [176] S. Mathieu, R. von Sachs, C. Ritter, V. Delouille, and L. Lefèvre, “Uncertainty Quantification in Sunspot Counts,” *ApJ*, vol. 886, p. 7, Nov. 2019.
 - [177] A. A. Pevtsov, I. Virtanen, K. Mursula, A. Tlatov, and L. Bertello, “Reconstructing solar magnetic fields from historical observations. I. Renormalized Ca K spectroheliograms and pseudo-magnetograms,” *A&A*, vol. 585, p. A40, Jan. 2016.
 - [178] T. Chatzistergos, N. A. Krivova, and K. L. Yeo, “Long-term changes in solar activity and irradiance,” *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 252, p. 106150, Nov. 2023.
 - [179] G. Vanden Broeck, S. Bechet, F. Clette, G. Rauw, K. P. Schröder, and M. Mittag, “Relationship between TIGRE solar S-index and USET Ca II K full disk images,” *A&A*, vol. 689, p. A95, Sept. 2024.
 - [180] O. Cohen, I. Sokolov, I. Roussev, C. Arge, W. Manchester, T. Gombosi, R. Frazin, H. Park, M. Butala, F. Kamalabadi, *et al.*, “A semiempirical magnetohydrodynamical model of the solar wind,” *The Astrophysical Journal*, vol. 654, no. 2, p. L163, 2006.
 - [181] D. Shiota and R. Kataoka, “Magnetohydrodynamic simulation of interplanetary propagation of multiple coronal mass ejections with internal magnetic flux rope (SUSANOO-CME),” *Space Weather*, vol. 14, no. 2, pp. 56–75, 2016.
 - [182] J. Pomoell and S. Poedts, “EUHFORIA: European heliospheric forecasting information asset,” *Journal of Space Weather and Space Climate*, vol. 8, p. A35, June 2018.

- [183] P. H. Scherrer, R. S. Bogart, R. I. Bush, J. T. Hoeksema, A. G. Kosovichev, J. Schou, W. Rosenberg, L. Springer, T. D. Tarbell, A. Title, C. J. Wolfson, I. Zayer, and MDI Engineering Team, "The Solar Oscillations Investigation - Michelson Doppler Imager," *Sol. Phys.*, vol. 162, pp. 129–188, Dec. 1995.
- [184] R. A. Howard, J. D. Moses, A. Vourlidas, J. S. Newmark, D. G. Socker, S. P. Plunkett, C. M. Korendyke, J. W. Cook, A. Hurley, J. M. Davila, W. T. Thompson, O. C. St Cyr, E. Mentzell, K. Mehalick, J. R. Lemen, J. P. Wuelser, D. W. Duncan, T. D. Tarbell, C. J. Wolfson, A. Moore, R. A. Harrison, N. R. Waltham, J. Lang, C. J. Davis, C. J. Eyles, H. Mapson-Menard, G. M. Simnett, J. P. Halain, J. M. Defise, E. Mazy, P. Rochus, R. Mercier, M. F. Ravet, F. Delmotte, F. Auchere, J. P. Delaboudiniere, V. Bothmer, W. Deutsch, D. Wang, N. Rich, S. Cooper, V. Stephens, G. Maahs, R. Baugh, D. McMullin, and T. Carter, "Sun Earth Connection Coronal and Heliospheric Investigation (SECCHI)," *Space Sci. Rev.*, vol. 136, pp. 67–115, Apr. 2008.
- [185] P. Rochus, F. Auchère, D. Berghmans, L. Harra, W. Schmutz, U. Schühle, P. Addison, T. Appourchaux, R. Aznar Cuadrado, D. Baker, J. Barbay, D. Bates, A. BenMoussa, M. Bergmann, C. Beurthe, B. Borgo, K. Bonte, M. Bouzit, L. Bradley, V. Büchel, E. Buchlin, J. Büchner, F. Cabé, L. Cadiergues, M. Chaigneau, B. Chares, C. Choque Cortez, P. Coker, M. Condamin, S. Coumar, W. Curdt, J. Cutler, D. Davies, G. Davison, J. M. Defise, G. Del Zanna, F. Delmotte, V. Delouille, L. Dolla, C. Dumesnil, F. Dürig, R. Enge, S. François, J. J. Fourmond, J. M. Gillis, B. Giordanengo, S. Gisot, L. M. Green, N. Guerreiro, A. Guilbaud, M. Gyo, M. Haber-reiter, A. Hafiz, M. Hailey, J. P. Halain, J. Hansotte, C. Hecquet, K. Heerlein, M. L. Hellin, S. Hemsley, A. Hermans, V. Hervier, J. F. Hochedez, Y. Houbrechts, K. Ihsan, L. Jacques, A. Jérôme, J. Jones, M. Kahle, T. Kennedy, M. Klaproth, M. Kolleck, S. Koller, E. Kotsialos, E. Kraaikamp, P. Langer, A. Lawrenson, J. C. Le Clech', C. Lenaerts, S. Liebecq, D. Linder, D. M. Long, B. Mampaey, D. Markiewicz-Innes, B. Marquet, E. Marsch, S. Matthews, E. Mazy, A. Mazzoli, S. Meining, E. Meltchakov, R. Mercier, S. Meyer, M. Monnecke, F. Monfort, G. Morinaud, F. Moron, L. Mountney, R. Müller, B. Nicula, S. Parenti, H. Peter, D. Pfiffner, A. Philippon, I. Phillips, J. Y. Plessier, E. Pylyser, F. Rabecki, M. F. Ravet-Krill, J. Rebel-

- lato, E. Renotte, L. Rodriguez, S. Roose, J. Rosin, L. Rossi, P. Roth, F. Rouesnel, M. Roullia, A. Rousseau, K. Ruane, J. Scanlan, P. Schlatter, D. B. Seaton, K. Silliman, S. Smit, P. J. Smith, S. K. Solanki, M. Spescha, A. Spencer, K. Stegen, Y. Stockman, N. Szwec, C. Tamiatto, J. Tandy, L. Teriaca, C. Theobald, I. Tychon, L. van Driel-Gesztelyi, C. Verbeeck, J. C. Vial, S. Werner, M. J. West, D. Westwood, T. Wiegmann, G. Willis, B. Winter, A. Zerr, X. Zhang, and A. N. Zhukov, "The Solar Orbiter EU instrument: The Extreme Ultraviolet Imager," *A&A*, vol. 642, p. A8, Oct. 2020.
- [186] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (Los Alamitos, CA, USA), pp. 14347–14356, IEEE Computer Society, Oct. 2021.
- [187] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 10674–10685, IEEE Computer Society, June 2022.
- [188] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 Conference Proceedings, SIGGRAPH '22*, (New York, NY, USA), Association for Computing Machinery, 2022.
- [189] W. Song, Y. Ma, H. Sun, X. Zhao, and G. Lin, "Improving the spatial resolution of solar images using super-resolution diffusion generative adversarial networks," *A&A*, vol. 686, p. A272, June 2024.
- [190] E. Amar and O. Ben-Shahar, "Image Synthesis for Solar Flare Prediction," *ApJS*, vol. 271, p. 29, Mar. 2024.
- [191] F. P. Ramunno, H.-J. Jeong, S. Hackstein, A. Csillaghy, S. Voloshynovskiy, and M. K. Georgoulis, "Magnetogram-to-Magnetogram: Generative Forecasting of Solar Evolution," in *Proceedings of SPAICE2024: The First Joint European Space Agency / IAA Conference on AI in and for Space* (D. Dold, A. Hadjiivanov, and D. Izzo, eds.), pp. 75–80, Oct. 2024.

- [192] T. Kim, E. Park, H. Lee, Y.-J. Moon, S.-H. Bae, D. Lim, S. Jang, L. Kim, I.-H. Cho, M. Choi, and K.-S. Cho, "Solar farside magnetograms from deep learning analysis of STEREO/EUVI data," *Nature Astronomy*, vol. 3, pp. 397–400, Mar. 2019.
- [193] W. Sun, L. Xu, Y. Zhang, D. Zhao, and F. Zhang, "Solar Active Region Magnetogram Generation by Attention Generative Adversarial Networks," *Research in Astronomy and Astrophysics*, vol. 23, p. 025003, Feb. 2023.
- [194] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 8798–8807, IEEE Computer Society, June 2018.
- [195] H.-J. Jeong, Y.-J. Moon, E. Park, and H. Lee, "Solar Coronal Magnetic Field Extrapolation from Synchronic Data with AI-generated Farside," *The Astrophysical Journal Letters*, vol. 903, p. L25, Nov. 2020.
- [196] H.-J. Jeong, Y.-J. Moon, E. Park, H. Lee, and J.-H. Baek, "Improved AI-generated Solar Farside Magnetograms by STEREO and SDO Data Sets and Their Release," *ApJS*, vol. 262, p. 50, Oct. 2022.
- [197] X. Li, S. P. Valliappan, D. Shukhobodskaya, M. D. Butala, L. Rodriguez, J. Magdalenic, and V. Delouille, "A Transfer Learning Method to Generate Synthetic Synoptic Magnetograms," *Space Weather*, vol. 22, p. e2023SW003499, Jan. 2024.
- [198] M. Dannehl, V. Delouille, and V. Barra, "An Experimental Study on EUV-To-Magnetogram Image Translation Using Conditional Generative Adversarial Networks," *Earth and Space Science*, vol. 11, p. e2023EA002974, Apr. 2024.
- [199] J. Son, J. Cha, Y.-J. Moon, H. Lee, E. Park, G. Shin, and H.-J. Jeong, "Generation of He I 1083 nm Images from SDO AIA Images by Deep Learning," *ApJ*, vol. 920, p. 101, Oct. 2021.
- [200] H. Lee, E. Park, and Y.-J. Moon, "Generation of Modern Satellite Data from Galileo Sunspot Drawings in 1612 by Deep Learning," *ApJ*, vol. 907, p. 118, Feb. 2021.

- [201] F. Gao, T. Liu, W. Sun, and L. Xu, "Generating Space-based SDO/HMI-like Solar Magnetograms from Ground-based H α Images by Deep Learning," *ApJS*, vol. 266, p. 19, June 2023.
- [202] G. Shin, Y.-J. Moon, E. Park, H. Jeong, H. Lee, and S.-H. Bae, "Generation of High-resolution Solar Pseudo-magnetograms from Ca II K Images by Deep Learning," *ApJ*, vol. 895, p. L16, May 2020.
- [203] R. Jarolim, A. M. Veronig, W. Pötzi, and T. Podladchikova, "Instrument-To-Instrument translation: Instrumental advances drive restoration of solar observation series via deep learning," *arXiv e-prints*, p. arXiv:2401.08057, Jan. 2024.
- [204] S. Bechet and F. Clette, "USET Sunspot Group catalog." VO resource provided by the KSB-ROB, 2016.
- [205] 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, *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.
- [206] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least Squares Generative Adversarial Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, (Los Alamitos, CA, USA), pp. 2813–2821, IEEE Computer Society, Oct. 2017.
- [207] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2332–2341, 2019.
- [208] A. Asensio Ramos, C. J. Díaz Baso, and O. Kochukhov, "Approximate Bayesian neural Doppler imaging," *A&A*, vol. 658, p. A162, Feb. 2022.
- [209] A. Asensio Ramos, "Solar multiobject multiframe blind deconvolution with a spatially variant convolution neural emulator," *A&A*, vol. 688, p. A88, Aug. 2024.
- [210] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. Van Gool, "DiffIR: Efficient Diffusion Model for Image Restoration," in *2023 IEEE/CVF International Conference on Computer Vision*

- (ICCV), (Los Alamitos, CA, USA), pp. 13049–13059, IEEE Computer Society, Oct. 2023.
- [211] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, R. Timofte, and L. V. Gool, “DiffI2I: Efficient Diffusion Model for Image-to-Image Translation,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 1–16, Nov. 5555.
 - [212] T. Chatzistergos, N. A. Krivova, and I. Ermolli, “Understanding the secular variability of solar irradiance: the potential of Ca II K observations,” *Journal of Space Weather and Space Climate*, vol. 14, p. 9, Mar. 2024.
 - [213] B. Li, K. Xue, B. Liu, and Y.-K. Lai, “Bbdm: Image-to-image translation with brownian bridge diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1952–1961, June 2023.
 - [214] W. Peebles and S. Xie, “Scalable diffusion models with transformers,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.
 - [215] Z. Wang, E. Simoncelli, and A. Bovik, “Multiscale structural similarity for image quality assessment,” in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, pp. 1398–1402 Vol.2, 2003.