Université de Montréal

Fast High-dimensional Posterior Inference with Deep Generative Models: Application to CMB Delensing

par

Mohammad-Hadi Sotoudeh

Département de physique
Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Physique

August 31, 2023

Université de Montréal

Faculté des arts et des sciences

Ce mémoire intitulé

Fast High-dimensional Posterior Inference with Deep Generative Models: Application to CMB Delensing

présenté par

Mohammad-Hadi Sotoudeh

a été évalué par un jury composé des personnes suivantes :

Yashar Hezaveh

(président-rapporteur)

Laurence Perreault-Levasseur

(directeur de recherche)

Patrick Dufour

(membre du jury)

Résumé

Nous vivons à une époque marquée par une abondance de données cosmologiques de haute résolution. Cet afflux de données engendré par les missions d'observation de nouvelle génération au sol et dans l'espace porte le potentiel de remodeler fondamentalement notre compréhension de l'univers et de ses principes physiques sous-jacents. Cependant, la complexité grande des données observées pose des défis aux approches conventionnelles d'analyse de données, soit en raison de coûts de calcul irréalisables, soit en raison des hypothèses simplificatrices utilisées dans ces algorithmes qui deviennent inadéquates dans des contextes haute résolution à faible bruit, conduisant à des résultats sous-optimaux.

En réponse, la communauté scientifique s'est tournée vers des méthodes innovantes d'analyse de données, notamment les techniques d'apprentissage automatique (ML). Les modèles de ML, lorsqu'ils sont bien entraînés, peuvent identifier de manière autonome des correlations significatives dans les données de manière plus efficace et sans hypothèses restrictives inutiles. Bien que les méthodes de ML aient montré des promesses en astrophysique, elles présentent également des problèmes tels que le manque d'interprétabilité, les biais cachés et les estimations d'incertitude non calibrées, ce qui, jusqu'a maintenant, a entrave leur application dans d'importantes découvertes scientifiques. Ce projet s'inscrit dans le cadre de la collaboration "Learning the Universe" (LtU), axée sur la reconstruction des conditions initiales de l'univers, en utilisant une approche de modélisation bayésienne et en exploitant la puissance du ML. L'objectif de ce projet est de développer un cadre pour mener une inférence bayésienne au niveau des pixels dans des problèmes multidimensionnels.

Dans cette thèse, je présente le développement d'un cadre d'apprentissage profond pour un échantillonnage rapide des postérieurs en dimensions élevées. Ce cadre utilise l'architecture "Hierarchical Probabilistic U-Net", qui combine la puissance de l'architecture U-Net dans l'apprentissage de cartes multidimensionnelles avec le rigoureux cadre d'inférence des autoencodeurs variationnels conditionnels. Notre modèle peut quantifier les incertitudes dans ses données d'entraînement et générer des échantillons à partir de la distribution a posteriori des paramètres, pouvant être utilisés pour dériver des estimations d'incertitude pour les paramètres inférés. L'efficacité de notre cadre est démontrée en l'appliquant au

problème de la reconstruction de cartes du fond diffus cosmologique (CMB) pour en retirer de l'effet de lentille gravitationnelle faible. Notre travail constitue un atout essentiel pour effectuer une inférence de vraisemblance implicite en dimensions élevées dans les domaines astrophysiques. Il permet d'exploiter pleinement le potentiel des missions d'observation de nouvelle génération pour améliorer notre compréhension de l'univers et de ses lois physiques fondamentales.

Mots clés: Inférence Bayésienne de Grande Dimension, Échantillonnage Postérieur, Apprentissage Profond, Modèles Génératifs, Cosmologie, Délentillage du CMB

Abstract

We live in an era marked by an abundance of high-resolution cosmological data. This influx of data brought about by next-generation observational missions on the ground and in space, bears the potential of fundamentally reshaping our understanding of the universe and its underlying physical principles. However, the elevated complexity of the observed data poses challenges to conventional data analysis approaches, either due to infeasible computational costs or the simplifying assumptions used in these algorithms that become inadequate in high-resolution, low-noise contexts, leading to suboptimal results.

In response, the scientific community has turned to innovative data analysis methods, including machine learning (ML) techniques. ML models, when well-trained, can autonomously identify meaningful patterns in data more efficiently and without unnecessary restrictive assumptions. Although ML methods have shown promise in astrophysics, they also exhibit issues like lack of interpretability, hidden biases, and uncalibrated uncertainty estimates, which have hindered their application in significant scientific discoveries. This project is defined within the context of the Learning the Universe (LtU) collaboration, focused on reconstructing the initial conditions of the universe, utilizing a Bayesian forward modeling approach and harnessing the power of ML. The goal of this project is to develop a framework for conducting Bayesian inference at the pixel level in high-dimensional problems.

In this thesis, I present the development of a deep learning framework for fast high-dimensional posterior sampling. This framework utilizes the Hierarchical Probabilistic U-Net architecture, which combines the power of the U-Net architecture in learning high-dimensional mappings with the rigorous inference framework of Conditional Variational Autoencoders. Our model can quantify uncertainties in its training data and generate samples from the posterior distribution of parameters, which can be used to derive uncertainty estimates for the inferred parameters. The effectiveness of our framework is demonstrated by applying it to the problem of removing the weak gravitational lensing effect from the CMB. Our work stands as an essential asset to performing high-dimensional implicit likelihood inference in astrophysical domains. It enables utilizing the full potential of next-generation

observational missions to improve our understanding of the universe and its fundamental physical laws.

Keywords: High-dimensional Bayesian Inference, Posterior Sampling, Deep Learning, Generative Models, Cosmology, CMB Delensing

Contents

List of Tables

List of Figures

List of Symbols

Physics

Particles

H	Hydrogen atom
p	Proton
e^-	Electron
γ	Photon

Physical Units

eV	Electronvolt
K	Kelvin
$^\circ$	Degree

$'$	Arcminute
	Physical Constants
k	Boltzmann Constant
G	Gravitational Constant
c	Speed of Light
	Basic Physical and Cosmological Quantities
Γ_T	Thomson Scattering Reaction Rate
n_e	Number Density of Electrons
σ_T	Thomson Scattering Cross Section
M	Mass
b	Closest Approach Distance (aka Impact Parameter)
a	Scale Factor
z	Redshift

H	Hubble Parameter (aka Expansion Rate of the Universe)
Ω_m	Matter Density Parameter
A_s	Amplitude of the Primordial Power Spectrum (aka Scalar Amplitude)
χ	Comoving Distance
χ_{CMB}	Comoving Distance of the Last Scattering Surface
$S_\kappa(\chi)$	Comoving Angular Diameter Distance Function
κ	Curvature of the Universe $(-1\ /\ 0\ /\ +1)$
R	Present-time Radius of Curvature of the Universe
T	Temperature
T_0	Present-time CMB Temperature
$\delta T/T_0$	CMB Temperature Fluctuation

Fourier Space & Power Spectra

$\boldsymbol{k}$	Wavenumber
(ℓ, m)	Sperical Harmonic Indices: (Multipole Moment, Angular Mode)
$Y_{\ell m}$	Laplace's Spherical Harmonic Function
$a_{\ell m}^{TT}$	Spherical Harmonic Coefficient of CMB Temperature Map
C_ℓ^{TT}	Primordial CMB Temperature Power Spectrum
$\mathcal{D}_\ell^{TT}$	Normalized Primordial CMB Temperature Power Spectrum
$C_\ell^{\tilde{T}\tilde{T}}$	Observed (Lensed) CMB Temperature Power Spectrum
$C_\ell^{\psi\psi}$	Projected Lensing Potential Power Spectrum

CMB Lensing Notation

$\boldsymbol{\beta}$	Deflection Angle
$\boldsymbol{\alpha}(\hat{\boldsymbol{n}})$	Observed Deflection Angle at Direction $\hat{\boldsymbol{n}}$

$\Psi(\boldsymbol{r})$	Weyl Potential at Location $\boldsymbol{r}$
$\psi(\hat{\boldsymbol{n}})$	Projected Lensing Potential at Direction $\hat{\boldsymbol{n}}$
$\tilde{T}$	Observed CMB Temperature
R^{ψ}	Half Mean-Squared Deflection

Math & Statistics

Operators

$\dot{\square}$	Time Derivative of $\square$
$\nabla_{\perp}(\cdot)$	Gradient's Perpendicular Component (Physical)
$\nabla_{\hat{\boldsymbol{n}}}(\cdot)$	Gradient's Perpendicular Component (Angular)
$\|\cdot\|_p$	Lp Norm
$\langle\cdot\rangle$ / $\mathbb{E}[\cdot]$ / $\bar{\cdot}$	Expectation / Average

Coverage Probability Test

$\mathcal{R}$	Credible Region
$l_{\mathcal{R}}$ / CL	Credibility Level
$c_{\mathcal{R}}$ / CP	Coverage Probability

Inference Frameworks

Variables

$\boldsymbol{x}$	Data / Observation
$\boldsymbol{y}$	Parameter
$\boldsymbol{z}$	Latent Variable

Bayesian Inference

$p(\boldsymbol{x}\|\boldsymbol{y})$	Likelihood
$p(\boldsymbol{x})$	Evidence
$p(\boldsymbol{y})$	Parameters Prior

$p(\boldsymbol{y}\|\boldsymbol{x})$	Parameters Posterior
	VAE Probabilistic Framework
$p(\boldsymbol{x}\|\boldsymbol{z})$	Data Likelihood
$p(\boldsymbol{x})$	Evidence
$p(\boldsymbol{z})$	Latents Prior
$p(\boldsymbol{z}\|\boldsymbol{x})$	Latents Posterior
$q(\boldsymbol{z}\|\boldsymbol{x})$	Variational Latents Posterior
	HPU-Net Probabilistic Framework
$p(\boldsymbol{y}\|\boldsymbol{z},\boldsymbol{x})$	Conditional Parameters Likelihood
$p(\boldsymbol{y}\|\boldsymbol{x})$	Conditional Evidence
$p(\boldsymbol{z}\|\boldsymbol{x})$	Conditional Latents Prior
$p(\boldsymbol{z}\|\boldsymbol{y},\boldsymbol{x})$	Conditional Latents Posterior
$q(\boldsymbol{z}\|\boldsymbol{y},\boldsymbol{x})$	Conditional Variational Latents Posterior

L	Number of Latent Scales

Machine Learning

Training

$\boldsymbol{\theta}$ / $\boldsymbol{\phi}$	Neural Network Parameters
$\mathcal{D}$	Dataset
$\mathcal{J}$	Objective Function (aka Loss Function)
$\mathcal{L}$	Instance-Level Loss Function
η	Learning Rate
b	Batch Size

Data Naming Conventions

GT	Ground Truth
$\hat{\cdot}$	Prediction
$\tilde{\mu}$	Empirical Mean

$\tilde{\sigma}$	Empirical Standard Deviation
μ	Theoretical Mean
σ	Theoretical Standard Deviation

List of Abbreviations

Physics

ACT	Atacama Cosmology Telescope
CAMB	Code for Anisotropies in the Microwave Background
CMB	Cosmic Microwave Background
COBE	Cosmic Background Explorer
DASI	Degree Angular Scale Interferometer
RMS	Root Mean Square
SED	Spectral Energy Distribution
SPT	South Pole Telescope
[I]SW Effect	[Integrated] Sachs–Wolfe Effect

SZ Effect	Sunyaev-Zel'dovich Effect
WMAP	Wilkinson Microwave Anisotropy Probe

Statistics

ABC	Approximate Bayesian Computation
CDF	Cumulative Density Function
GRF	Gaussian Random Field
ILI	Implicit Likelihood Inference
MCMC	Markov Chain Monte Carlo
PDF	Probability Density Function
SBI	Simulation-Based Inference
TARP	Tests of Accuracy with Random Points

Machine Learning

ConvNet	Convolutional Neural Network
ELBO	Evidence Lower Bound
GAN	Generative Adversarial Network
HPU-Net	Hierarchical Probabilistic U-Net
KL Divergence	Kullback–Leibler Divergence
ML	Machine Learning
MLP	Multilayer Perceptron
MSE	Mean Squared Error
[G]NLL	[Gaussian] Negative Log-Likelihood
[c]VAE	[Conditional] Variational Autoencoder

Acknowledgment

To Laurence, for her unconditional support and gardening this young seedling.

To my parents, Zahra and Reza, for *raising* me with love and their warm support throughout this *complex* journey.

To Yashar, for being such a comprehensive role model.

To my friends at Ciela Institute, my *family* in Montreal.

To Iran, for the desire for freedom and the moment we are all happy.

To everybody and everything that made me, me!

To my teachers, for *intrinsically* motivating a young student to pursue his passion in astronomy.

To Astronomy, for the joy involved in every moment of studying the skies.

Chapter 1

Introduction

We live in the era of unprecedentedly large, high-resolution, and low-noise cosmological datasets. The wealth of data provided by next-generation observational missions such as the James Webb Space Telescope, Euclid, Roman Space Telescope, SPHEREx, Vera Rubin Observatory, Simons Observatory, and CMB-S4 holds the promise of revolutionizing our understanding of the universe and its underlying physics. However, the complexity of the observed data presents challenges in applying traditional data analysis techniques to explore this vast information landscape. First, running conventional algorithms on exponentially larger data is computationally infeasible. Furthermore, such algorithms often rely on simplifying assumptions no longer valid in high-resolution, low-noise regimes, leading to suboptimality and/or biased predictions.

This situation has led the community to explore innovative data analysis approaches, with a high potential lying in machine learning (ML) techniques. If properly trained, these models can automatically extract meaningful patterns, relationships, and features from the data much more efficiently and without relying on restrictive assumptions. In recent years, ML methods have been successfully applied to various areas in astrophysics for classification, detection, accelerating simulations, and statistical inference. While the results look promising, it should be noted that ML methods are prone to lack of interpretability, hidden biases, and rigorous uncertainty estimates, which introduce hurdles to using them for major scientific discoveries. For instance, various ML methods used for Bayesian simulation-based inference often yield overconfident (i.e., smaller-than-expected) uncertainty measures [**33**].

This project is defined within the context of the Learning the Universe (LtU) international collaboration, which aims to reconstruct the initial conditions of the universe. The term "initial conditions" is used to refer to: 1) a handful of cosmological parameters that describe the matter-energy content of the universe, and 2) a large number of parameters that describe the initial spatial distribution of this content shortly after the Big Bang. LtU

tackles this problem using a Bayesian forward modeling approach. Breaking it down further, the Bayesian approach aims to infer the posterior distribution of parameters (here, initial conditions) given some observations. Furthermore, the forward model is a computational model that can generate synthetic observations given a set of initial conditions. Using a dataset of different initial conditions and their corresponding synthetic observations, as well as a well-defined learning objective, one can train an ML model to learn the inverse of the forward model. This model can then be used to infer the initial conditions from actual observations.

Apart from having observational noise that makes parameter estimates uncertain, such complex problems are often underconstrained, meaning that given the available information, there might exist several acceptable answers, or more precisely, a manifold of acceptable answers that are consistent with observations. Hence, the ML model must be able to account for multimodal posterior distributions. This task is arduous when observations and parameters are high-dimensional (e.g., images consisting of many pixels). Throughout this project, we developed a deep learning framework for fast Bayesian inference in high-dimensional problems. We use a deep generative architecture called Hierarchical Probabilistic U-Net (HPU-Net) that combines U-Net architecture with the Variational Autoencoder (VAE) approximate inference framework to perform this task. To demonstrate its efficacy, we apply our method to the problem of removing the weak gravitational lensing effect from the Cosmic Microwave Background (CMB).

The rest of this thesis is structured as follows: Chapter 2 provides theoretical background on CMB and gravitational lensing. The discussion then delves into the topic of CMB lensing and delineates the cosmological scope of the project. The chapter ends by formally defining the problem and outlining the employed framework for inference. Moving forward, Chapter 3 is dedicated to exploring the deep learning aspects of this project. It commences with the basic concepts of neural networks. Then, it covers deep probabilistic models, and VAEs in particular, which will set the stage for introducing the HPU-Net architecture and its underlying statistical framework. The chapter ends with a discussion of methods for evaluating the learned posterior distribution. Subsequently, in Chapter 4, we present the scientific paper that details our contribution and main results. The thesis concludes with a summary of findings and future remarks presented in Chapter 5.

Chapter 2

Gravity's Fingerprint

In this chapter, we delve into fundamental astrophysical concepts of our study and lay its statistical groundwork. The first two sections present the cosmic microwave background (CMB) and gravitational lensing as independent topics. Subsequently, Section 2.3 discusses the weak lensing effect of large scale structure on CMB. Finally, Section 2.4 presents the formal definition of the scientific problem and Bayesian inference approach to tacke this problem.

2.1. Cosmic Microwave Background

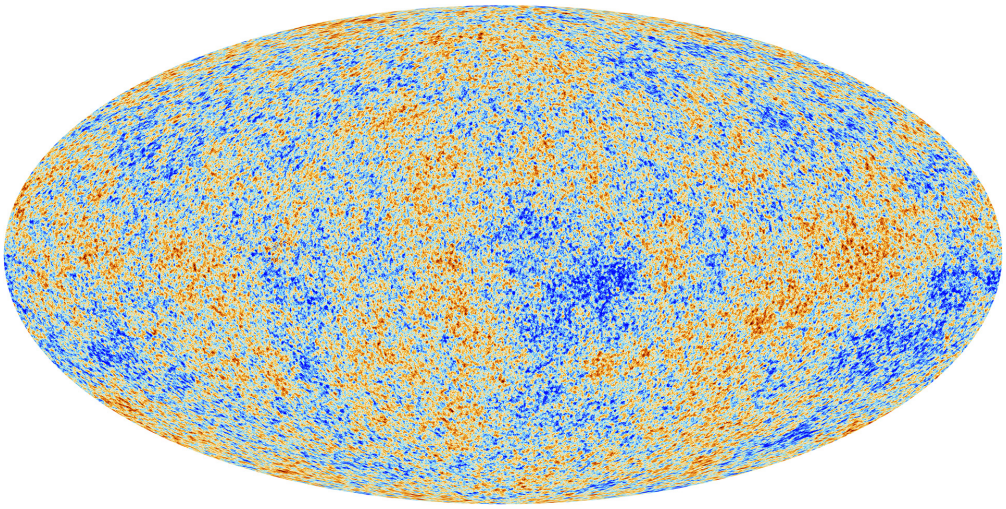

Fig. 2.1. Cosmic Microwave Background (CMB) observed by the Planck satellite. The figure presents a color map illustrating observed CMB temperatures from different directions in the sky, where blue indicates lower temperatures and red indicates higher temperatures. Credit: ESA and the Planck Collaboration.

2.1.1. Discoveries & Frontiers

It was in 1964 that Arno Penzias and Robert Wilson, two radio astronomers working at Bell Telephone Laboratories in New Jersey, were conducting experiments using a sensitive radio antenna to detect faint radio signals. In their experiments, they kept encountering an unexpected noise in their measurements that could not be attributed to any terrestrial or atmospheric sources, and it seemed to persist regardless of the direction of their antenna. While investigating various possible radiation sources, they were unaware that they were on the brink of discovering the relic radiation left over from the early stages of the universe. This radiation called the Cosmic Microwave Background (CMB), became the most compelling evidence for the Big Bang theory and one of the most important discoveries in cosmology.

Since then, numerous ground-based and space-borne telescopes have been dedicated to studying the properties of the CMB. The Cosmic Background Explorer (COBE), launched by NASA in 1989, achieved the first major milestone. COBE made groundbreaking measurements of the CMB's temperature fluctuations, providing the first evidence for the *almost* perfect isotropy of the CMB and confirming the Big Bang's predictions [**61**]. In the late 1990s, the Degree Angular Scale Interferometer (DASI) operated at the Amundsen-Scott South Pole Station. DASI focused on measuring the polarization of the CMB [**53**], revealing essential information about the physical state of early universe and supporting the theory of cosmic inflation. DASI's observations further confirmed the standard cosmological model.

In 2001, NASA launched the Wilkinson Microwave Anisotropy Probe (WMAP) [**7**], delivering even more precise measurements of CMB temperature fluctuations. WMAP's data refined our understanding of the composition, age, and expansion rate of the universe. The Atacama Cosmology Telescope (ACT) began its operations in the Atacama Desert in Chile in 2007. ACT focused on detecting small-scale temperature anisotropies in the CMB [**22**], refining measurements of cosmological parameters, and providing insights into dark energy and dark matter properties. Also, at the South Pole, the South Pole Telescope (SPT) [**12**] has been instrumental in studying the CMB's polarization and small-scale temperature anisotropies, providing valuable constraints on neutrino masses and early universe conditions.

Launched by the European Space Agency in 2009, the Planck satellite [**16**] conducted a comprehensive and precise study of the CMB until 2013, which remains the most precise CMB full-sky survey to date. Figure 2.1 presents the temperature map constructed using Planck data, and Figure 2.2 compares the resolution of COBE, WMAP, and Planck. In subsequent years, various experiments and upgraded versions of existing telescopes, such as QUaD [**36**], ACTPol [**78**], BICEP [**46**], BICEP2 [**65**], and SPTpol [**5**], continued to advance our knowledge of the CMB.

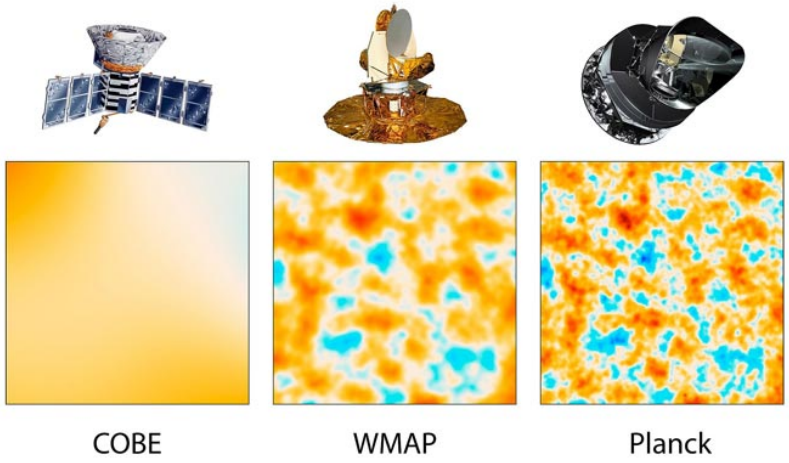

Fig. 2.2. Resolution improvement of the CMB space missions over time. Credit: NASA, JPL-Caltech, and ESA.

Looking ahead, the most important future CMB missions, including the Simons Observatory [**3**] and the CMB-S4 [**2**] experiment, hold great promise for expanding our understanding of cosmology.

2.1.2. Physics of the CMB

Formation of the CMB

Before the radiation of CMB photons, the universe was a hot and dense expanding fluid of photons and baryons. During those times, interactions between photons and baryons were so frequent that photons were unable to travel freely without being absorbed and re-emitted by their surrounding particles. The two main interactions between photons and baryons were

(1) The equilibrium bound-free reactions that kept the universe ionized[1]:

$$\mathrm{H} + \gamma \rightleftharpoons p + e^- \tag{2.1}$$

(2) Thomson scattering of photons by free electrons:

$$\gamma + e^- \rightarrow \gamma + e^- \tag{2.2}$$

As the universe expanded and cooled, two things happened:

(1) First, with the energy scale (kT) of the photon-baryon fluid dropping below the ionization energy of hydrogen atoms (13.6 eV $\equiv$ 3760 K) around redshift $z = 1400$ [2],

[1] In fact, the heavier elements produced during big bang nucleosynthesis also had a contribution, which are neglected in this discussion for simplicity.

[2] A more sophisticated analysis involves using the Saha equation to find the temperature where the ionization fraction of hydrogen dropped below 50%.

photons were no longer energetic enough to ionize atoms. As a result, electrons and protons combined and formed neutral hydrogen atoms. This process, known as *recombination*, significantly reduced the number of free electrons in the photon-baryon fluid.

(2) In parallel, the expansion of the universe kept decreasing the energy density of the photon-baryon fluid; accordingly, the rate of the Thomson scattering reactions ($\Gamma_T = n_e \sigma_T c$) was diminished, leading to a more transparent environment for photons. Eventually, around $z = 1100$, with the reaction rate dropping below the expansion rate of the universe ($H := \dot{a}/a$), the photon-baryon fluid reached a state where the electrons were being diluted by expansion more rapidly than photons could reach them. This led to photons becoming *decoupled* from baryons. Shortly after decoupling, at the moment[3] of *last scattering*, photons became able to travel freely without being scattered.

Prior to the last scattering, the frequent scattering and thermalization processes put photons in thermal equilibrium with baryons. They were constantly absorbed and re-emitted by their surrounding baryons, making the photon-baryon fluid a perfect blackbody. After decoupling, the blackbody spectrum was maintained, but the energy of the photons and hence the characteristic temperature of the blackbody radiation decreased inversely with the scale factor, i.e., $T \propto a^{-1}$. Today, we receive CMB photons perfectly following blackbody radiation with nearly the same temperature $T_0 = 2.73$ K from every direction in the sky[4].

Temperature Fluctuations

Although CMB radiation exhibits remarkable isotropy, it is not entirely isotropic across the sky. Each region in the sky could have slightly lower or higher temperatures than the average, leading to cold and hot patches seen in Figure 2.1. The root mean square of these fluctuations is

$$\left\langle \left(\frac{\delta T}{T_0} \right)^2 \right\rangle^{1/2} \sim 10^{-5} \tag{2.3}$$

These temperature fluctuations (or anisotropies) are believed to be originated by quantum fluctuations during inflation. There exist numerous physical processes that affect the intensity and/or pattern of fluctuations. This can happen during or before the last scattering (primary effects) or throughout the path of CMB photons to us (secondary effects), with

[3]To be precise, since the last scattering of different photons hapenned at slightly different times, last scattering is more of a momentary epoch rather than a singular "moment".

[4]Assuming an entirely radiation-dominated early universe, the isotropy of CMB is surprising, as the extent of the causally connected regions at the time of last scattering was much smaller than the size of the observable universe. Inflationary models suggest an epoch in the early universe with acclerating expansion to overcome this so-called horizon problem.

each process leaving its imprint on the observed CMB. As a result, CMB studies can be done based on two approaches:

- By correcting secondary effects, one can use the recovered primordial CMB to study the pre-recombination physics or measure cosmological parameters by analyzing anisotropy patterns.
- By analyzing primary and secondary effects, one can study the physical processes leading to those effects, as well as the state of the universe at the corresponding times.

Table 2.1 presents well-known examples of primary and secondary anisotropies, focusing on their effect on the temperature map.

Table 2.1. Examples of CMB Primary and Secondary Anisotropies

Type	Name	Physical Process	Effect on CMB
Primary	Initial Fluctuations	Quantum fluctuations in the early universe generating small density perturbations.	Presence of hot and cold spots on CMB.
	Sachs-Wolfe (SW) Effect	Gravitational redshift of photons while climbing in or falling into the baryonic potential wells at the time of the last scattering.	A decrease or increase in observed temperatures.
	Baryon Acoustic Oscillations	Sound waves in the early universe create periodic overdensities in the matter distribution, which imprint corresponding oscillations on the photons.	Characteristic "bumps" in the CMB temperature power spectrum.
Secondary	Integrated Sachs-Wolfe (ISW) Effect	Gravitational redshift of photons while passing through time-varying gravitational potentials along their path.	A decrease or increase in observed temperatures.
	Sunyaev-Zel'dovich (SZ) Effect	Scattering of CMB photons off hot electrons in galaxy clusters, gaining energy in the process.	Brighter (hotter) CMB in the direction of galaxy clusters.
	Gravitational Lensing	Gravitational bending of CMB photons' path by large-scale structures, without altering their energies.	Distortion in the apparent positions of CMB sources.

2.1.3. Characterizing Anisotropies

In order to investigate CMB anisotropies, one needs mathematical tools to describe the observed features and their statistical properties. Several analysis methods exist for this purpose, including correlation function, power spectrum, peak statistics, wavelet analysis, and Minkowski functionals. In this study, we primarily use power spectrum to describe the statistical properties of CMB.

Power Spectrum

The power spectrum quantifies the amplitude of CMB fluctuations at different angular scales. It represents the distribution of power, or the variance, of the fluctuations as a function of the angular scale $\ell = 180°/\theta$ (i.e., multipole moment). Mathematically, the temperature power spectrum C_ℓ^{TT} is defined as the average of the square of the magnitude of the ℓ^{th} coefficient of the spherical harmonic expansion of CMB temperature anisotropies, expressed as:

$$C_\ell^{TT} = \frac{1}{2\ell+1} \sum_m \left| a_{\ell m}^{TT} \right|^2, \tag{2.4}$$

where the summation is performed over all m (angular modes) for a given ℓ, and the factor of $1/(2\ell+1)$ normalizes the power spectrum. Each $a_{\ell m}^{TT}$ is a spherical harmonic coefficient that describes the decomposition of the CMB temperature map into spherical harmonics:

$$a_{\ell m}^{TT} = \iint Y_{\ell m}(\theta, \phi) \, \frac{\delta T}{T_0}(\theta, \phi) \, \sin\theta \, d\theta \, d\phi, \tag{2.5}$$

where $Y_{\ell m}$ is the Laplace's Spherical Harmonic at index (ℓ, m). The power spectrum is a key quantity in cosmology and is often used to constrain models, estimate cosmological parameters, and compare theoretical predictions with observational data. It provides essential information about the statistical properties and angular distribution of CMB temperature anisotropies, revealing important details about the early universe and its evolution. Figure 2.3 shows the temperature power spectrum based on Planck's data.

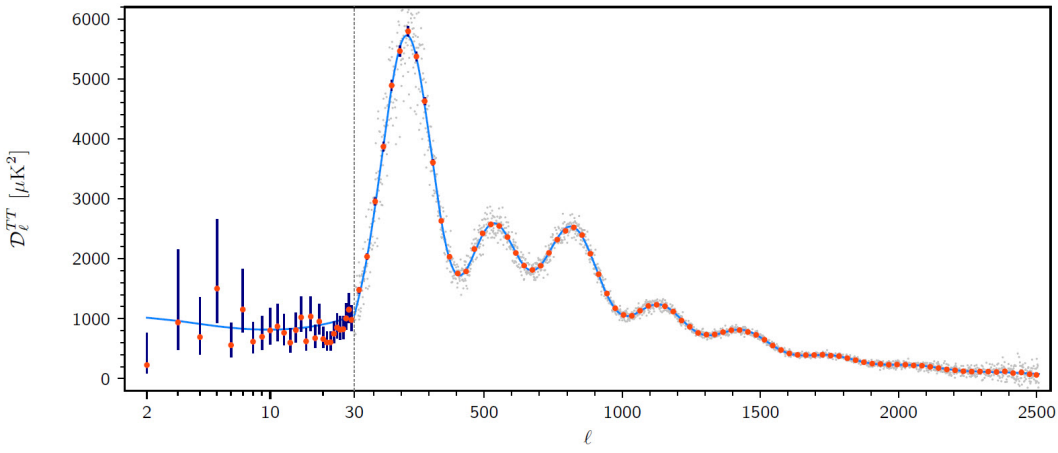

Fig. 2.3. Planck 2018 temperature power spectrum $\mathcal{D}_\ell^{TT} = \ell(\ell+1)\, C_\ell^{TT} \,/\, 2\pi$. The red dots are measurements made with Planck, shown with dark blue error bars. The light blue curve represents the best fit to the ΛCDM model. The x-axis is logarithmic up to $\ell = 30$ (the vertical dotted line) and linear at higher ℓ. Credit: [**4**].

2.2. Gravitational Lensing

According to general relativity, mass and energy warp the fabric of spacetime. In addition, every freely moving object follows a geodesic, which is the path of least spacetime curvature. Hence, the gravity of nearby massive objects might bend the path of photons propagating through spacetime. As a result, the photons received on Earth might not be emitted from their arriving direction. This phenomenon is known as *gravitational lensing*, and it occurs when light from distant sources, such as stars or galaxies (the source), passes near massive objects like black holes or galaxy clusters (the lens). The massive object's gravitational field acts as a lens, bending and distorting the light's path. This effect can lead to magnified, distorted, or multiple images of the background source. Figure 2.4 shows an example of gravitational lensing.

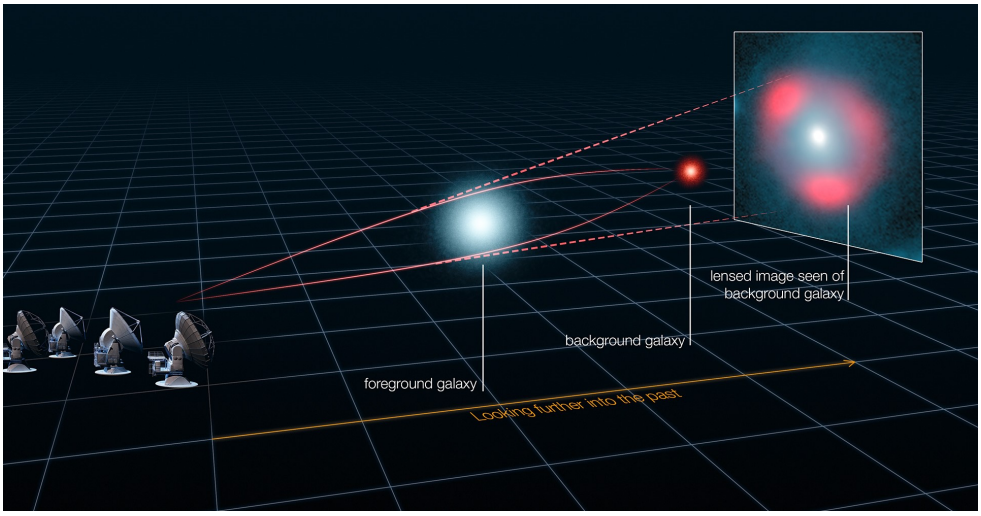

Fig. 2.4. An example of gravitational lensing: The light of a distant background galaxy (source) is bent by the foreground galaxy (lens), causing a distorted image of the source. Credit: ALMA (ESO/NRAO/NAOJ), L. Calçada (ESO), Y. Hezaveh et al.

Gravitational lensing is a powerful tool to study both the lens and source objects, as well as the intervening environment between the source and the observer. It can be used to infer the physical properties of the lens, such as the mass and spin of black holes [**71, 64, 68**] or the dark matter distribution of galaxies and galaxy clusters [**15, 41, 38, 42, 35**]. Also, it can magnify the light from distant objects (e.g., supernovae or ancient galaxies), allowing us to study them in more detail and even detect otherwise faint or distant objects [**26, 27, 14, 31, 62, 13, 34, 37**]. It can serve as a probe to constrain cosmological parameters [**23, 57, 72, 25, 29, 32, 9, 47, 18**], such as the Hubble constant, dark energy properties, and the density of the universe. Depending on the relative positions of the source, lens, and observer, as well as the distribution of mass within the lens, three different regimes of gravitational lensing exist, which are presented with examples in Figure 2.5.

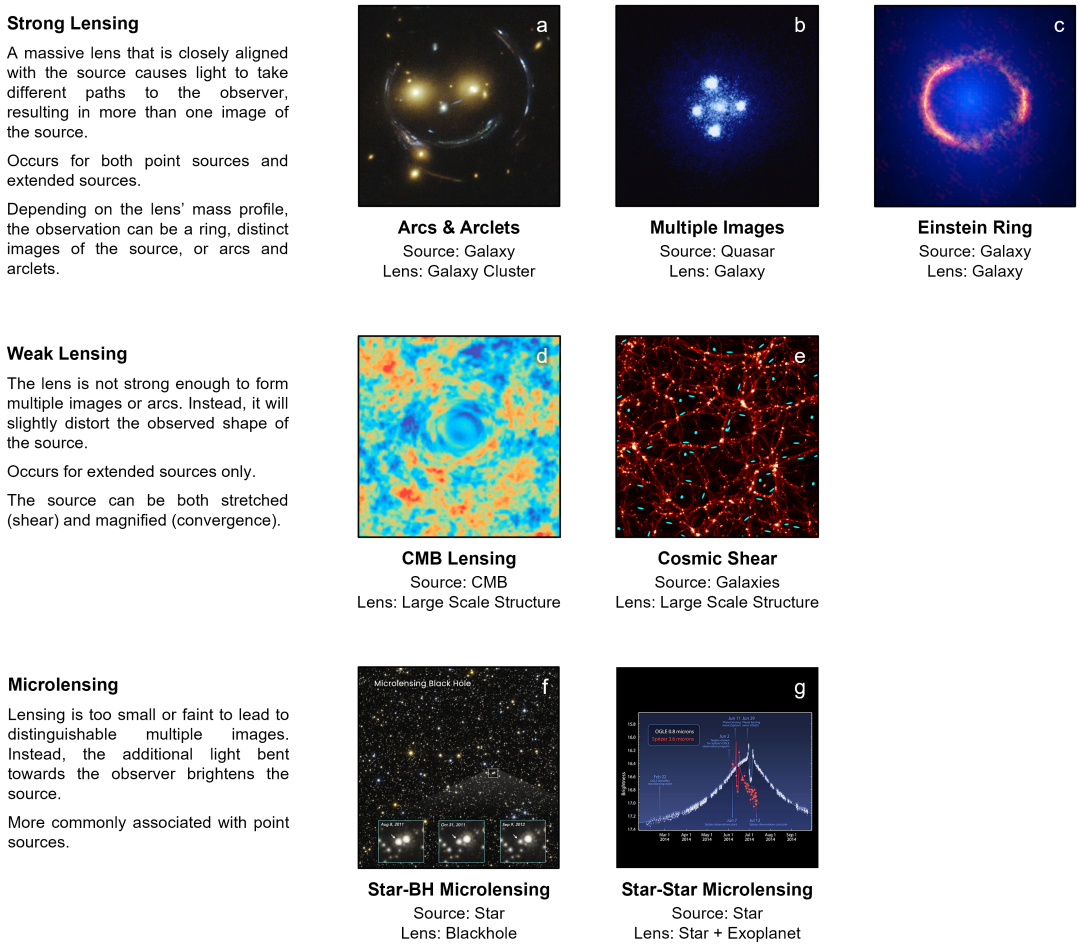

Fig. 2.5. Different types of gravitational lensing with examples for each category. Image Credits: (a) NASA/ESA/JPL-Caltech, (b) NASA/ESA/STScI, (c) NRAO/ESO/NAOJ/NASA/ESA, (d) W. Hu / T. Okamoto, (e) Canada France Hawaii Telescope, (f) NASA/ESA/STScI, (g) NASA/JPL-Caltech/Warsaw University Observatory

2.3. CMB Lensing

As presented in Table 2.1 and Figure 2.5, CMB lensing is a secondary effect on cosmic microwave background and is considered an instance of weak gravitational lensing. When CMB photons propagate in space, they pass through cosmic structures (e.g., galaxies and galaxy clusters), the gravity of which can slightly deflect photons from their original path. Hence, the observed directions of CMB photons do not represent their original emission direction. Based on a rough estimate, the deflection angles are $\sim 2'$ (RMS) and they are correlated over sky areas as large as $\sim 2^\circ$ [5], comparable to the degree-scale primary fluctuations of CMB [**55, 30**]. The lensing also introduces small amounts of non-Gaussianity (non-zero higher order moments) and statistical anisotropy (non-zero off-diagonal covariance elements) into the primordial signal.

[5]This corresponds to the angular size of a typical galaxy cluster at redshift 2.

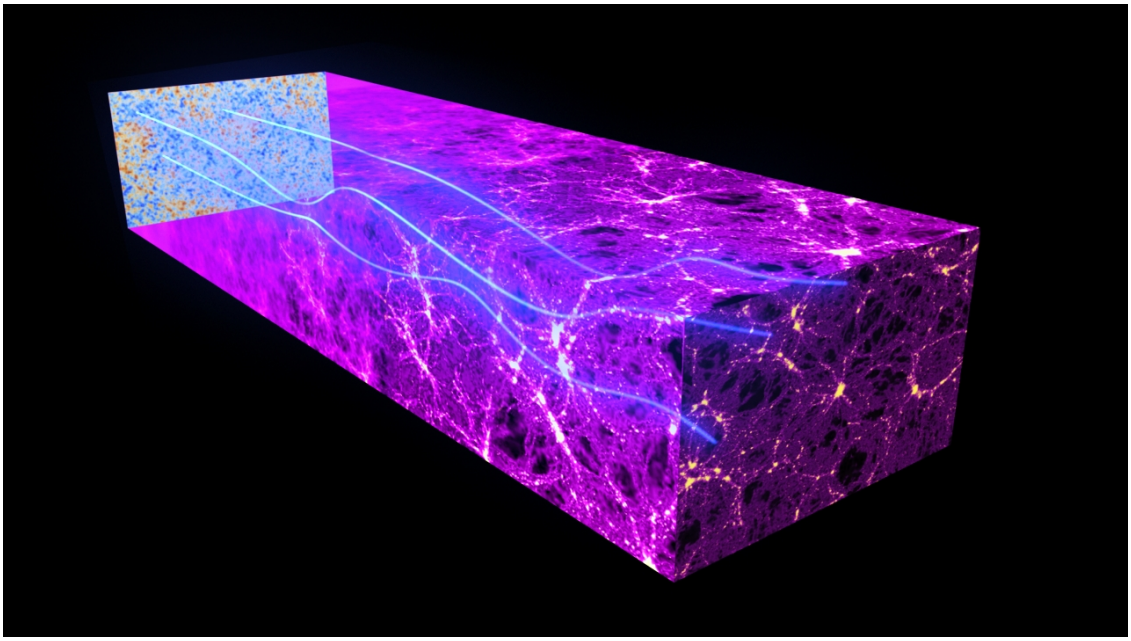

Fig. 2.6. CMB Lensing. Credit: ESA and the Planck Collaboration.

Correcting the lensing effect is important for obtaining unbiased estimates of cosmological parameters from the CMB. Furthermore, analyzing the additional information introduced by lensing enables probing the state of the universe during the course of deflection events. Finally, the lensing alters the polarization of CMB photons, most importantly by introducing a B-mode pattern that can be confused with the primordial B-mode signal from gravitational waves[6]. In this study, we are interested in correcting the lensing effects on the temperature power spectrum. Hereafter, we assume that the encountered sky extent and angular sizes are small enough for the validity of flat sky and small angle approximations.

2.3.1. Lensing Equations

Deflection Field

In this part, we quantify CMB lensing by deriving a formula for the deflection angle of CMB photons. Figure 2.7 depicts the deflection of a CMB photon by an overdense structure, viewed in the comoving coordinate system. According to general relativity, the deflection occurs in a plane (i.e., the *deflection plane*). At each part of the path, the infinitesimal deflection $d\beta$ from the initial path is given by

$$d\beta = -2\nabla_{\perp}\Psi \;\; d\chi, \tag{2.6}$$

[6]In the absence of inflationary tensor perturbations, primordial CMB is expected to exhibit only E-mode polarization. Detecting B-mode polarization serves as the distinctive signature of tensor perturbations during inflation, the primary source of primordial gravitational waves.

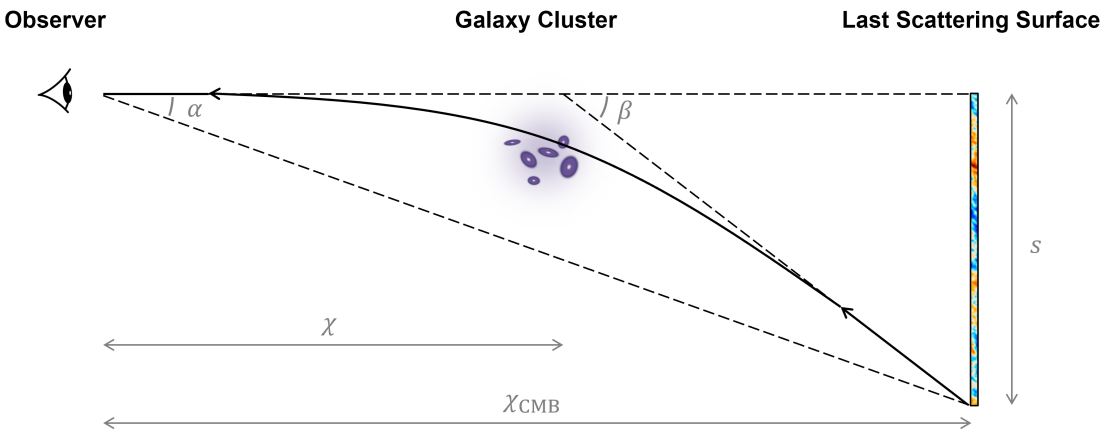

Fig. 2.7. Deflection of a CMB photon by an overdense structure.

where $\nabla_\perp$ is the normal component of the gradient (perpendicular to the path), Ψ is the Weyl potential, and $d\chi$ is the infinitesimal comoving distance traveled by the photon. If all the gravitational influence was from a single overdense region, we could have calculated the total deflection by integrating Equation 2.6 over the photon's path:

$$\beta = -2 \int_{\text{path}} \nabla_\perp \Psi \; d\chi = \frac{4GM}{bc^2}, \tag{2.7}$$

where M and b are the mass of the overdense region and the closest approach distance (in the absence of lensing), respectively. Then, we could have used the geometric relation[7]

$$S_\kappa(\chi_{\text{CMB}} - \chi)\, \beta = S_\kappa(\chi_{\text{CMB}})\, \alpha \tag{2.8}$$

to relate β to the *observed deflection* α. $S_\kappa(\chi)$ comes from the Robertson-Walker metric and depends on the geometry of the universe,

$$S_\kappa(\chi) = \begin{cases} R \sin(\chi/R) & \kappa = +1 \text{ (Closed)} \\ \chi & \kappa = 0 \quad \text{(Flat)} \\ R \sinh(\chi/R) & \kappa = -1 \text{ (Open)} \end{cases}, \tag{2.9}$$

with R being the present-day radius of curvature of the universe. Finally, combining Equations 2.7 and 2.8 would lead to

$$\alpha = \frac{S_\kappa(\chi_{\text{CMB}} - \chi)}{S_\kappa(\chi_{\text{CMB}})} \frac{4GM}{bc^2}. \tag{2.10}$$

In reality, as depicted in Figure 2.8, CMB photons pass through numerous overdensities (potential wells) and underdensities (potential hills) throughout their journey to the observer. In this case, we need to compute the exact photon trajectories (i.e., perform ray tracing) to calculate the deflection angle. Since multiple lenses exist, the deflections do not occur

[7]Recall that the small angle formula is valid.

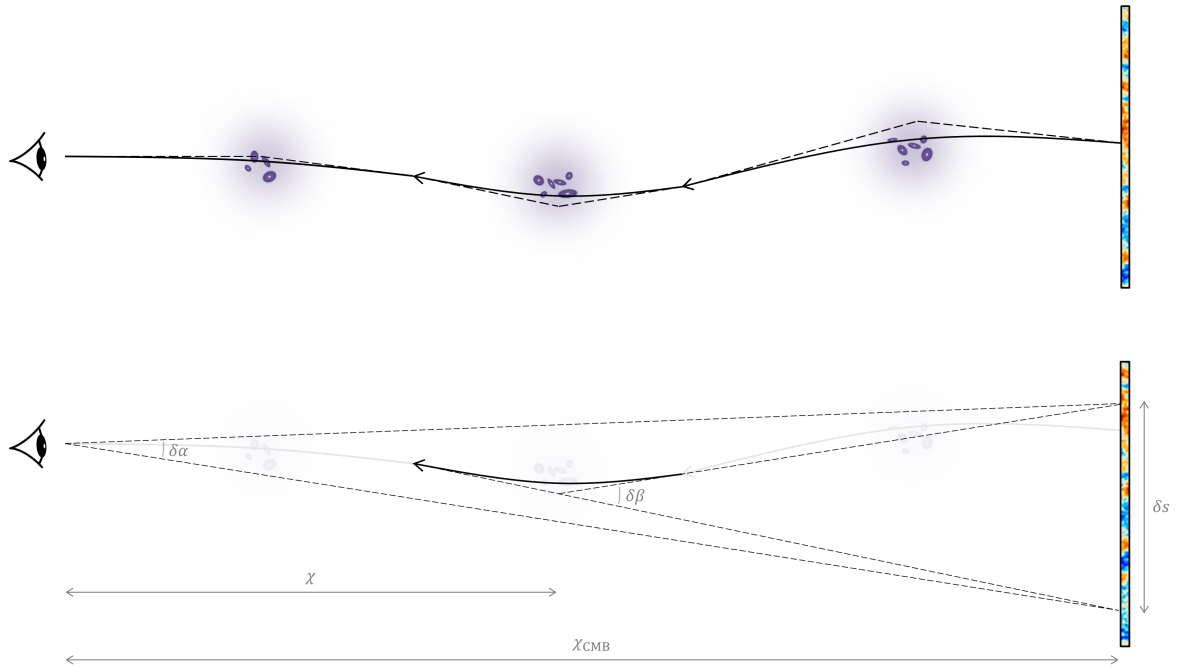

Fig. 2.8. Deflection of a CMB photon by intervening cosmic structures (top), with one deflection highlighted (bottom).

in a single plane[8], which necessitates defining a 2D coordinate system on the sky plane to quantify deflections[9]. In this coordinate system, the infinitesimal vector deflection $d\boldsymbol{\beta}$ from the initial path is given by

$$d\boldsymbol{\beta} = -2\nabla_{\perp}\Psi \; d\chi, \tag{2.11}$$

where this time $\nabla_{\perp}$ represents the normal component**s** of the gradient. In a similar fashion to Equation 2.8, we can write

$$S_{\kappa}(\chi_{\mathrm{CMB}} - \chi)\, d\boldsymbol{\beta} = S_{\kappa}(\chi_{\mathrm{CMB}})\, d\boldsymbol{\alpha} \tag{2.12}$$

By substituting $d\boldsymbol{\beta}$ from Equation 2.11, we can find the infinitesimal observed deflection:

$$d\boldsymbol{\alpha} = -2\, \frac{S_{\kappa}(\chi_{\mathrm{CMB}} - \chi)}{S_{\kappa}(\chi_{\mathrm{CMB}})}\, \nabla_{\perp}\Psi \; d\chi, \tag{2.13}$$

and integrate it over the photon's path to calculate the total observed deflection:

$$\boldsymbol{\alpha} = -2 \int_{\mathrm{path}} \frac{S_{\kappa}(\chi_{\mathrm{CMB}} - \chi)}{S_{\kappa}(\chi_{\mathrm{CMB}})}\, \nabla_{\perp}\Psi \; d\chi. \tag{2.14}$$

However, there is a caveat to this approach: Performing ray tracing is computationally expensive, and sometimes impossible. There exist approximate approaches to simplify this situation. The Born approximation does so by assuming the gravitational deflections are so

[8]This is why lensing in the presence of multiple lenses is referred to as *multiplane lensing*.
[9]For the moment, the orientation of this coordinate system does not matter.

small that the photon's path can be considered as a perturbation around the line of sight. Accordingly, the potential at each location can be approximated using a first-order expansion around the closest point on the line of sight. As a result, integrations over the photon's path can be approximated with integrations over the line of sight (i.e., the unperturbed path). For instance, the total observed deflection can be approximated as:

$$\boldsymbol{\alpha} = -2 \int_{\text{path}} \frac{S_\kappa(\chi_{\text{CMB}} - \chi)}{S_\kappa(\chi_{\text{CMB}})} \nabla_\perp \Psi \; d\chi \approx -2 \int_{\text{los}} \frac{S_\kappa(\chi_{\text{CMB}} - \chi)}{S_\kappa(\chi_{\text{CMB}})} \nabla_\perp \Psi \; d\chi, \tag{2.15}$$

where $\nabla_\perp$ in the last equality represents the gradient components perpendicular to the line of sight. Figure 2.9 shows a graphical illustration of this approach.

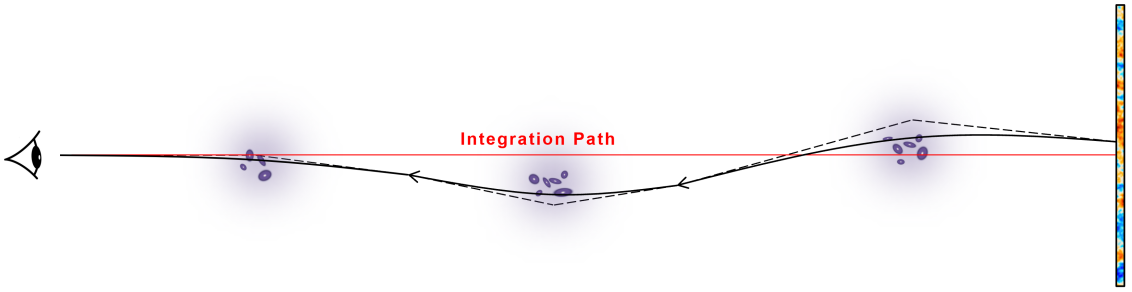

Fig. 2.9. Using Born approximation, one can integrate along the line of sight instead of performing computationally expensive ray tracing.

Lensing Potential

Since we can safely approximate the source CMB radiation to be instantaneously emitted, it is convenient to aggregate all lens information in a 2D map of the lensing potential on the sky plane. To do so, we first use the relation

$$\nabla_\perp(\cdot) = \frac{\nabla_{\hat{\boldsymbol{n}}}(\cdot)}{S_\kappa(\chi)} \tag{2.16}$$

to convert the spatial gradient to an angular gradient on the sphere. We may now define the *lensing potential*,

$$\psi(\hat{\boldsymbol{n}}) := -2 \int \frac{S_\kappa(\chi_{\text{CMB}} - \chi)}{S_\kappa(\chi_{\text{CMB}})\, S_\kappa(\chi)} \Psi \; d\chi, \tag{2.17}$$

so that the observed deflection angle is given by

$$\boldsymbol{\alpha}(\hat{\boldsymbol{n}}) = \nabla_{\hat{\boldsymbol{n}}} \psi. \tag{2.18}$$

2.3.2. Observable Effects

We now focus on how the lensing potential affects the observations of CMB. As stated earlier, we are interested in the effects on the temperature field. In the flat-sky limit, this effect can

be described as a remapping of the primary (unlensed) CMB map given by

$$\tilde{T}(\boldsymbol{x}) = T\Big(\boldsymbol{x} + \boldsymbol{\alpha}(\boldsymbol{x})\Big), \tag{2.19}$$

where $\boldsymbol{x}$ is a direction in the sky, and $\tilde{T}$ and T represent observed and primary CMB maps, respectively. It is often useful to Taylor expand the lensing displacements,

$$\tilde{T}(\boldsymbol{x}) = T(\boldsymbol{x}) + \alpha^a \nabla_a T(\boldsymbol{x}) + \frac{1}{2}\alpha^a \alpha^b \nabla_a \nabla_b T(\boldsymbol{x}) + \cdots, \tag{2.20}$$

with a and b referring to different dimensions in the coordinate system, to gain intuition about the lensing effects.

Mode Coupling

Lensing causes independent temperature modes to become correlated through the lensing potential. To prove this, we start from the Fourier transform of the temperature map:

$$\tilde{T}(\boldsymbol{k}) = \int \tilde{T}(\boldsymbol{x}) e^{-i\boldsymbol{k}\cdot\boldsymbol{x}} d\boldsymbol{x}. \tag{2.21}$$

Keeping at most the first-order terms, we will have

$$\tilde{T}(\boldsymbol{k}) \approx \int \Big(T(\boldsymbol{x}) + \alpha^a \nabla_a T(\boldsymbol{x})\Big)\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}\, d\boldsymbol{x} = T(\boldsymbol{k}) + \int \alpha^a \nabla_a T(\boldsymbol{x})\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}\, d\boldsymbol{x}. \tag{2.22}$$

Using this equation, the average $\Big\langle\, \tilde{T}(\boldsymbol{k})\, \tilde{T}(\boldsymbol{k}')\, \Big\rangle$ over an ensemble of CMB fluctuations for fixed lenses can be approximated as:

$$\begin{aligned} \Big\langle\, \tilde{T}(\boldsymbol{k})\, \tilde{T}(\boldsymbol{k}')\, \Big\rangle \approx{}& \Big\langle\, T(\boldsymbol{k})\, T(\boldsymbol{k}')\, \Big\rangle \\ &+ \Big\langle\, T(\boldsymbol{k}) \int \alpha^a \nabla_a T(\boldsymbol{x})\, e^{-i\boldsymbol{k}'\cdot\boldsymbol{x}}\, d\boldsymbol{x}\, \Big\rangle \\ &+ \Big\langle\, T(\boldsymbol{k}') \int \alpha^a \nabla_a T(\boldsymbol{x})\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}\, d\boldsymbol{x}\, \Big\rangle. \end{aligned} \tag{2.23}$$

Assuming the unlensed, primary CMB is statistically isotropic and Gaussian distributed, its independent modes are decoupled. Hence, the first term is equal to

$$\Big\langle\, T(\boldsymbol{k})\, T(\boldsymbol{k}')\, \Big\rangle = \delta(\boldsymbol{k} - \boldsymbol{k}')\, C_\ell^{TT}, \tag{2.24}$$

with C_ℓ^{TT} representing the unlensed temperature power spectrum and $\ell \equiv |\boldsymbol{k}|$. However, the integrals in the second and third terms are non-zero. By substituting α^a with $\nabla^a \psi$ using Equation 2.18 and evaluating the integrals, we can find out that:

$$\Big\langle\, \tilde{T}(\boldsymbol{k})\, \tilde{T}(\boldsymbol{k}')\, \Big\rangle \approx \frac{1}{2\pi}\, \psi(\boldsymbol{k} + \boldsymbol{k}') \cdot \Big[\boldsymbol{k} C_\ell^{TT} + \boldsymbol{k}' C_{\ell'}^{TT}\Big]. \tag{2.25}$$

As the lensing potential interacts with CMB, it leads to a mixing of different spatial scales (i.e., modes) of CMB temperature and polarization fluctuations. This mixing (i.e., mode-coupling) introduces correlations between fluctuations on different scales, resulting in off-diagonal elements into the covariance matrix of the observed CMB. The characteristic spacing of these elements is $\delta\ell = 50$, given by the peak of the deflection angle power spectrum.

Power Spectrum

The lensing also alters the temperature power spectrum. Using the Taylor expansion of CMB temperature (Equation 2.19) and under the flat sky approximation, the lensed temperature power spectrum to first order in the lensing potential power spectrum $C_\ell^{\psi\psi}$ is given by:

$$C_\ell^{\tilde{T}\tilde{T}} = (1 - \ell^2 R^\psi) C_\ell^{TT} + \int \frac{d^2\boldsymbol{k}'}{(2\pi)^2} \Big[\boldsymbol{k}' \cdot (\boldsymbol{k} - \boldsymbol{k}')\Big]^2 C_{|\boldsymbol{k}-\boldsymbol{k}'|}^{\psi\psi} C_{\ell'}^{TT}, \tag{2.26}$$

where R^ψ is half of the total mean-squared deflection, defined by:

$$R^\psi := \frac{1}{2}\langle |\nabla\psi|^2 \rangle \sim 3 \times 10^7. \tag{2.27}$$

In this project, we aim to reconstruct the unlensed temperature power spectrum. We do so by removing the lensing effect from observed CMB maps, and computing the power spectra of the resulting delensed maps. Having the physical foundations of our work discussed, we now turn to introduce the formal definition of the delensing problem and our statistical approach in the next section.

2.4. Problem Definition and Statistical Framework

2.4.1. Inverse Problems

Every physical problem involves inferring unknown quantities based on observed data. This task can take various forms, such as discovering underlying physical laws through an investigation of relationships between observables, employing well-defined mathematical models to predict the evolution of a system's state, or deducing unobservable parameters from their observed effects. These examples represent only a subset of the various analytical approaches employed in physics.

Inverse problems are a class of problems that involve inferring underlying parameters giving rise to a specific set of observations. They stand in contrast to "forward" or "direct" problems, where predetermined inputs are fed into a well-defined model to predict corresponding outcomes. In the realm of inverse problems, the outcomes are provided, and the objective is to deduce the inputs that led to them. This is done using a model or system that might not be entirely known or could be subject to uncertainties. In astrophysics, inverse problems play a crucial role in understanding the properties and evolution of the universe. Figure 2.10 presents some examples of inverse problems in astrophysics.

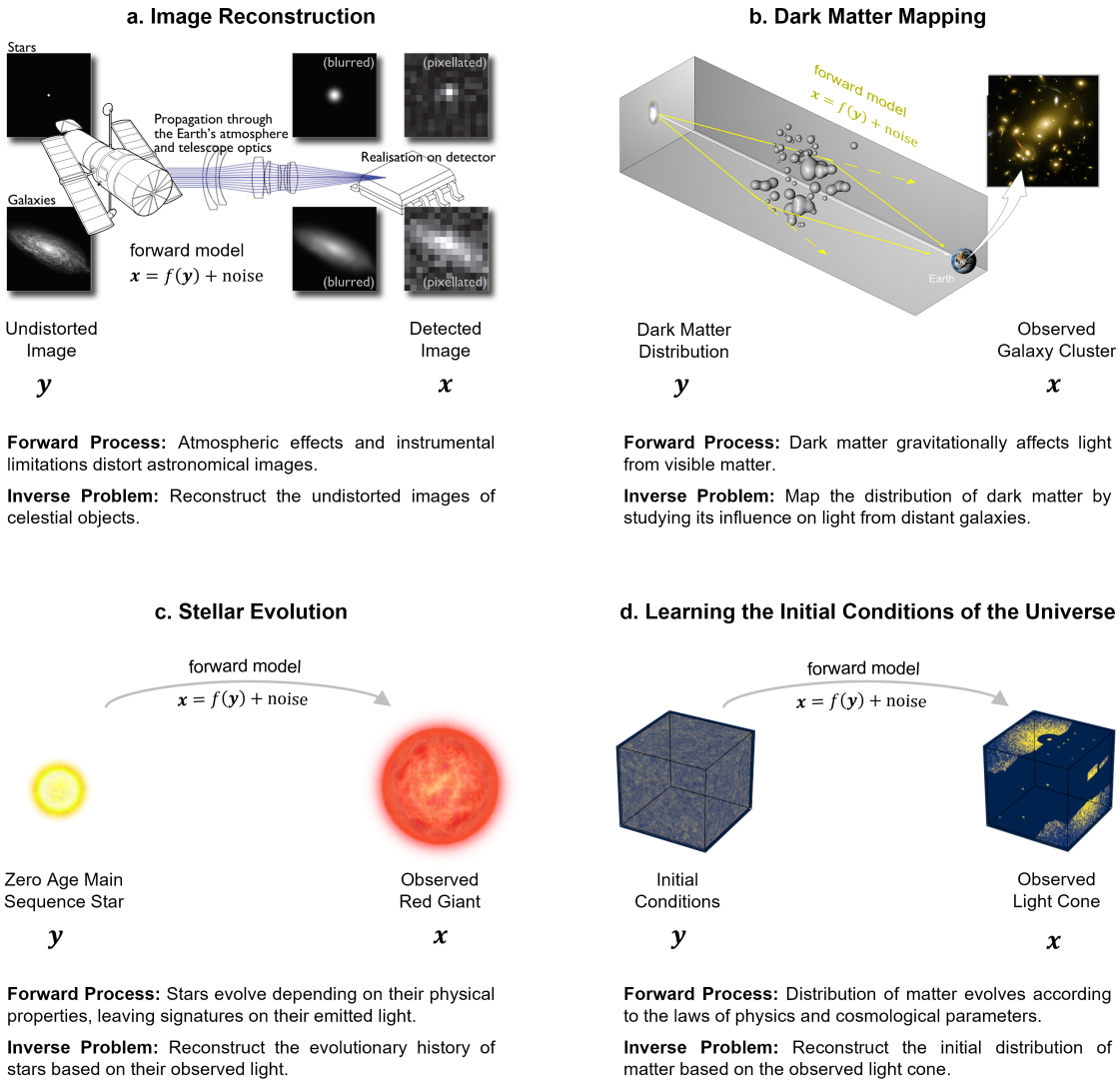

Fig. 2.10. Examples of inverse problems in astrophysics. Image Credits: (a) R. Mandelbaum et al. [**59**], (b) ESA, (c) Viktor Hahn, (d) Adapted from Learning the Universe (LtU) Collaboration internal material.

Formally speaking, an inverse problem is defined using five components:

- **Forward Model**: The set of equations that govern the evolution of the system.
- **Model Parameters**: Inputs of the forward model.
- **Physical State**: The state of the system predicted by the forward model.
- **Observations:** The observable implication of the physical state on the environment.
- **Noise Model:** The mathematical representation of uncertainties that affect the observations.

The goal of the inverse problem is to infer model parameters from the observations.

CMB Delensing as an Inverse Problem

CMB Delensing can be formulated as an inverse problem with the following components:

- **Forward Model**: The equation to calculate the photon's deflection angle, i.e., Equation 2.15.
- **Model Parameters**: An unlensed CMB temperature map, i.e., a map with all secondary effects present but weak gravitational lensing.
- **Physical State / Observations**[10]: A CMB temperature map with all secondary effects, including weak gravitational lensing, present.
- **Noise Model:** We assume a diagonal Gaussian noise structure for pixels, where each pixel's noise is independent of others[11].

Having established the components of the CMB Delensing problem, the next subsection outlines our employed approach to tackle this problem.

2.4.2. Inference

Due to incompleteness, noise, and uncertainties in observational data, inverse problems are often ill-posed, i.e., they may have multiple solutions or lack stability. *Statistical inference* frameworks can effectively address these challenges by characterizing and quantifying uncertainties. They offer a rigorous approach to tackling inverse problems.

Bayesian Framework

The Bayesian framework is a statistical approach that deals with uncertainty by using probability distributions to represent beliefs. It relies on the Bayes theorem to express the belief about an uncertain parameter based on initial knowledge and observed data. This approach starts with an initial belief represented by the *prior* probability distribution $p(\boldsymbol{y})$, which incorporates any existing information or prior knowledge about the uncertain parameter $\boldsymbol{y}$. With observed data $\boldsymbol{x}$ at hand, the Bayesian framework updates the prior by considering the *likelihood* of observing the data given different parameter values $p(\boldsymbol{x}|\boldsymbol{y})$. This is done using the Bayes theorem:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})}{p(\boldsymbol{x})} \tag{2.28}$$

The result $p(\boldsymbol{y}|\boldsymbol{x})$ is the *posterior* probability distribution, which represents the updated beliefs about $\boldsymbol{y}$. The distribution $p(\boldsymbol{x})$ is called *evidence* or *marginal likelihood* and, in

[10]For our level of rigor, we assume the temperature map corresponds to both the physical state and observations. In fact, the actual observation is the intensity of the radiation, which the temperature field is inferred from. However, in our inverse problem, we assume that the temperature field is already calculated and given to us.

[11]Nevertheless, our framework is designed to be extensible to accommodating non-diagonal noise as well.

principle, can be calculated by integrating the joint distribution of data and parameters:

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{y} = \int p(\boldsymbol{x}|\boldsymbol{y}) p(\boldsymbol{y})\, d\boldsymbol{y} \tag{2.29}$$

The Bayesian approach to CMB delensing aims to access the posterior distribution of unlensed CMB maps given an observed (lensed) map. More specifically, considering that a map is represented by a group of pixels, the goal is to access the joint distribution of unlensed pixel values conditioned on the observations. The large number of pixels makes this task an instance of *high-dimensional inference*. In such a problem space, direct modeling of the posterior is challenging due to overwhelming computational complexity. A more favored approach called *posterior sampling* aims to generate samples from the posterior[12] instead of directly modeling it. This approach can efficiently handle complex and high-dimensional posteriors and ones with intricate geometries.

Numerous techniques exist for posterior sampling. Some approaches like Markov Chain Monte Carlo (MCMC) [**63, 8**], Nested Sampling [**75**], and Variational Inference [**43, 10**] rely on evaluating the likelihood $p(\boldsymbol{x}|\boldsymbol{y})$, which in many cases is unknown, intractable, or difficult to compute, prohibiting the direct application of the Bayes theorem. However, the group of methods known as Implicit Likelihood Inference (ILI)[13] do not need direct access to the likelihood. Instead, likelihood is implicitly encoded in the data used for inference. For instance, in Simulation-Based Inference (SBI) [**76, 17**] (a subclass of ILI), the likelihood is implicitly defined using a forward model (i.e., a simulator). This computational model generates simulated data based on a given set of parameters, effectively replacing the need for directly modeling the likelihood function.

Early approaches to SBI, such as Approximate Bayesian Computation (ABC) [**77, 60, 74**], do not involve machine learning (ML). Although these methods can approximate complex likelihood functions, they remain computationally expensive, particularly in high-dimensional problems [**1**]. The reason is they rely on running simulations during inference. Furthermore, they require repeating the entire inference chain when more observations become available. On the contrary, incorporating ML into SBI enables efficient high-dimensional inference. ML methods involve training a surrogate model (e.g., a neural network) for the simulator in advance, thus eliminating the need for repeated simulations during inference. This amortization significantly accelerates the inference process in SBI [**17**].

To summarize, we intend to delens CMB by performing high-dimensional Bayesian inference at the pixel level (i.e., infer the value of pixels of the unlensed CMB temperature map). This task is computationally intractable using conventional methods. Furthermore, the

[12] In our case, the posterior samples are potential realizations of the unlensed CMB map.
[13] Implicit Likelihood Inference is also referred to as Likelihood-free Inference.

complexities involved in accessing and computing the likelihood even make SBI approaches challenging. We propose a deep learning framework for posterior sampling to address these challenges. The deep learning foundations of our work will be laid in the next chapter.

Chapter 3

Sampling the Unseen, A *Deep* Dive

This chapter aims to cover the basic concepts of deep learning and introduce the neural network architecture used for posterior sampling. It begins by introducing the fundamental components of neural networks in Section 3.1. Then, Section 3.2 discusses the applications of neural networks to model probability distributions, mainly focusing on Variational Autoencoders. The first two sections lay the groundwork for presenting the Hierarchical Probabilistic U-Net architecture in Section 3.3. The chapter will conclude with Section 3.4 exploring methods to evaluate the model's performance.

3.1. Neural Networks

Neural networks are powerful mathematical models for quantifying complex relationships. Over the past few decades, they have garnered significant attention and have become a cornerstone of modern machine learning and artificial intelligence research. Their countless applications range from pattern recognition and natural language processing to facilitating scientific discoveries. Figure 3.1 displays some of the remarkable advancements enabled by deep learning, including COVID-19 vaccine and drug discovery [**58, 48**], producing realistic images using generative models like StyleGAN [**45**], predicting previously unknown protein structures with AlphaFold [**44**], "enhancing communication and problem-solving capabilities with large language models like ChatGPT"[1] [**67, 66, 11**], and powering self-driving cars through reinforcement learning [**6, 82**].

[1]This phrase and its cited references are based on ChatGPT's generated text and suggestions!

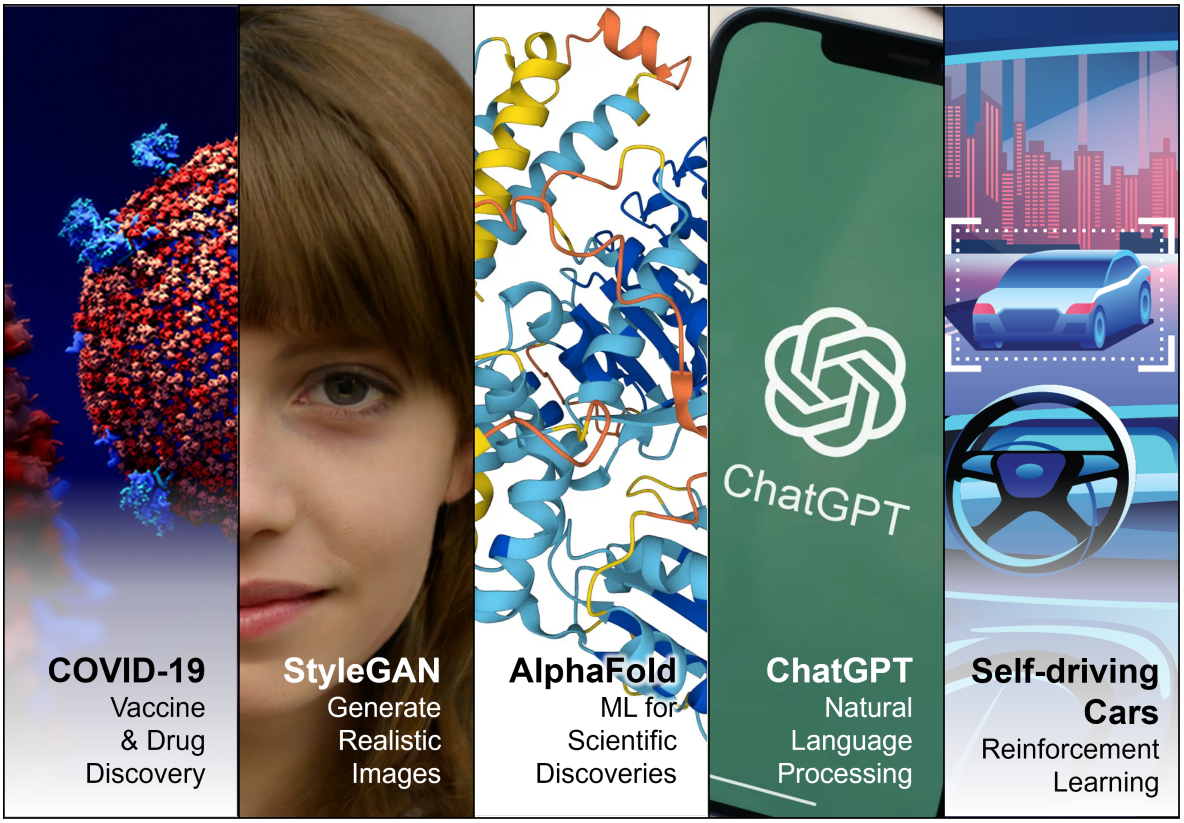

Fig. 3.1. Some of the remarkable advancements enabled by deep learning. Image Credits (left to right): Design Cells/Science Photo Library, E. Salvaggio, AlphaFold Protein Structure Database, D. Thomazini/Shutterstock, IDTechEx.

3.1.1. Perceptron

Perceptrons (aka *artificial neurons* or simply *neurons*) are the fundamental components of a neural network. They resemble neurons in the human nervous system[2]. As summarized by Figure 3.2, a perceptron receives its input(s) $\boldsymbol{x}$, applies a linear transformation $\boldsymbol{w} \cdot \boldsymbol{x} + b$, and passes the result to a non-linear function f, leading to the output y of the perceptron. $\boldsymbol{w}$, b, and f are called weight(s), bias, and activation function, respectively. A neural *network* is mainly constructed by stacking perceptrons together.

3.1.2. Layers

A neural network typically consists of several *layers*. Each layer is a group of perceptrons receiving the same input. Depending on the specific task and architecture, neural networks can have various types of layers, some of which are presented below.

[2]The idea of using a mathematical model inspired by the human brain for pattern recognition and learning tasks roots in the work of Frank Rosenblatt, published in his paper [**70**] in 1958.

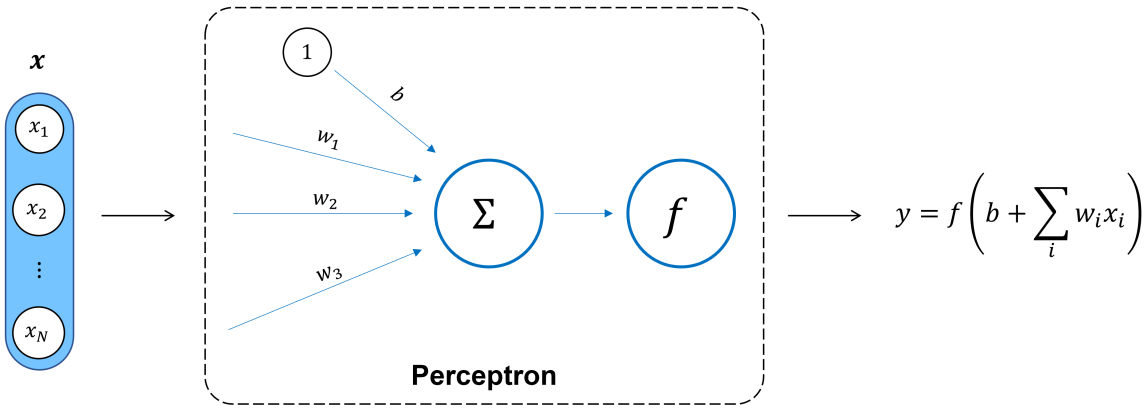

Fig. 3.2. Inside of a perceptron.

Fully Connected Layers

In a *fully connected* layer, each neuron receives input from every neuron in the previous layer. By stacking fully connected layers, one can create the simplest type of neural network, called a Multi-Layer Perceptron (MLP). Figure 3.3 showcases a typical neural network architecture consisting of four MLPs. Some early applications of MLPs in astrophysics include estimating galaxy redshifts from their Spectral Energy Distribution (SED) [**79**] and galaxy classification [**73**].

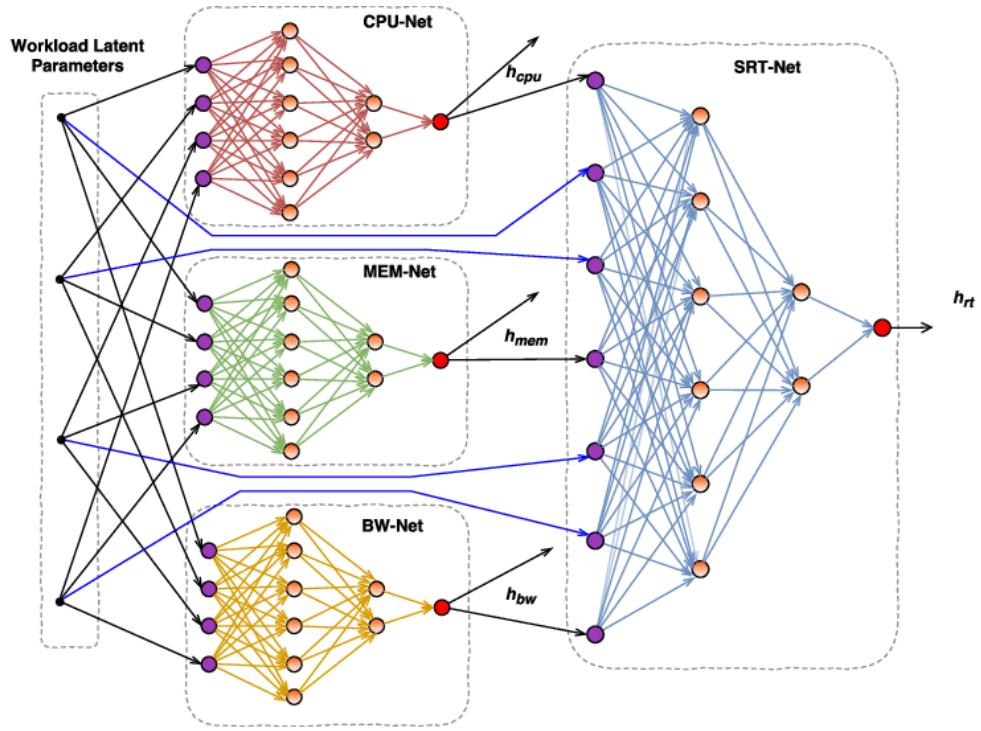

Fig. 3.3. A neural network consisting of four multi-layer perceptrons. Purple and red neurons indicate inputs and outputs, respectively. Credit: [**21**].

Convolutional Layers

Although MLPs afford plentiful complexity to learn intricate relationships, they are suboptimal for detecting patterns within a sequence. Convolutional Neural Networks (ConvNets) [**24, 51, 50**] exploit translational invariance[3] through the use of *convolutional* layers to capture local patterns and spatial dependencies within their input. Figure 3.4 illustrates the difference between fully connected and convolutional layers.

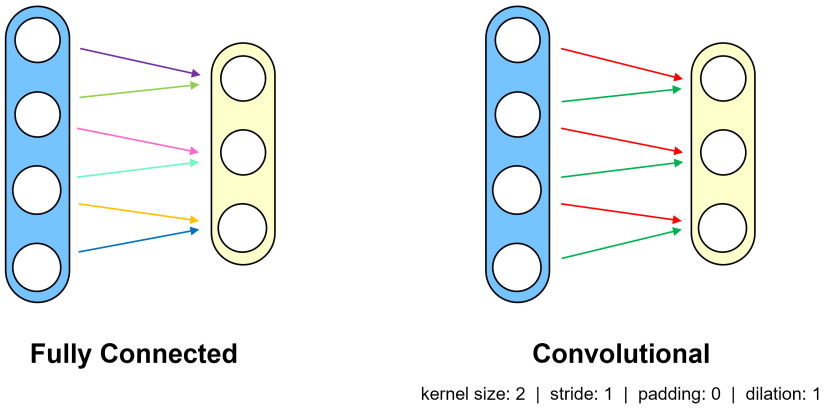

Fig. 3.4. Comparison of a fully connected layer with a convolutional layer. Each arrow color represents a unique value. In a convolutional layer, parameters are shared between neurons; hence, their weights are depicted with the same color.

In a convolutional layer, weights (and bias) are shared between neurons. Consequently, each neuron calculates the inner product of a specific input region with a template known as a *filter*. This computation provides a similarity measure, allowing convolutional layers to capture local patterns. The output of a convolutional layer is termed a *feature map*, where each element corresponds to the activation of a specific filter at a spatial location. This concept is illustrated in Figure 3.5, where applying a layer with shared weights equates sliding a filter over the input sequence, a process known as *convolving* the filter with the input.

Stacking convolutional layers enables the network to extract increasingly complex features and representations. Furthermore, sharing weights significantly reduces the number of learnable parameters compared to fully connected layers, making ConvNets more computationally efficient than MLPs. Figure 3.6 shows a typical ConvNet architecture. Some early applications of ConvNets in astrophysics and cosmology include identifying pulsars [**83**], object classification from 1D spectra [**40**], and galaxy classification [**19**].

[3]Translational invariance refers to the ability of the network to recognize patterns regardless of their position in the input data. This is particularly important in image processing tasks where the location of an object or a pattern within an image can vary.

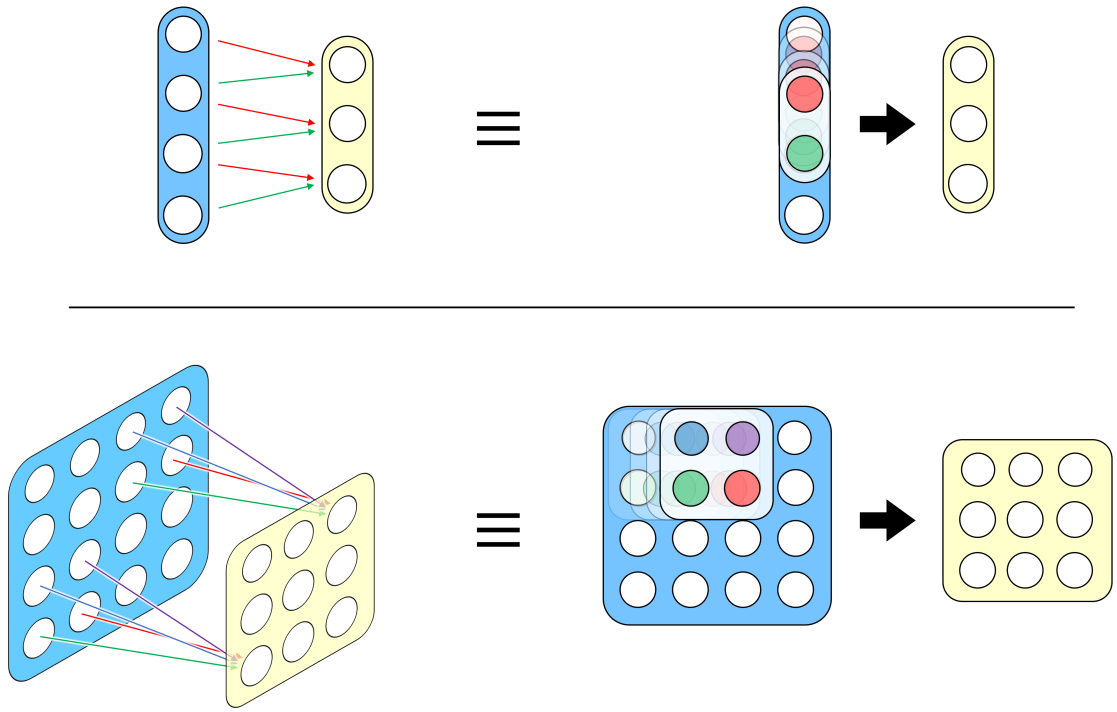

Fig. 3.5. Equivalence of weight sharing with convolution in 1D (top) and 2D (bottom).

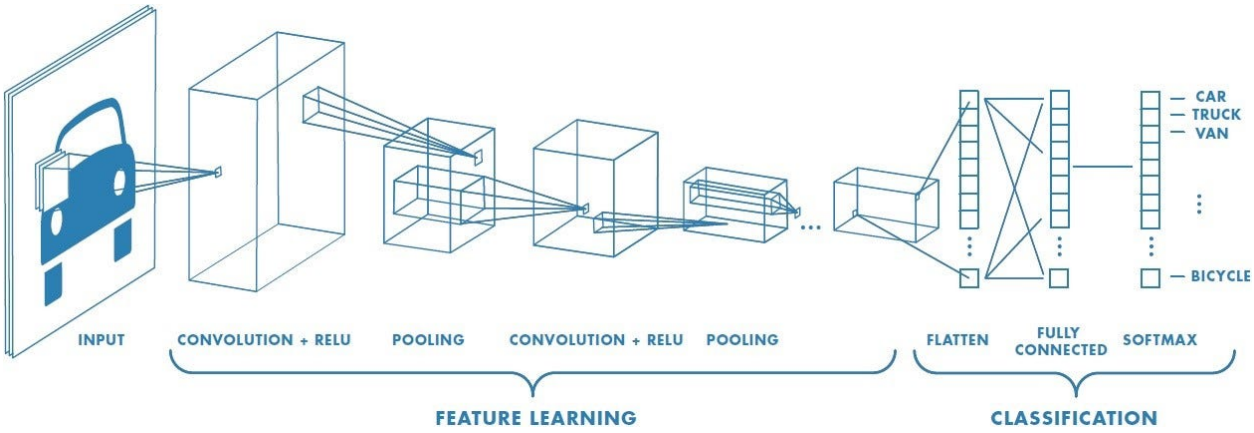

Fig. 3.6. A typical convolutional neural network. Credit: Sumit Saha.

A convolutional layer has several properties; among them are the number of filters, kernel size (spatial extent of filters), and stride (step size to slide filters across input). These properties affect the layer's output size and its neurons' receptive field[4]. It is common to apply padding to the input (add pixels to the edges) to reach a desired output size, or use a dilated filter (introduce gaps between filter elements) to expand the receptive field and

[4]Receptive field is the spatial extent of the network's input that influences the output of a neuron. Note that the network's input differs from the layer's immediate input.

capture information from a wider area. Figure 3.7 uses an example to introduce the main characteristics of a convolutional layer.

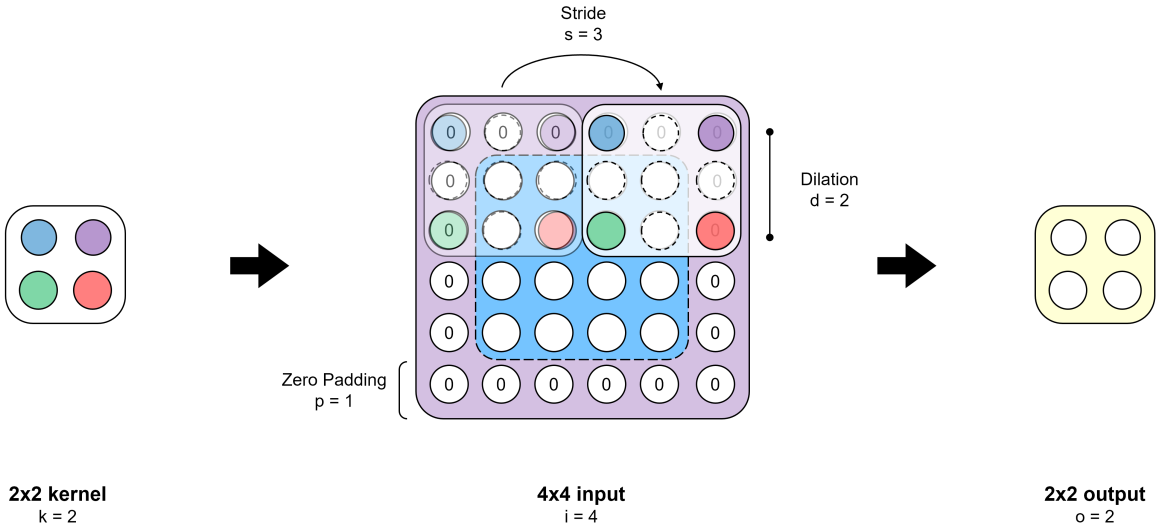

Fig. 3.7. Main properties of convolutional layers.

The output size o of a convolutional layer can be calculated using the following formula:

$$o = \left\lfloor \frac{i + 2p - k - (k-1)(d-1)}{s} \right\rfloor + 1 \tag{3.1}$$

For a comprehensive guide on convolution arithmetics, see [**20**].

Downsampling and Upsampling

ConvNets often include layers that reduce the spatial dimensions of feature maps. This helps the network to reduce computational complexity and capture high-level, abstract features by progressively decreasing the spatial resolution. Furthermore, they make the network more robust to small translations in the input by summarizing local information into a more compact representation.

For ConvNets with dense predictions[5], having layers to restore the spatial resolution of feature maps becomes essential. These layers enable lower-resolution feature maps to be merged or concatenated with higher-resolution feature maps from previous layers, which is necessary to recover fine-grained details lost during downsampling, allowing for localization and reconstruction of objects. Figure 3.8 introduces commonly used downsampling and upsampling layers in ConvNets. For a more detailed discussion of different neural network layers, including normalization and regularization layers, see [**28, 39**].

[5]When a neural network produces dense predictions, it generates output values for multiple locations or elements in the input data, rather than producing a single output. This is common in tasks such as image segmentation, object detection, and high-dimensional inference

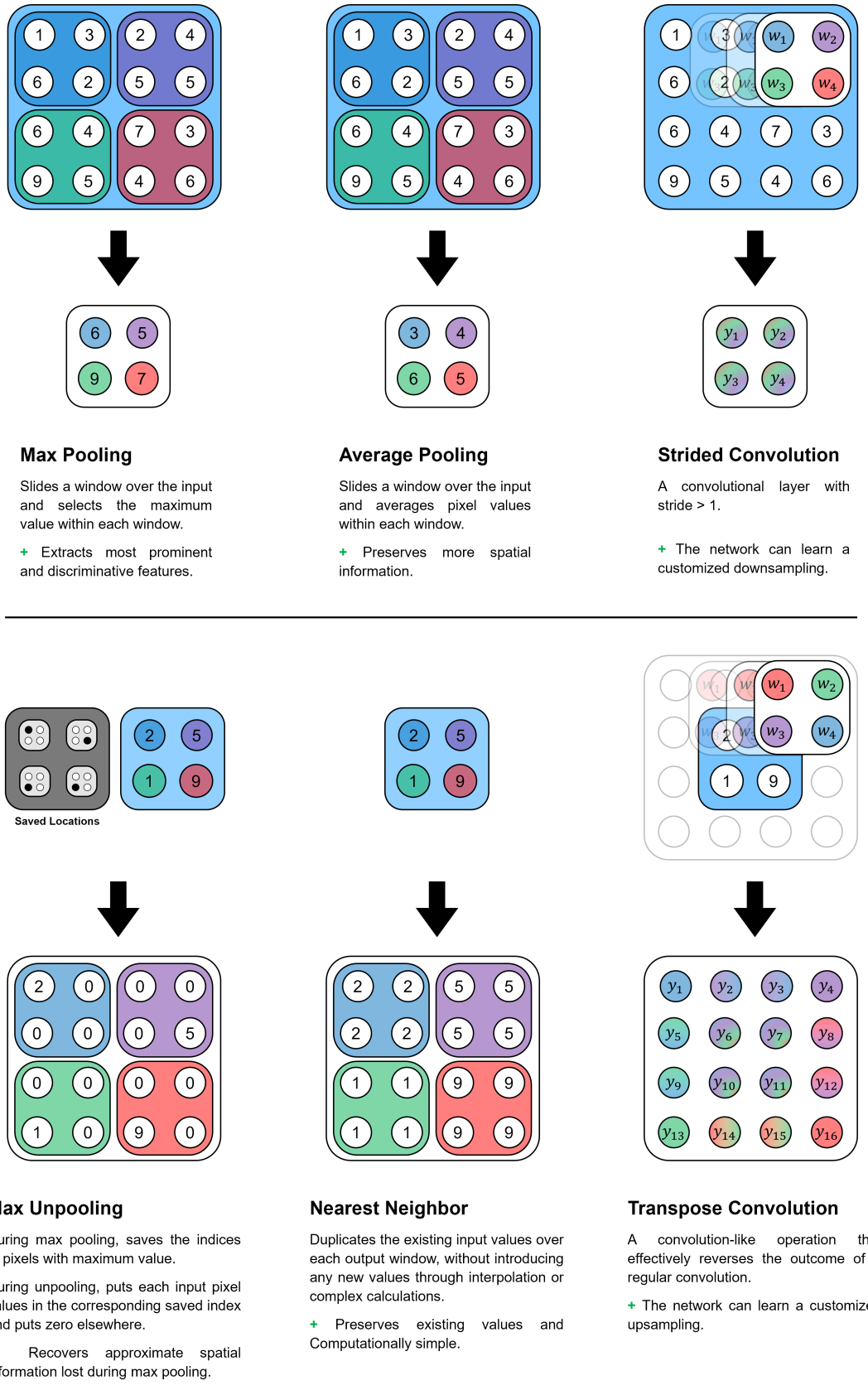

Fig. 3.8. Commonly used downsampling (top) and upsampling (bottom) layers in convolutional neural networks with the key advantage of each method.

3.1.3. Computer Scientists As Architects

One of the tasks of computer scientists is finding the proper neural network *architecture* for the specific task at hand. Architecture refers to the structure and design of a neural network, including the types, arrangement, and connectivity of its various layers. The choice of architecture has a significant impact on the performance and efficiency of the network, as different architectures are tailored to address specific problem types. In this section, a ConvNet architecture known as U-Net is introduced to illustrate how design choices can make a network suitable for specific tasks. It will also lay the foundations to introduce our high-dimensional inference model.

U-Net

Introduced in [**69**], a U-Net is a specialized ConvNet designed for image-to-image tasks, i.e., where both input and output are images. It consists of three main components: a *contracting path* that extracts information from the input through convolutional and downsampling layers, an *expanding path* that uses the extracted features to generate the output image through convolutional layers and up-sampling operations, as well as *skip connections* that connect corresponding layers between the contracting and expanding paths. The skip connections preserve spatial information and enhance the accuracy of the network by allowing a direct flow of information at each resolution. Figure 3.9 shows a neural network with U-Net architecture.

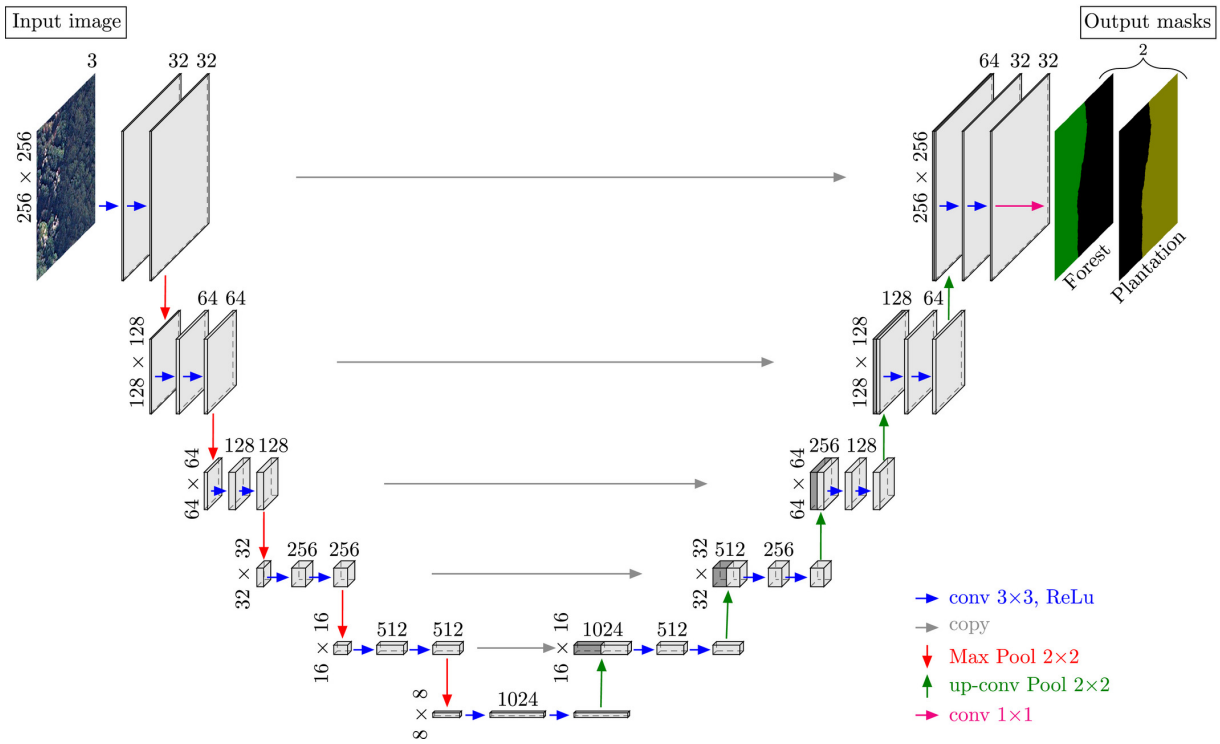

Fig. 3.9. A neural network with U-Net architecture. Credit: [**80**].

3.1.4. Training Neural Networks

So far, we have defined a neural network and discussed its various components. However, it is still unclear how neural network parameters (i.e., weights and biases) are "tuned" to yield the desired output. The process of achieving optimal values for these parameters is known as *training* and involves several components which are discussed subsequently.

Loss Function

The first component of training is the *loss function*, $\mathcal{L}(\hat{\boldsymbol{y}}, \boldsymbol{y})$. It has to be a **differentiable** function that measures the performance of the network. It receives the network's prediction $\hat{\boldsymbol{y}}$ and the desired output $\boldsymbol{y}$ for a training example and yields a quantitative comparison between them: The lower the loss function value, the closer the network is to optimal performance. With the loss function, one can find out how different sets of network parameters compare and which one(s) results in "better" network predictions[6].

The choice of loss function depends on the network's specific task. For a detailed discussion of different loss functions and their usages, see [**81**]. Two well-known loss function families for regression[7] tasks are Lp-norm and Negative Log-Likelihood (NLL) loss functions.

Lp-norm is a family of loss functions that measure the difference between predicted and target values based on the p-norm of the error:

$$\mathcal{L}_{\text{Lp}}(\hat{\boldsymbol{y}}, \boldsymbol{y}) := \frac{1}{p} \, \|\hat{\boldsymbol{y}} - \boldsymbol{y}\|_p^p \tag{3.2}$$

By selecting an appropriate value of p, Lp loss allows the model to prioritize different aspects of the error, leading to different characteristics in the model's behavior and sensitivity to outliers. One of the most famous choices is $p = 2$, more famously known as Mean Squared Error (MSE) loss:

$$\mathcal{L}_{\text{MSE}}(\hat{\boldsymbol{y}}, \boldsymbol{y}) := \frac{1}{2} \, \|\hat{\boldsymbol{y}} - \boldsymbol{y}\|_2^2 \tag{3.3}$$

Negative Log-likelihood loss functions are commonly used in regression tasks when the target variable follows a specific probability distribution. For instance, when the target is assumed to follow a Gaussian distribution, one can use the Gaussian NLL loss:

$$\mathcal{L}_{\text{GNLL}}(\hat{\boldsymbol{\mu}}, \hat{\Sigma}, \boldsymbol{y}) = \frac{1}{2}(\ln |\hat{\Sigma}| + (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^\top \hat{\Sigma}^{-1} (\boldsymbol{y} - \hat{\boldsymbol{\mu}})) \tag{3.4}$$

The above formulas are for the instance-level loss, i.e., they show how the loss function is calculated for *one* training example, $(\boldsymbol{x}, \boldsymbol{y})$. During training, the network's performance is usually evaluated based on several training examples (aka, a training *batch*). In this case,

[6] By enabling comparative assessment, the loss function *defines* what is considered as better.

[7] Regression is a statistical technique used to model and predict numerical values based on the relationship between variables.

the individual loss values are averaged together. This value will serve as an estimate of the expected loss over the entire training dataset $\mathcal{D}_{\text{train}} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$, i.e.,

$$\mathcal{J}(\boldsymbol{\theta}, \mathcal{D}_{\text{train}}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}_{\text{train}}}\Big[\mathcal{L}(\hat{\boldsymbol{y}}, \boldsymbol{y})\Big] \approx \frac{1}{b}\sum_{i=1}^{b}\mathcal{L}\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \boldsymbol{y}_i\Big), \tag{3.5}$$

where b is the batch size (i.e., the number of training examples in the batch) and $\boldsymbol{\theta}$ is the network's parameters. The learning objective of the network can be summarized in the following equation:

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta}, \mathcal{D}_{\text{train}}), \tag{3.6}$$

For a detailed discussion on batch training, see Chapter 8 of [**28**].

Optimization

Due to non-linear and complex relationships between neural network parameters and their output, determining optimal parameters through closed-form solutions is often impossible. Hence, one should use an *optimization algorithm* to traverse the *loss landscape*[8] and search for optimal parameters. Furthermore, since the loss functions are typically non-convex and high-dimensional, finding the global optimum is computationally infeasible. Therefore, approximate optimization methods are used to find suboptimal solutions that effectively minimize the loss function. Figure 3.10 displays a visualization of a loss landscape.

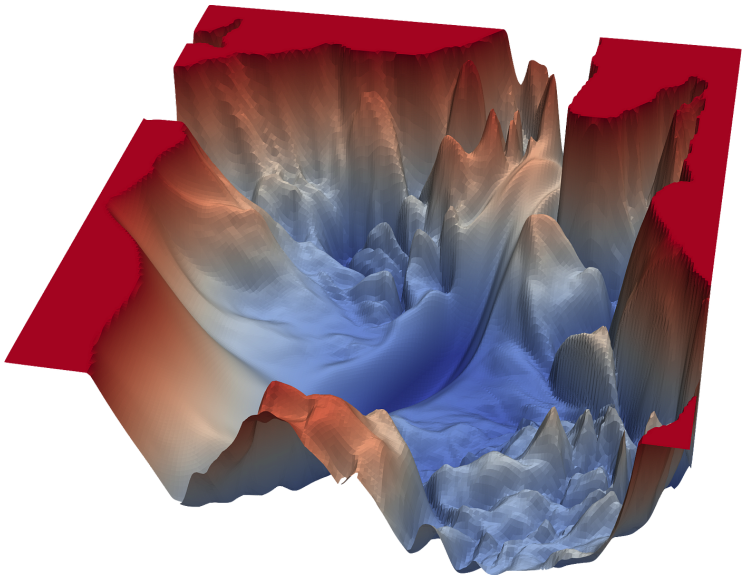

Fig. 3.10. A typical loss landscape of VGG-56 - a convolutional neural network variant primarily used for image classification tasks. For more information about how this plot was generated, see [**56**].

[8]Loss landscape refers to the geometric representation of the loss function in the high-dimensional space of a neural network's parameters. It provides a visualization of how the loss function changes with respect to different values of parameters.

An optimization algorithm adjusts the network's parameters to reduce the loss function and improve its performance. It iteratively updates the parameters based on the computed gradients or higher-order moments of the loss function with respect to the parameters. These methods are often accompanied by a *regularization* technique to improve generalization and prevent overfitting[9]. Regularization techniques introduce additional terms or constraints to the loss function, encouraging the network to have smaller weights, sparse solutions, or smoother decision boundaries.

The simplest optimization algorithm is called Gradient Descent (Algorithm 1), which updates the parameters in the direction opposite to the gradient of the loss function:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla \mathcal{J}(\boldsymbol{\theta}_t), \tag{3.7}$$

with η being a factor called the learning rate that determines the step size of parameter updates. It might be kept constant or adjusted using a predefined *schedule* throughout training.

Algorithm 1 Mini-Batch Gradient Descent Optimization Algorithm

Initialize parameters: $\boldsymbol{\theta}$
Initialize learning rate: η
Initialize number of epochs: E
Initialize batch size: b
for $e \leftarrow 1$ to E **do**
    Shuffle training data
    for $t \leftarrow 1$ to $\frac{\#\text{ training examples}}{B}$ **do**
        Randomly select mini-batch: mb
        Compute gradient using the examples in mb: $g_t \leftarrow \nabla \mathcal{J}(\boldsymbol{\theta}_t)$
        Update parameters: $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta\, g_t$
    end for
end for

Other optimization algorithms include adaptive learning rate methods (e.g., AdaGrad, RMSprop, Adam), which dynamically adjust the learning rate during training to accelerate convergence and handle different scales of gradients; momentum-based methods, which incorporate a momentum term that accelerates the optimization process by accumulating gradients over time; and second-order optimization methods which consider information such as Hessian matrix to update parameters. For a detailed discussion about optimization algorithms, see Chapter 8 of [**28**].

[9]Overfitting happens when a machine learning model performs too well on training data but fails to generalize to new, unseen data.

The convergence of training depends on several choices: optimization algorithm and its *hyperparameters*[10], initial values of the parameters, batch size, and the employed regularization technique. It is left to the designer to make appropriate choices based on the problem and "tune" the corresponding hyperparameters using techniques like grid search, random search, and Bayesian optimization to reach an effective training configuration.

Backpropagation

The last piece of training is a method to compute the loss function's gradient with respect to each weight. The most widely used and effective method for this task is known as *backpropagation*. It relies on the chain rule of derivatives to efficiently compute gradients layer by layer. It starts from the output layer and propagates the gradients backward through the network. For a detailed discussion on backpropagation, see Section 6.5 of [**28**].

3.2. Probabilistic World

The introduction of probability theory revolutionized our perception of the world, fundamentally altering how we analyze and interpret events with varying outcomes. It also had a profound impact on physical models. Before its emergence, deterministic models assumed absolute certainty and predictability in physical events. The probability theory revolutionized this perspective by acknowledging the inherent uncertainty and variability in physical systems. In this section, we explore how exceptional modeling capabilities of neural networks can be combined with the principles of probability theory to construct a powerful inference framework for physical problems.

3.2.1. Deep Probabilistic Models

Deep probabilistic models leverage the remarkable expressive power of neural networks to model probabilistic relationships. As illustrated by Figure 3.11, deep probabilistic models can be trained to predict probability, estimate probability density, or directly sample from a probability distribution. Our model falls under the last category (also known as *deep generative models*), which we will delve into further to explore their characteristics.

Generative Models

Generative models are designed to generate new data samples that resemble their training data by capturing their underlying probability distribution[11]. Figure 3.12 showcases some of the most prominent deep generative models.

[10]In machine learning, hyperparameters are externally set configuration settings that impact the behavior and performance of a learning algorithm, in contrast to "parameters" which are internal and learned from the training data. A typical example of a hyperparameter is the learning rate and its schedule (i.e., how it varies throughout training).

[11]This distribution can possibly be the posterior distribution of some physical parameters.

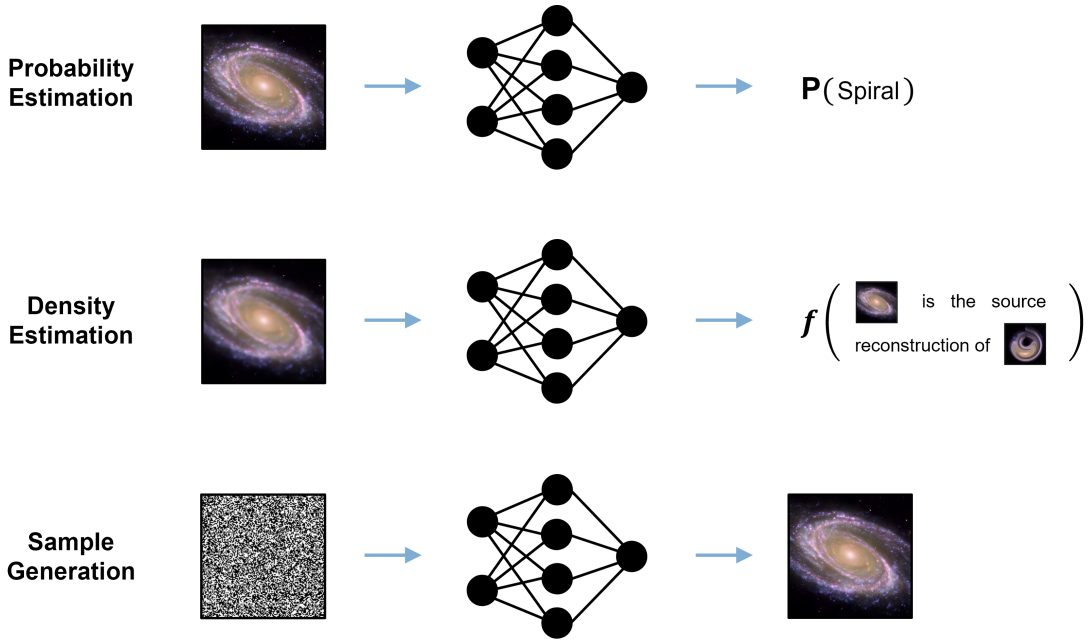

Fig. 3.11. Applications of deep probabilistic models.

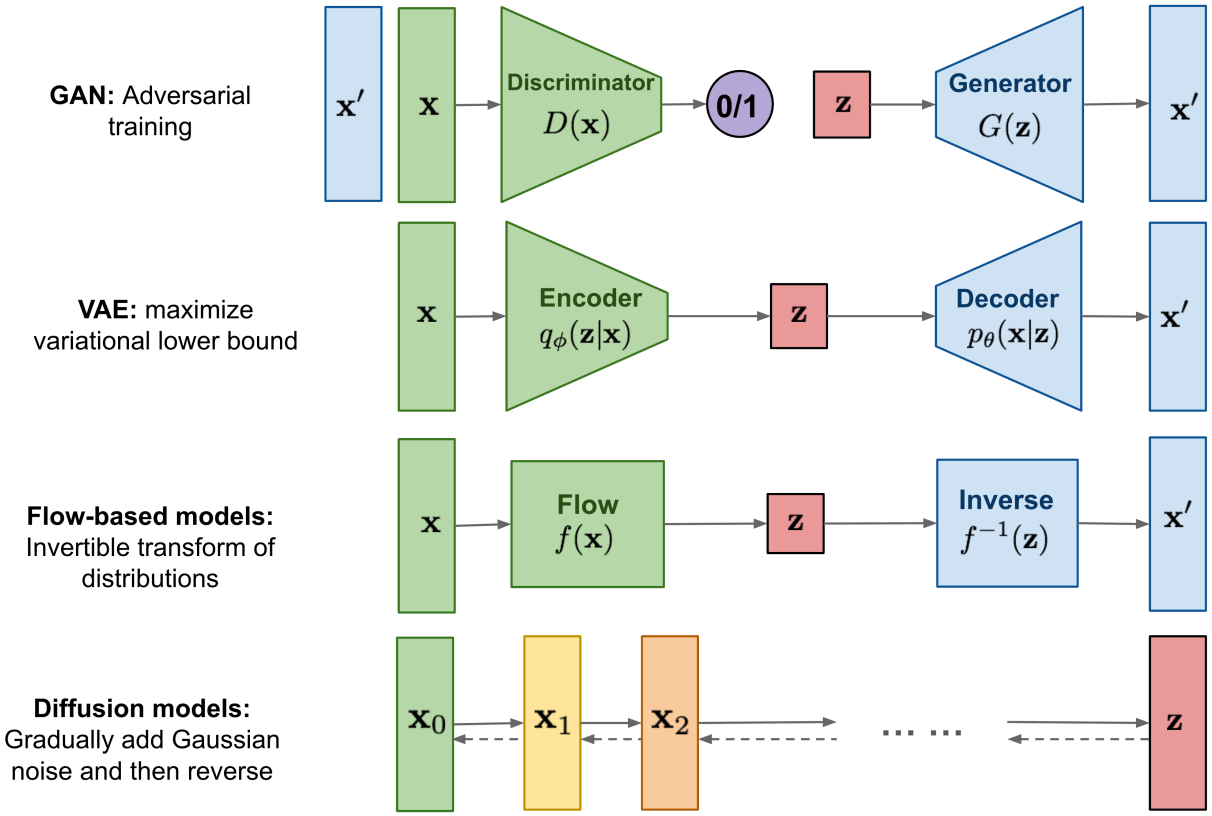

Fig. 3.12. Overview of deep generative models. Credit: Lilian Weng.

Here is how each of the models presented in Figure 3.12 work:

(1) **Variational AutoEncoders (VAEs)** learn the underlying data distribution using an encoder-decoder architecture. They approximate data distribution using a technique called variational inference. We will shortly investigate VAEs in more detail.

(2) **Generative Adversarial Networks (GANs)** consist of two neural networks, a generator and a discriminator. The generator transforms random noise into synthetic data samples, while the discriminator tries to distinguish between real and synthetic samples. Through an adversarial training process, the generator learns to produce increasingly realistic data by competing and improving against the discriminator.

(3) **Normalizing Flows** approximate complex probability distributions by applying a series of invertible transformations to a simple base distribution. These transformations progressively warp the base distribution to capture the characteristics of the target distribution.

(4) **Diffusion Models** iteratively refine a noise source until it closely approximates the target data distribution. This is done through a process known as the diffusion process, where noise is gradually added to the initial input.

3.2.2. VAEs, A Closer Look

In this subsection, we delve into the underlying theoretical framework of VAEs. It forms the foundation of our inference model. We start with the manifold hypothesis, the core assumption of VAEs. Subsequently, we introduce the approximate inference framework employed to train VAEs. We finish by introducing Conditional VAEs, which enable learning conditional posteriors.

Manifold Hypothesis

High-dimensional data, such as images, suffer from overparameterization when represented in pixels. Consider a set of 100×100 human face images, which correspond to a 10000-dimensional space, called *data space*. If we were to pick an image and randomly modify its pixel values, it is highly likely that we would deviate from the distribution of valid human faces and end up with a meaningless image. This phenomenon amplifies as the number of pixels increases. In other words, the exponential growth of possible configurations with increasing dimensions poses challenges for accurately modeling the underlying data distribution. This phenomenon can be regarded as an instance of the *curse of dimensionality*.

However, it is reasonable to assume that within this vast high-dimensional space, there exists a manifold - a lower-dimensional structure - that captures the essence of valid human faces. By traversing this manifold, we can explore points that correspond to different facial

features, expressions, and shapes while still resembling a human face. This assumption is known as the *manifold hypothesis*. The existence of such a manifold and the ability to access it are essential for VAEs to generate new images that represent data distribution.

The latent manifold is modeled within a lower-dimensional space, called *latent space*, and is defined by two transformations: an encoder and a decoder. They convert images between the data space and the latent space. The optimal encoder-decoder pair to access the latent manifold is initially unknown and needs to be identified among a given family of potential transformations. This can be achieved by aiming to preserve maximum information during the encoding process, resulting in minimal reconstruction error when decoding the corresponding latent representation[12]. Figure 3.13 summarizes the latent space concepts and terminology. We often denote data space variables by $\boldsymbol{x}$ and latent space variables by $\boldsymbol{z}$.

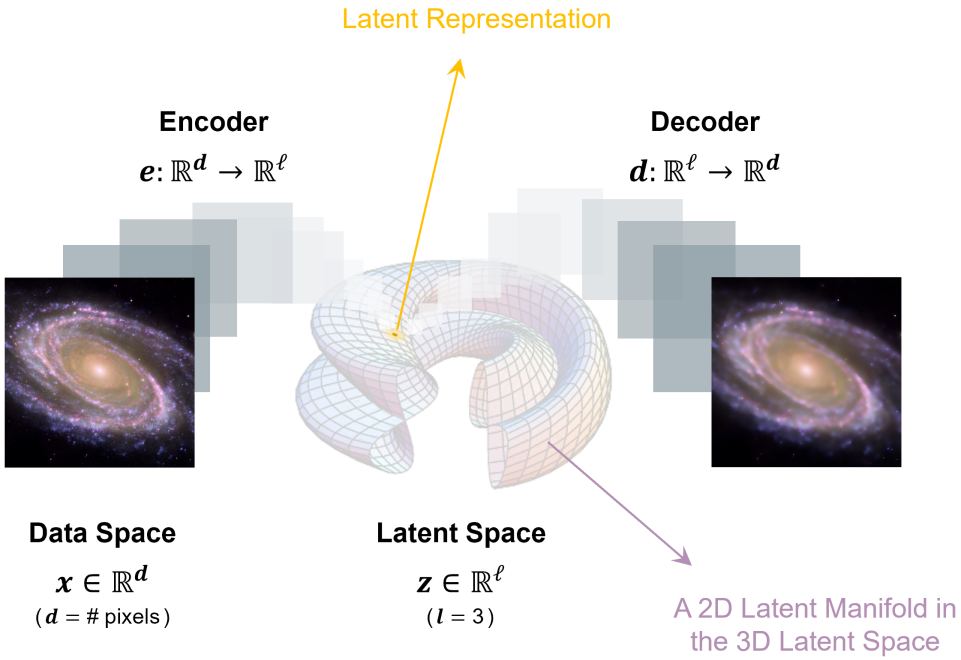

Fig. 3.13. Latent space concepts and terminology.

Probabilistic Framework

In the realm of machine learning, one approach to learning the latent manifold is through Variational Autoencoders (VAEs). VAEs combine the power of neural networks with probabilistic inference to learn the encoding and decoding transformations. To generate a data sample using a VAE, one must first draw a sample from a prior distribution defined in latent space and then use a neural network (i.e., VAE's decoder) to map the generated sample to data space. The prior distribution and the decoder together model the joint distribution of

[12]The representation of an image in latent space is called a *latent representation*.

the data space and latent space variables, $p(\boldsymbol{x},\boldsymbol{z})$, which can be decomposed as

$$p(\boldsymbol{x},\boldsymbol{z}) = p(\boldsymbol{z})\ p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}), \tag{3.8}$$

where the prior $p(\boldsymbol{z})$ is a simple distribution in the latent space, and the likelihood $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ is modeled using the decoder. The subscript $\boldsymbol{\theta}$ represents the decoder's parameters. The VAE's objective is to learn $\boldsymbol{\theta}$, such that the marginal likelihood[13] (i.e., evidence),

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x},\boldsymbol{z})\ d\boldsymbol{z} = \int p(\boldsymbol{z})\ p_{\theta}(\boldsymbol{x}|\boldsymbol{z}) d\boldsymbol{z}, \tag{3.9}$$

of training data is maximized.

Approximate Inference

It appears that we have every required element to train a VAE. However, there is one caveat: The integral in Equation 3.9 is intractable; in other words, it is too computationally expensive to integrate over all possible values of the latent variables. Hence, accessing the latent representation of data samples by directly maximizing evidence is not feasible. For the same reason, we cannot use the Bayes theorem to calculate the posterior distribution of latent variables,

$$p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) = \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{x})}, \tag{3.10}$$

since $p(\boldsymbol{x})$ appears in the denominator. To address this issue, one can employ approximate inference techniques. One such approach is variational inference[14], wherein a variational distribution, denoted as $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$, approximates the actual posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$. The approximate posterior is modeled by a neural network (i.e., VAE's encoder) parametrized by $\boldsymbol{\phi}$[15]. By doing so, the challenge of dealing with an intractable integral is replaced with an optimization problem aimed at obtaining an optimal approximation for the posterior. The objective function is derived by establishing a lower bound on $p(\boldsymbol{x})$, maximizing which ensures enhancement of $p(\boldsymbol{x})$. The derivation of this lower bound starts by introducing and subtracting a term from the logarithm of $p(\boldsymbol{x})$:

$$\ln p(\boldsymbol{x}) = \ln p(\boldsymbol{x}) + \int q(\boldsymbol{z}|\boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})}\right) d\boldsymbol{z} - \int q(\boldsymbol{z}|\boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})}\right) d\boldsymbol{z} \tag{3.11a}$$

$$= \int q(\boldsymbol{z}|\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{z} + \int q(\boldsymbol{z}|\boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})}\right) d\boldsymbol{z} - \int q(\boldsymbol{z}|\boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})}\right) d\boldsymbol{z} \tag{3.11b}$$

[13]It is called marginal likelihood since it is derived by marginalizing out the latent variables from $p(\boldsymbol{x},\boldsymbol{z})$.
[14]The name refers to the optimization problem formulation that involves finding the best approximation by minimizing a *variational* divergence or distance measure.
[15]The encoder can model a family of variational distributions. Each combination of its parameters corresponds to a member of this family.

$$= \int q(\boldsymbol{z}|\boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})}\right) d\boldsymbol{z} - \int q(\boldsymbol{z}|\boldsymbol{x}) \left[\ln\left(\frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})}\right) - \ln p(\boldsymbol{x})\right] d\boldsymbol{z} \tag{3.11c}$$

$$= \int q(\boldsymbol{z}|\boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})}\right) d\boldsymbol{z} - \int q(\boldsymbol{z}|\boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{x},\boldsymbol{z})}{p(\boldsymbol{x})q(\boldsymbol{z}|\boldsymbol{x})}\right) d\boldsymbol{z} \tag{3.11d}$$

$$= \int q(\boldsymbol{z}|\boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})}\right) d\boldsymbol{z} - \int q(\boldsymbol{z}|\boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{x})}\right) d\boldsymbol{z} \tag{3.11e}$$

We can use the definition of the Kullback–Leibler (KL) divergence[16] to express the second term of Equation 3.11e as $-D_{\mathrm{KL}}\Big(q(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z}|\boldsymbol{x})\Big)$. By designating the first term as $\mathrm{ELBO}(q,p)$, we can rewrite the equation as

$$\ln p(\boldsymbol{x}) = \mathrm{ELBO}(q,p) + D_{\mathrm{KL}}\Big(q(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z}|\boldsymbol{x})\Big). \tag{3.12}$$

Since KL divergence is always greater than or equal to zero, we can conclude that

$$\mathrm{ELBO}(q,p) \leq \ln p(\boldsymbol{x}). \tag{3.13}$$

$\mathrm{ELBO}(q,p)$ is called the Evidence Lower Bound, and its negative $\mathcal{L}_{\mathrm{ELBO}} := -\mathrm{ELBO}(q,p)$ can serve as a loss function for VAEs. With some mathematical manipulation, we can decompose $\mathcal{L}_{\mathrm{ELBO}}$ into two easily interpretable terms:

$$\mathcal{L}_{\mathrm{ELBO}} = -\int q(\boldsymbol{z}|\boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})}\right) d\boldsymbol{z} \tag{3.14a}$$

$$= -\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\Big[\ln p(\boldsymbol{x},\boldsymbol{z}) - \ln q(\boldsymbol{z}|\boldsymbol{x})\Big] \tag{3.14b}$$

$$= -\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\Big[\ln p(\boldsymbol{x}|\boldsymbol{z}) + \ln p(\boldsymbol{z}) - \ln q(\boldsymbol{z}|\boldsymbol{x})\Big] \tag{3.14c}$$

$$= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\Big[-\ln p(\boldsymbol{x}|\boldsymbol{z})\Big] + D_{\mathrm{KL}}\Big(q(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z})\Big) \tag{3.14d}$$

$$= \mathcal{L}_{\mathrm{rec}} + \mathcal{L}_{\mathrm{KL}} \tag{3.14e}$$

The first term is called the reconstruction term. It evaluates the fidelity of generated data to the original input. The second term acts as a regularizer. It encourages smooth and structured latent space representations by penalizing the divergence between prior and approximate posterior. To further reduce computational costs, one typically uses only one latent sample to estimate the expectation $\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\,\cdot\,]$.

Figure 3.14 illustrates the training and sampling processes of a VAE. During sampling, a random sample is drawn from the prior distribution of latent variables and fed into the decoder to generate a sample from the data distribution. In the training phase, the encoder

[16] KL divergence is a measure of the difference between two probability distributions.

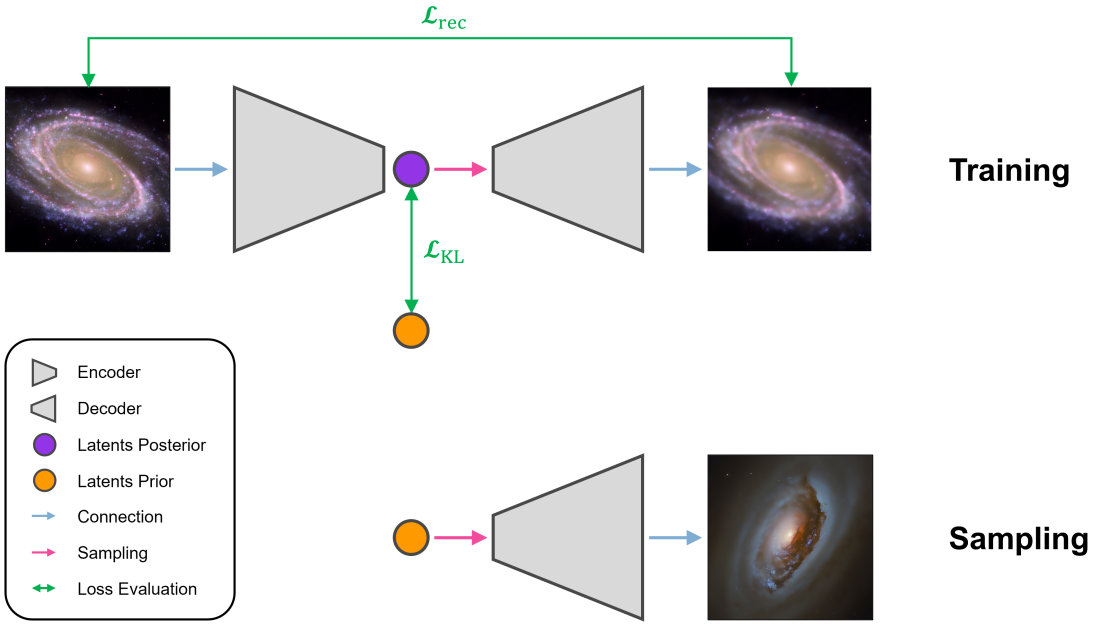

Fig. 3.14. VAE's training & sampling processes.

is supplied with data examples from the training set to predict the latents' posterior distribution, followed by drawing samples from the posterior and passing them to the decoder. Subsequently, the decoder's outputs are compared with the inputs to calculate the reconstruction loss $\mathcal{L}_{\text{rec}}$. This loss is combined with $\mathcal{L}_{\text{KL}}$, the KL divergence between the prior and posterior distributions of the latents, to form the total loss function $\mathcal{L}_{\text{ELBO}}$. The loss function is then used for backpropagation, updating both the decoder's and the encoder's parameters.

Need for Conditions

A VAE has the ability to generate new examples that represent its training data. However, when data examples can be divided into different classes, the VAE cannot generate examples belonging to a specific class. The generation process in a VAE is solely driven by latent variables and is independent of any specific conditions or attributes. Hence,

- A VAE trained on the MNIST dataset[17] cannot generate specific digits.
- A VAE trained on human faces cannot generate faces with a given hair color.
- A VAE trained on galaxy images cannot generate galaxy images that can be regarded as valid source reconstructions of a particular lens-source system.

[17]The MNIST dataset [**52**] is a collection of handwritten digit images commonly used for training and testing machine learning algorithms in the field of computer vision. It consists of 60,000 training images and 10,000 testing images, each labeled with its represented digit.

To generate or manipulate data based on specific characteristics, a VAE must incorporate additional conditional information during its training and sampling processes. In a Conditional VAE (cVAE), the encoder and decoder take conditional information in addition to their regular input. This leads to a more structured latent space where features relevant to the provided conditional information are organized and localized in distinct regions.

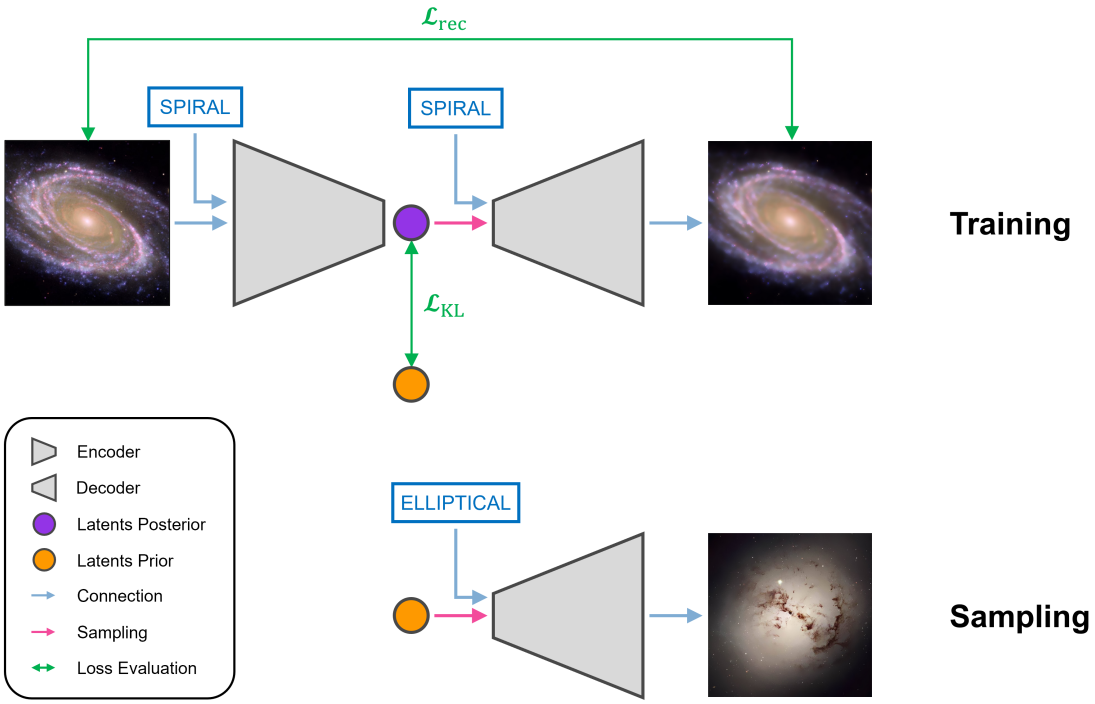

Fig. 3.15. cVAE's training & sampling processes.

Figure 3.15 illustrates the training and sampling processes of a cVAE. They are identical to that of a VAE, except that this time the encoder and decoder are conditioned on a class variable. In other words, they receive an additional label as input that specifies the category to which their input belongs.

3.3. U-Nets Can Be Uncertain

In Section 3.1, the U-Net architecture was presented as a way to learn "deterministic" high-dimensional mappings. Furthermore, VAEs and cVAEs were discussed in detail in Section 3.2. A cVAE can generate new examples from random noise based on given conditional information. In many inference tasks in astrophysics, the goal is to reconstruct some physical parameters[18] $\boldsymbol{y}$ based on observations $\boldsymbol{x}$, where both observations and parameters are

[18]Note that parameters here differs from neural network parameters. It means physical quantities of interest that we aim to infer.

many-dimensional[19] images. In these problem spaces, inputs (or observations) are often not sufficient to narrow down predictions to a single acceptable answer[20]. Instead, there exists a manifold of consistent parameters for each observation.

An appropriate inference model for these tasks is one capable of learning the manifold of acceptable answers and using it to produce consistent samples within the parameter space. The model must learn high-dimensional mappings while accurately accounting for uncertainties and variations in its predictions. To achieve this objective, a Hierarchical Probabilistic U-Net (HPU-Net) [**49**] is constructed by merging the U-Net architecture with the cVAE framework. This fusion results in a model capable of extracting information across multiple scales, encoding complex probability distributions, and learning high-dimensional probabilistic mappings. This combination offers a powerful tool for posterior sampling.

Building HPU-Nets upon VAEs offers two advantages compared to other prominent deep generative models. First, VAEs can be trained more efficiently on large, high-dimensional datasets. In contrast, diffusion models, for example, require a large number of iterations to generate a sample. Additionally, VAEs have a well-understood theoretical framework, which equips HPU-Nets with a sound mathematical formalism to describe their probabilistic behavior. This especially proves useful in physical applications, where it is important to learn probability distributions in a principled and robust way. The architecture and training process of HPU-Nets are described in the following subsections.

3.3.1. Architecture

Figure 3.16 describes the HPU-Net architecture. The model's contracting path is exactly the same as a regular U-Net. However, in the expanding path, some of the scales in the hierarchy are equipped with a latent space with the same dimensionality as the corresponding feature maps. At those *latent scales*, there exist three additional steps after regular convolutional layers and before upsampling. First, additional convolutional layers will predict the required parameters to sample from the latent space[21]. Then, a sample will be drawn from the latent space, which will finally be appended to the set of existing feature maps at that scale. The latent spaces enable the model to quantify uncertainties at different resolutions and exhibit probabilistic behavior.

[19]There is a subtlety in using the term "dimension" in our discussion. Although a 32×32 image is referred to as a 2-dimensional image, it is important to note that each pixel within the image can be viewed as a distinct physical parameter. If we were to analyze and infer the pixel values, we would essentially be working within a 1024-dimensional space, with each dimension representing a unique pixel.
[20]This can be due to noisy observations, lack of information, or inherent uncertainties.
[21]For example, if we choose the latent spaces to be pixel-wise Gaussian distributions, the convolutional layers would predict the means and standard deviations per pixel.

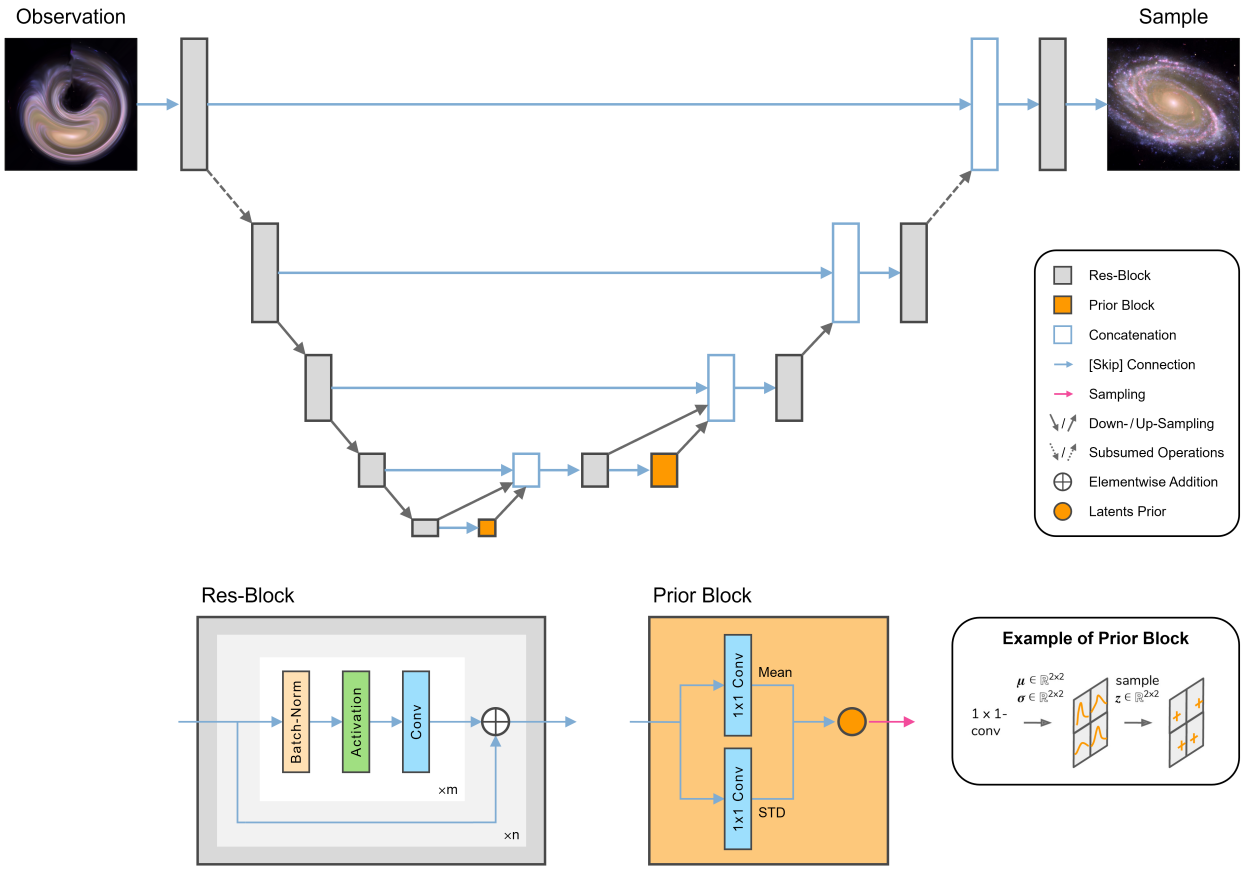

Fig. 3.16. HPU-Net architecture. Figure adapted from [**49**] with modifications.

3.3.2. Probabilistic Framework

Before discussing the training procedure of HPU-Nets, it is essential to explore their underlying mathematical foundation. Generating a prediction using an HPU-Net involves drawing a sample from a prior distribution defined in latent space and decoding it to the parameter space. Unlike vanilla VAEs, the prior is no longer a simple distribution. Instead, it should be learned, spans over several scales of the network, and is conditioned on observations. Given the hierarchical structure of latent spaces, the prior factorizes as

$$p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_L|\boldsymbol{x}) = p(\boldsymbol{z}_L|\boldsymbol{z}_{<L}, \boldsymbol{x}) \cdot \ldots \cdot p(\boldsymbol{z}_1|\boldsymbol{x}), \tag{3.15}$$

where L is the total number of latent spaces. Each factor is modeled by the contracting path and the skip connections and expanding path components that affect the corresponding latent space. Figure 3.17 illustrates how different components of the network contribute to modeling each factor of the prior.

The skip connections and the expanding path serve an additional purpose, which is decoding latent variables into a prediction of parameters. The network components together model

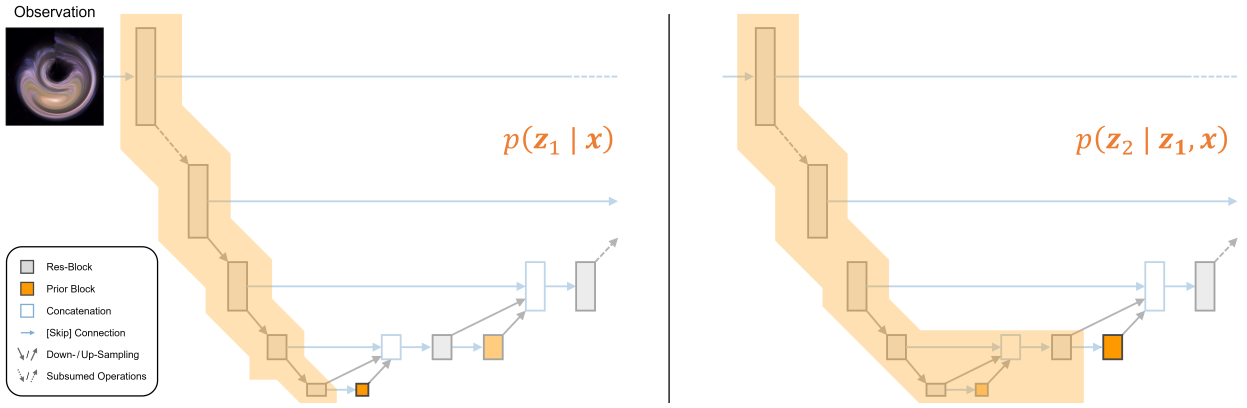

Fig. 3.17. Implementation of different factors of the prior distribution in HPU-Net.

the joint distribution of parameters $\boldsymbol{y}$ and latent variables $\boldsymbol{z}$, conditioned on observations $\boldsymbol{x}$,

$$p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\; p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{x}), \tag{3.16}$$

where $\boldsymbol{\theta}$ refers to the network's parameters. The HPU-Net's objective is to learn $\boldsymbol{\theta}$ such that the marginal likelihood of parameters $\boldsymbol{y}$ - conditioned on observations $\boldsymbol{x}$ - is maximized[22]:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \int p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{x})\; d\boldsymbol{z} = \int p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\; p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{x})\; d\boldsymbol{z} \tag{3.17}$$

The integral in Equation 3.17 is intractable, rendering it difficult to access the posterior distribution of latent variables $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x})$. Similar to VAEs, however, we can use a variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x})$ to approximate the latents' posterior. This approximate posterior is modeled by a neural network which is used during training and will be introduced shortly. It is possible to use the approximate posterior to derive an ELBO. The derivation starts by adding and subtracting a term from $\ln p(\boldsymbol{y}|\boldsymbol{x})$ and mathematically manipulating the result:

$$\begin{aligned}
\ln p(\boldsymbol{y}|\boldsymbol{x}) &= \ln p(\boldsymbol{y}|\boldsymbol{x}) + \int q(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x})}\right) d\boldsymbol{z} - \int q(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x})}\right) d\boldsymbol{z} \\
&= \int q(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x}) \ln p(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{z} + \int q(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x})}\right) d\boldsymbol{z} - \int q(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x})}\right) d\boldsymbol{z} \\
&= \int q(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x})}\right) d\boldsymbol{z} - \int q(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x}) \left[\ln\left(\frac{p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x})}\right) - \ln p(\boldsymbol{y}|\boldsymbol{x})\right] d\boldsymbol{z} \\
&= \int q(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x})}\right) d\boldsymbol{z} - \int q(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{y}|\boldsymbol{x}) q(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x})}\right) d\boldsymbol{z}
\end{aligned} \tag{3.18a}$$

[22] In Bayesian inference terminology, $p(\boldsymbol{y}|\boldsymbol{x})$ is the posterior distribution of parameters.

$$= \int q(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x})}\right) d\boldsymbol{z} - \int q(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x})}\right) d\boldsymbol{z} \tag{3.18b}$$

$$= \mathrm{ELBO}(q,p) + D_{\mathrm{KL}}\Big(q(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x}) \parallel p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x})\Big) \tag{3.18c}$$

Since KL divergence is always greater than or equal to zero, we can conclude that:

$$\mathrm{ELBO}(q,p) \leq \ln p(\boldsymbol{y}|\boldsymbol{x}). \tag{3.19}$$

The negative of ELBO $\mathcal{L}_{\mathrm{ELBO}} := -\mathrm{ELBO}(q,p)$ can serve as a loss function for HPU-Nets. Similar to VAEs, $\mathcal{L}_{\mathrm{ELBO}}$ can be decomposed into the reconstruction and regularization terms:

$$\mathcal{L}_{\mathrm{ELBO}} = -\int q(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x})}\right) d\boldsymbol{z} \tag{3.20a}$$

$$= -\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x})}\Big[\ln p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x}) - \ln q(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x})\Big] \tag{3.20b}$$

$$= -\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x})}\Big[\ln p(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{x}) + \ln p(\boldsymbol{z}|\boldsymbol{x}) - \ln q(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x})\Big] \tag{3.20c}$$

$$= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x})}\Big[-\ln p(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{x})\Big] + D_{\mathrm{KL}}\Big(q(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x}) \parallel p(\boldsymbol{z}|\boldsymbol{x})\Big) \tag{3.20d}$$

$$= \mathcal{L}_{\mathrm{rec}} + \mathcal{L}_{\mathrm{KL}} \tag{3.20e}$$

3.3.3. Training

Figure 3.18 illustrates the training process of an HPU-Net. As mentioned in the last subsection, an additional network is required to model the latents' variational posterior. This network called the *Posterior Net*, is only used during training and has almost the same structure as the network used for sampling (i.e., the *Prior Net*), with two exceptions: 1) the Posterior Net receives both the observation and the actual value of the parameter as input, and 2) since the sole purpose of the Posterior Net is drawing samples from the variational posterior, it has a truncated decoder. In other words, the layers whose **only** purpose is decoding latents are not present in the Posterior Net.

During training, both the Prior Net and Posterior Net predict the required parameters to sample from their latent spaces. However, samples are only drawn from the Posterior Net's latents and injected in place of the Prior Net's latent samples. Revisiting Equation 3.20, the reconstruction term is focused on assisting the networks to generate authentic predictions, and the KL term is focused on enriching the Prior Net latents by incorporating the information encoded in the posterior net. It also prevents the Posterior Net from overfitting to the training data.

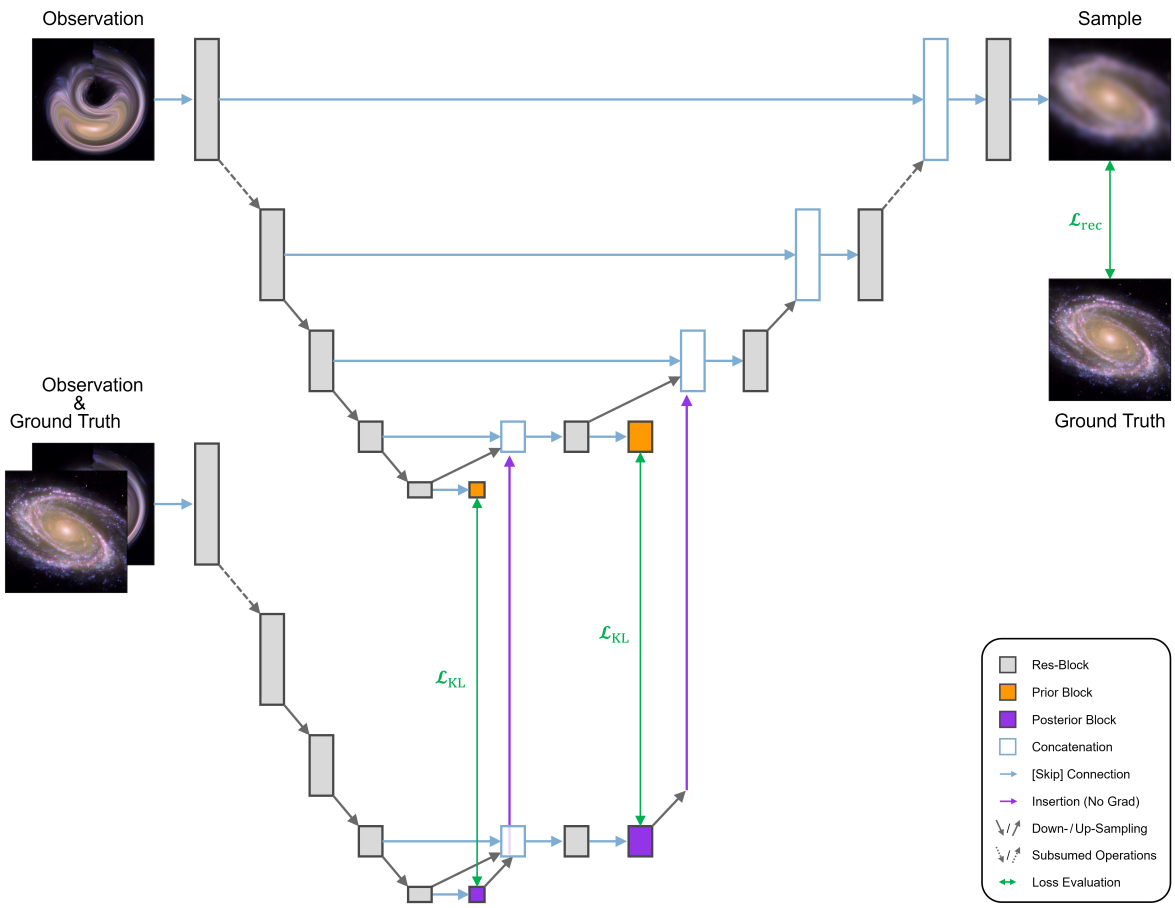

Fig. 3.18. HPU-Net's training process.

3.4. Evaluating Samples

Once a posterior sampling model is trained, we are interested to assess the fidelity of its encoded posterior to the true posterior distribution. In principle, two probability distributions are equal if and only if their probability density functions (PDFs) or cumulative distribution functions (CDFs) are equal. However, the model only generates posterior samples, and there is no way to access neither the PDF nor the CDF. To make matters worse, the only information we often have about the true posterior is just "one" posterior sample, corresponding to the true parameter value in the test dataset. Depending on the available information, various methods can be employed to evaluate the learned posterior distribution. In this study, we rely on the following methods to assess the statistical performance of the HPU-Net:

(1) **Comparing Moments**: The moments of a probability distribution are statistical measures that describe various aspects of its shape and characteristics. By comparing the moments of learned and true posteriors, we can see how the model's central tendency, variability, and other shape-related characteristics align with the true values. This method is applicable to two scenarios:

(a) If the true posterior is accessible, we can directly compare moments between the learned and true posteriors. This matching can be conducted up to any order of moments allowed by the number of posterior samples generated by the model.

(b) The true parameter value (i.e., ground truth, GT) can serve as a point estimate of the true posterior's mean, enabling a first-order moment comparison between the true and learned posteriors.

(2) **Assessing Power Spectrum**: Power spectrum provides insights into the frequency content of a signal. In cosmology, the power spectrum of cosmic structures (e.g., galaxies or dark matter) or the CMB is a fundamental prediction of various cosmological models. It is constrained by cosmological parameters and provides information about the distribution of matter and radiation across different scales. By comparing the power spectra of the model's predictions with the target power spectrum, one can evaluate the model's accuracy in generating maps with correct spatial features. This will complement the findings from comparing moments.

(3) **Coverage Probability Test:** This test measures the accuracy of the interval estimates provided by the model. It evaluates the model's uncertainty estimates to determine whether they are calibrated. We will shortly discuss this test in detail.

3.4.1. Coverage Probability Test

When employing statistical analysis to infer an unknown parameter, interval estimates provide more informative insights compared to relying solely on point estimates. In Bayesian statistics, a *credible region* defines a range of plausible values for a parameter at a designated confidence level. For instance, a credible region with 60% *credibility level* indicates that the true parameter will be within this range with a 60% probability. This implies that across multiple Bayesian analyses conducted on distinct datasets from the same population, the true parameter is expected to fall within the credible region approximately 60% of the time[23]. One measure of the learned posterior's quality is the accuracy of its credible regions, i.e., whether they capture the true parameter with the correct frequency or not. This is evaluated using the coverage probability test through the following steps:

(1) Use the learned posterior to form a large number of different credible regions with the **same** credibility level.

(2) Calculate the fraction of times the true parameter falls within the credible region to estimate the *coverage probability*, i.e., the probability that the credible region covers the true parameter.

[23]From this definition, it is evident that credible regions are not unique. For a given probability distribution, there exists an infinite number of credible regions with a certain credibility level.

(3) Compare the coverage probability (CP) with the credibility level (CL). If they match, we conclude that the model is calibrated. If CP $>$ CL (i.e., the credible regions are covering the true parameter more frequently than they should), we call the model *underconfident* or *conservative*, meaning that it is constructing larger-than-expected credible regions. If CP $<$ CL, the model is called *overconfident*. Figure 3.19 illustrates what calibrated, conservative, and overconfident models look like.

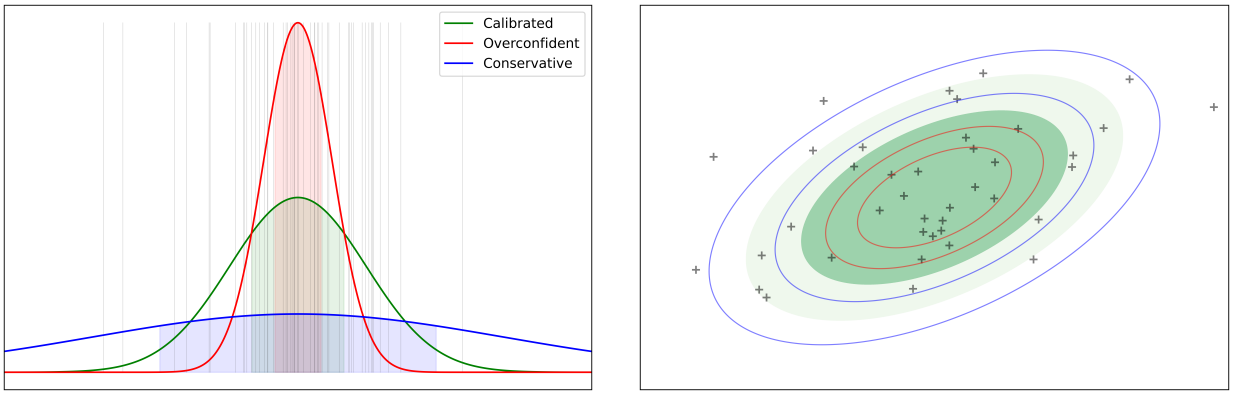

Fig. 3.19. Illustration of calibrated, overconfident, and conservative models in 1D (left) and 2D (right). Each subplot displays 40 samples from the target distribution—vertical lines in 1D and plus signs in 2D. Calibrated, overconfident, and conservative models are denoted by green, red, and blue, respectively. Shaded regions (1D) and inner ellipsoids (2D) indicate 50% confidence intervals. Calibrated models capture around half of the samples, while overconfident and conservative models fail to encompass the correct fraction.

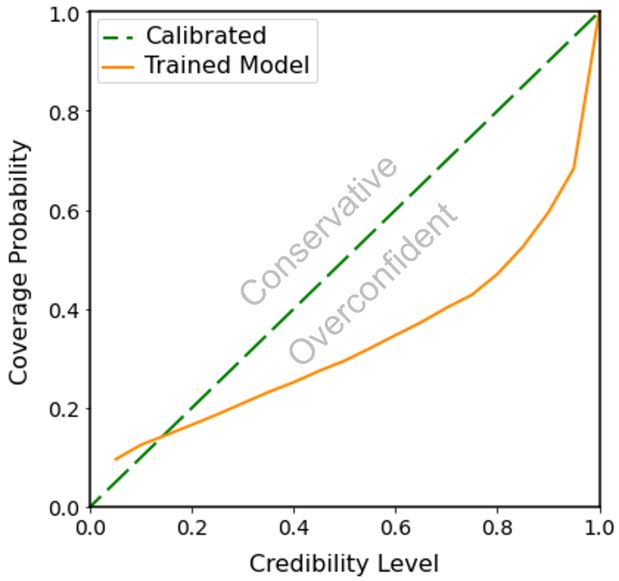

Fig. 3.20. Example of a coverage probability curve.

By repeating this process for regions with different credibility levels, one can generate a coverage probability curve (see Figure 3.20 for an example), illustrating the calibration performance across various credibility levels. For a calibrated model, the curve will resemble a diagonal line.

The specific steps of the coverage test may differ depending on the circumstances. For example, numerous ways exist to define credible regions. Furthermore, how we calculate or estimate the credibility level hinges on whether we have direct access to posterior densities or are limited to posterior samples. Additionally, the computational costs significantly influence the test's design in high-dimensional scenarios. In this study, we will use the Test of Accuracy with Random Points (TARP) [**54**] to perform the coverage probability test. This test estimates the credibility level based on the posterior samples generated by the model; hence, it is suitable for assessing HPU-Net's performance. Furthermore, dealing with distances rather than probability densities makes the test feasible for high-dimensional problems. The test involves the following steps (see Figure 4.4 for a visual summary):

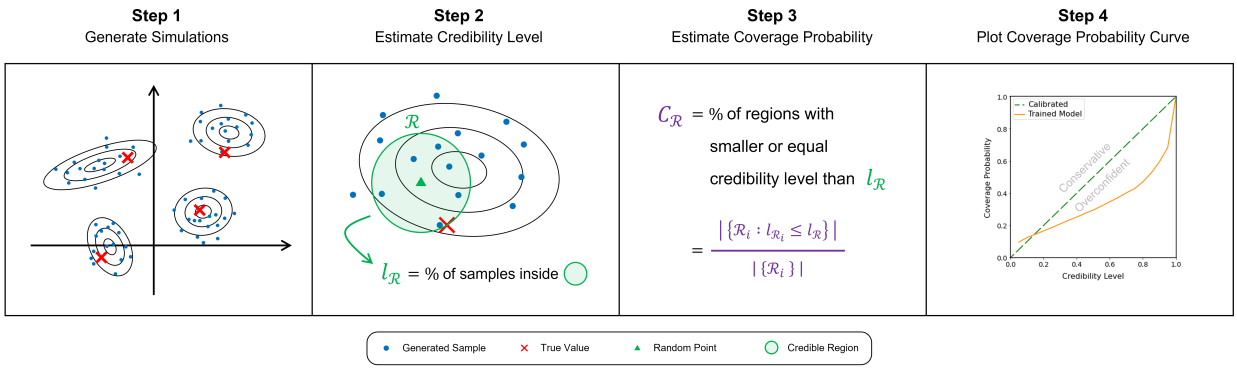

Fig. 3.21. TARP coverage probability test in four steps.

(1) Take a set of input-GT pairs (i.e., a test set). Feed the HPU-Net with the inputs to generate k predictions for each GT.

(2) For each test example, sample a random point in the parameter (output) space. Then, define the credible region $\mathcal{R}$ as the hypersphere centered on the random point and extended to the GT[24]. The fraction of predictions closer than the GT to the random point (i.e., the ones that fall inside $\mathcal{R}$) will approximate the credibility level $l_{\mathcal{R}}$.

(3) For each credible region $\mathcal{R}$ with credibility level $l_{\mathcal{R}}$, estimate the coverage probability $c_{\mathcal{R}}$ by calculating the fraction of credible regions with smaller or equal credibility level than $l_{\mathcal{R}}$[25]. Figure 3.22 shows why this works.

[24]This is the *smallest* possible hypersphere centered on the random point that contains the GT.
[25]In other words, we estimate the coverage probability using the CDF of the estimated credibility levels. A calibrated model is represented by a diagonal CDF, i.e., that of a uniform distribution.

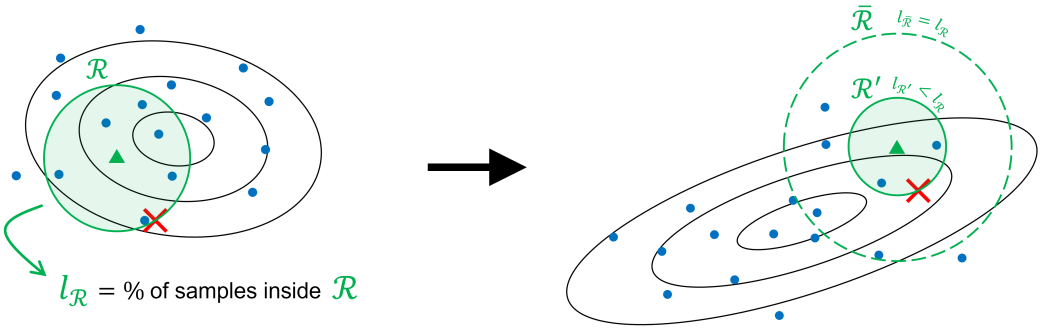

Fig. 3.22. Visual explanation of estimating coverage probabilities in TARP.

Estimating the coverage probability for region $\mathcal{R}$ involves constructing an $l_{\mathcal{R}}$-credible region $\bar{\mathcal{R}}$ for each test example, followed by calculating the fraction of these regions that encompass the GT. Our focus is on credible regions in the form of hyperspheres centered on the random point. For a region $\mathcal{R}'$ with credibility level $l_{\mathcal{R}'}$, in cases where $l_{\mathcal{R}} > l_{\mathcal{R}'}$, the corresponding $\bar{\mathcal{R}}$ for that example will be larger than $\mathcal{R}'$. Given that $\mathcal{R}'$ is inherently the smallest region containing the GT, it is guaranteed that $\bar{\mathcal{R}}$ will include the GT for that particular example.

(4) Generate a coverage probability curve by plotting $c_{\mathcal{R}}$ vs. $l_{\mathcal{R}}$.

It is important to note while coverage tests provide valuable insights into the calibration of uncertainties and the accuracy of credible regions, they do not directly measure the overall accuracy of a model. For instance, a model that generates samples from the "prior" distribution yields well-calibrated credible regions. However, it completely disregards the useful information in the input data to constrain its output.

With the astrophysical and deep learning foundations of our work established, we now turn to present our high-dimensional posterior sampling framework, its application to CMB delensing, and the main results in the next chapter.

First Article.

A Deep Generative Framework for Fast High-dimensional Posterior Sampling: Application to CMB Delensing

by

Mohammad-Hadi Sotoudeh[1,2,3], Pablo Lemos[1,2,3,4], and Laurence Perreault-Levasseur[1,2,3,4]

([1]) Department of Physics, University of Montreal
([2]) Mila - Quebec AI Institute
([3]) Ciela - Montreal Institute for Astrophysics and Machine Learning
([4]) Center for Computational Astrophysics, Flatiron Institute

This article will be submitted to The Astrophysical Journal.

Abstract

The next generation of telescopes and simulations are set to vastly increase the volume and resolution of the available astrophysical data. Performing Bayesian inference to derive insights from this data encounters challenges due to the exponential growth in data complexity. While existing posterior sampling methods bypass the costs of fully modeling the posterior distribution, they often prove impractical due to high computational complexity or overly simplified theoretical assumptions that neglect small-scale physics. This paper introduces a deep generative framework for fast posterior sampling based on the Hierarchical Probabilistic U-Net architecture. We apply this framework to remove the effect of weak gravitational lensing from CMB and evaluate the learned posterior by examining the generated samples' power spectra and conducting the coverage probability test. While our model's uncertainty estimates are slightly conservative, it can accurately delens CMB maps, such the power spectrum of unlensed (target) maps mainly lie within the 2σ range defined by the variability in the model's generated samples. We also demonstrate our model's robustness against changes to cosmological parameters, making it suitable for real-observation scenarios.

Keywords: High-dimensional Bayesian Inference, Posterior Sampling, Deep Learning, Generative Models, Cosmology, CMB Delensing

1. Introduction

Throughout the present decade, modern instruments and simulations will significantly enhance the available astrophysical data in several aspects. Space missions and ground-based telescopes such as the James Webb Space Telescope, Euclid, Roman Space Telescope, SPHEREx, Vera Rubin Observatory, Simons Observatory, and CMB-S4 will not only bring about exponential growth to the size of the observed data but also will increase the resolution and data acquisition rate to unprecedented levels.

Utilizing this data to infer physical and astrophysical parameters relies on Bayesian inference techniques, in which the posterior distribution $p(\boldsymbol{y}|\boldsymbol{x})$ of parameters $\boldsymbol{y}$ conditioned on observed data $\boldsymbol{x}$ is computed by combining the initial beliefs about the parameters encoded in the prior distribution $p(\boldsymbol{y})$, and the likelihood $p(\boldsymbol{x}|\boldsymbol{y})$ which captures the probability of observing the data given the parameters. However, given the vast volume and high dimensionality of the observed data and parameters, directly modeling the posterior distribution is intractable. Further complicating matters, even performing approximate methods like Markov Chain Monte Carlo (MCMC), Nested Sampling, Variational Inference, and Approximate Bayesian Computation (ABC) that generate posterior samples becomes infeasible in the data-intensive regime. This happens for two reasons: First, some posterior sampling

methods rely on evaluating the likelihood, which is only possible through simplifying assumptions that neglect the complexities of small-scale - but now detectable through observations - physics. Furthermore, even implicit likelihood approaches that indirectly approximate likelihood might suffer from excessive computational costs. As a result, exploiting the full potential of the observed data requires innovative data analysis approaches that do not require direct likelihood modeling and are computationally feasible in high-dimensional scenarios.

In recent years, Machine Learning (ML) methods have gained popularity to address challenges posed by traditional data analysis algorithms in physics. However, many of them are limited to providing point estimates of parameters with no measure of uncertainty of their predictions. This paper presents a deep generative framework to perform fast high-dimensional Bayesian inference. Our model is based on the Hierarchical Probabilistic U-Net (HPU-Net) architecture that combines the U-Net architecture [**28**], appropriate for learning high-dimensional mappings, with the approximate inference framework of Variational Autoencoders (VAEs). To demonstrate the physical application of our framework, we apply it to the problem of CMB delensing, i.e., removing the effect of weak gravitational lensing from the cosmic microwave background.

This paper is organized as follows: In Section 2, we introduce the model's architecture and learning objective. Section 3 covers the experiments conducted with our model. We start by introducing the metrics we used to evaluate the model's performance. Then, we present the theoretical foundations, data generation process, and main results for the two experiments conducted. Subsequently, Section 4 highlights the main findings of this project and discusses future directions to improve our work. Finally, Section 5 concludes the paper by summarizing the key insights and contributions of this study.

The authors confirm contribution to the paper as follows: MH.S. contributed to the model's design, dataset generation, training the model, analyzing the results, and writing the paper. P.L. contributed to research planning and supervision of dataset generation. L.PL. contributed to research planning, supervision of model's design and dataset generation, and analysis of results.

2. Model

2.1. Deep Generative Models

Our inference framework is built upon deep generative models, a class of deep learning models designed to generate new data samples resembling existing training examples. The presence of sampling layers in generative networks enables them to quantify uncertainties and expose variability in their predictions. Some well-known deep generative models include Generative

Adversarial Networks (GANs) [**9**], Variational Autoencoders (VAEs) [**17**], Normalizing Flows [**26**] and Diffusion Models [**14**].

We base our model on VAEs for several reasons. First, they are sampled more efficiently on large datasets with a lower computational cost. In contrast, conventional diffusion models, for example, require a large number of iterations to generate a sample [**4**], especially in high-dimensional applications. Furthermore, VAEs have a well-understood approximate inference framework that describes their probabilistic behavior and can be used to address their potential limitations. For instance, [**27**] have leveraged this framework to propose an alternative optimization algorithm for VAEs based on constrained optimization to train them in a more principled way and improve their common output blurriness issue.

We use the Hierarchical Probabilistic U-Net (HPU-Net) [**18**] architecture for inference. It is constructed by augmenting a U-Net with latent spaces and training the resulting model using a scheme similar to the approximate inference framework of Conditional Variational Autoencoders (cVAEs). HPU-Net is suitable for learning high-dimensional probabilistic mappings, and it is designed for underconstrained problems where the available data is not sufficient for drawing a single acceptable answer. In this context, the network aims to learn the latent manifold associated with acceptable answers and employ it to generate parameters consistent with observations. This makes the network capable of modeling multimodal distributions, accounting for intrinsic uncertainties, and addressing physical applications where parameter estimates are always prone to some level of uncertainty due to observational noise.

2.2. Network Architecture

Our general goal can be defined as learning the posterior distribution $p(\boldsymbol{y}|\boldsymbol{x})$ of some physical parameters $\boldsymbol{y}$ given observed data $\boldsymbol{x}$. To achieve this goal, We use two separately trained networks:

(1) **Mean Net:** This network is a regular (deterministic) U-Net with the goal of learning the *mean of the posterior distribution* ($\bar{\boldsymbol{y}}$).

(2) **Noise Net:** This network is an HPU-Net that aims to learn the distribution of *deviations* ($\boldsymbol{n} := \boldsymbol{y} - \bar{\boldsymbol{y}}$), i.e., the difference between posterior samples and the posterior mean. Sampling a deviation from this network and adding it to the posterior mean makes a posterior sample.

Figure 4.1 illustrates our framework's architecture. Both the Mean Net and Noise Net have a base U-Net architecture that includes a contracting path, an expanding path, and skip connections. The contracting path is responsible for reducing the spatial dimensions of the input while capturing high-level abstract features. It is composed of convolutional blocks (consisting of convolutional, batch normalization, and activation layers) and downsampling

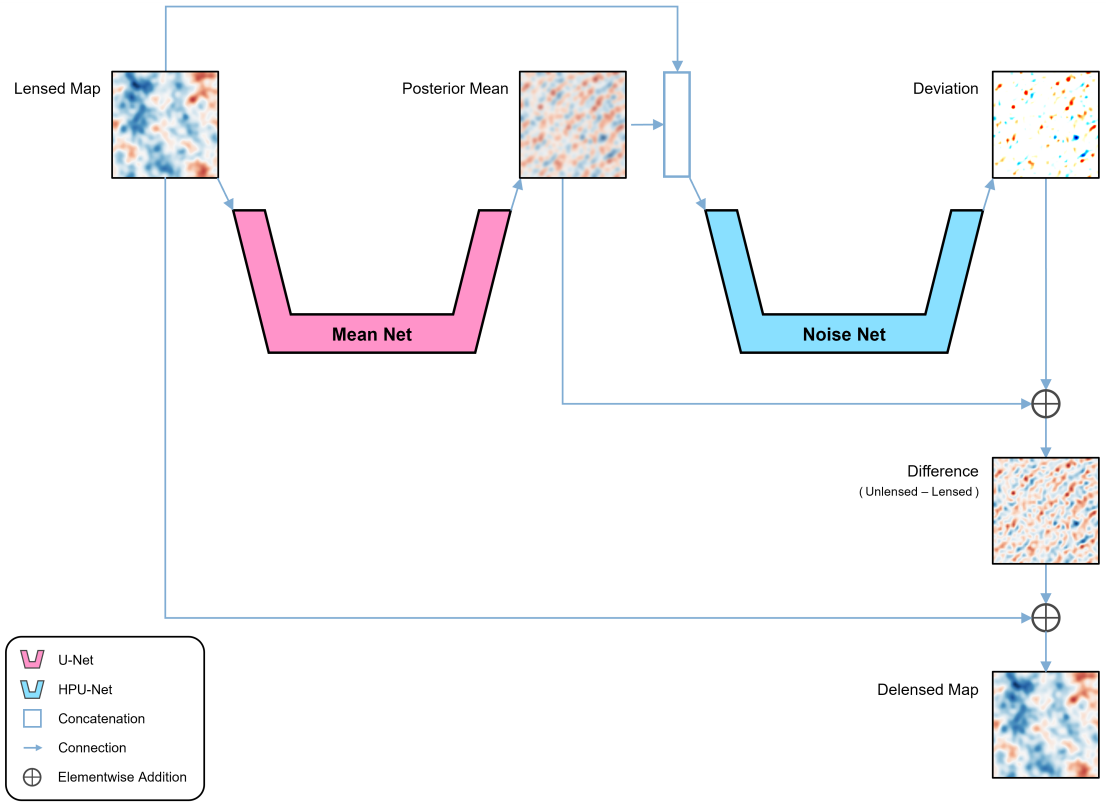

Fig. 4.1. Diagram of our framework's architecture. It consists of two neural networks: The Mean Net, which is a U-Net that learns the posterior mean, and the Noise Net, which is a Hierarchical Probabilistic U-Net that generates deviation samples (i.e., the difference between posterior samples and the posterior mean).

operations. In order to facilitate the flow of gradients through the network, the convolutional blocks utilize residual connections suggested by [**13**]. The expanding path reconstructs a finely detailed output from the compact representation obtained in the contracting path. It includes similar convolutional blocks, as well as upsampling operations. Finally, skip connections preserve spatial information and enhance the accuracy of the network by allowing a direct flow of information at each resolution.

In the Noise Net, certain scales in the expanding path have a latent space with the same dimensionality as the corresponding feature maps. Throughout the propagation of feature maps in the expanding path, they are concatenated with samples drawn from corresponding latent spaces. This enables the network to quantify uncertainties at different resolutions and exhibit probabilistic behavior.

2.3. Learning Objective

Each neural network represents a function f using its parameters $\boldsymbol{\theta}$. It is trained on a set of examples $\mathcal{D}_{\text{train}} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$ to minimize a loss function $\mathcal{J}(\boldsymbol{\theta}, \mathcal{D}_{\text{train}})$. The loss function

measures the error between the model's predictions $\hat{\boldsymbol{y}} := f_{\boldsymbol{\theta}}(\boldsymbol{x})$ and the ground truth (GT, target, or true) values $\boldsymbol{y}$. We can summarize the learning objective of the network in the following equation:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta}, \mathcal{D}_{\text{train}}), \tag{4.21}$$

$\mathcal{J}(\boldsymbol{\theta}, \mathcal{D}_{\text{train}})$ is defined using the instance-level loss $\mathcal{L}(\hat{\boldsymbol{y}}, \boldsymbol{y})$, which determines how the prediction is compared to the ground truth for one training example:

$$\mathcal{J}(\boldsymbol{\theta}, \mathcal{D}_{\text{train}}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}_{\text{train}}}\Big[\mathcal{L}(\hat{\boldsymbol{y}}, \boldsymbol{y})\Big] \tag{4.22}$$

In other words, the loss function value is obtained by averaging the instance-level loss over the training dataset. To improve computational efficiency, this average is typically calculated over a batch of size b of training examples rather than the entire dataset, i.e.,

$$\mathcal{J}(\boldsymbol{\theta}, \mathcal{D}_{\text{train}}) \approx \frac{1}{b}\sum_{i=1}^{b} \mathcal{L}(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_i). \tag{4.23}$$

Mean Net Objective

The Mean Net is trained to minimize the Mean Squared Error (MSE) between its output and the target value:

$$\mathcal{L}_{\text{MeanNet}} = \frac{1}{2}\left\|\hat{\bar{\boldsymbol{y}}} - \boldsymbol{y}\right\|_2^2 \tag{4.24}$$

According to [**2**], if a deterministic neural network (i.e., with no sampling layers) is trained with MSE loss, its optimal solution will be the posterior mean[1]. Hence, if properly trained, the Mean Net is guaranteed to learn the posterior mean. Figure 4.2 illustrates the training process of the Noise Net.

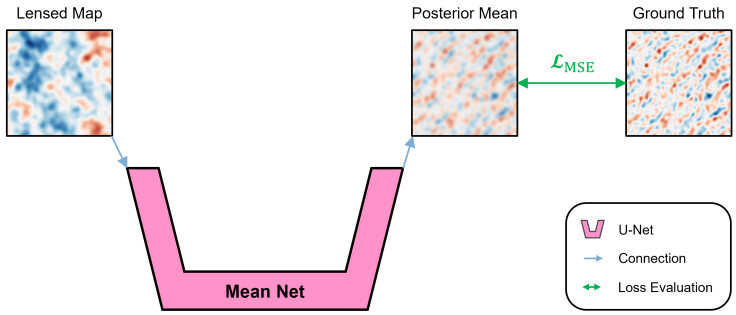

Fig. 4.2. Training process of the Mean Net.

[1]This is why we used the "bar" in $\hat{\bar{\boldsymbol{y}}}$ to denote model's prediction.

Noise Net Objective

The Noise Net has the same functionality as the conglomerate of the prior and decoder of a VAE. It models the joint distribution of deviations $\boldsymbol{n}$ and latent variables[2] $\boldsymbol{z}$, conditioned on observed data $\boldsymbol{x}$:

$$p(\boldsymbol{n}, \boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{x})\; p(\boldsymbol{n}|\boldsymbol{z}, \boldsymbol{x}). \tag{4.25}$$

The Noise Net's latent spaces have a hierarchical structure, i.e., the latent variables at each decoding scale depend on the previous scales' latent variables. Hence, the joint conditional probability of the latent variables, also known as the conditional prior of latents, $p(\boldsymbol{z}|\boldsymbol{x})$ can be factorized as

$$p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_L|\boldsymbol{x}) = p(\boldsymbol{z}_L|\boldsymbol{z}_{<L}, \boldsymbol{x}) \cdot \ldots \cdot p(\boldsymbol{z}_2|\boldsymbol{z}_1, \boldsymbol{x}) \cdot p(\boldsymbol{z}_1|\boldsymbol{x}), \tag{4.26}$$

where L is the total number of latent spaces. The network aims to maximize the Evidence Lower Bound (ELBO) on the posterior $p(\boldsymbol{n}|\boldsymbol{x})$, which can be decomposed into the following terms:

$$\begin{aligned} \mathcal{L}_{\text{NoiseNet}} &= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{n},\boldsymbol{x})}\Big[-\ln p(\boldsymbol{n}|\boldsymbol{z},\boldsymbol{x})\Big] + D_{\text{KL}}\Big(q(\boldsymbol{z}|\boldsymbol{n},\boldsymbol{x}) \parallel p(\boldsymbol{z}|\boldsymbol{x})\Big) \\ &= \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}} \end{aligned} \tag{4.27}$$

The first term in Equation 4.27 is called the reconstruction term. It evaluates the fidelity of the generated data to the expected output. By assuming that $p(\boldsymbol{n}|\boldsymbol{z},\boldsymbol{x})$ is a multivariate Gaussian distribution, the reconstruction term will reduce to the Gaussian Negative Log-likelihood (GNLL) Loss:

$$\mathcal{L}_{\text{rec}} = \frac{1}{2}\Big(\ln|\hat{\Sigma}| + (\boldsymbol{n} - \hat{\boldsymbol{\mu}})^\top \hat{\Sigma}^{-1}(\boldsymbol{n} - \hat{\boldsymbol{\mu}})\Big), \tag{4.28}$$

with $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ being the mean and covariance matrix of $p(\boldsymbol{n}|\boldsymbol{z},\boldsymbol{x})$ and $\boldsymbol{n}$ being a real posterior sample. If we further assume that the Gaussian distribution is diagonal[3], Equation 4.28 will be simplified to

$$\mathcal{L}_{\text{rec}} = \frac{1}{2}\sum_i \left[\frac{(\boldsymbol{n}_i - \hat{\boldsymbol{\mu}}_i)^2}{\hat{\sigma}_i^2} + \hat{\sigma}_i^2\right], \tag{4.29}$$

where $\hat{\boldsymbol{\mu}}_i$ and $\hat{\sigma}_i^2$ are the i^{th} component of $\hat{\boldsymbol{\mu}}$ and the i^{th} diagonal element of $\hat{\Sigma}$, respectively. Each index corresponds to an output pixel, and the summation is performed over all output pixels. During training, the $\hat{\boldsymbol{\mu}}_i$s and σ_i^2s are estimated by drawing multiple samples from the Noise Net and calculating their pixel-wise mean and variance. To clarify further, considering the difference between the GT and the posterior mean as a real posterior sample, the network aims to produce samples with similar statistical properties. It does so by learning the mean and variance of $p(\boldsymbol{n}|\boldsymbol{z},\boldsymbol{x})$ such that the probability of sampling the GT is maximized.

[2]Often, latent variables are simply referred to as *latents*.

[3]This means that by knowing the observations $\boldsymbol{x}$ and latent variables $\boldsymbol{z}$, we can determine each output pixel's value independent of other pixels.

The second term in Equation 4.27 is the Kullback–Leibler (KL) divergence between the variational posterior $q(\boldsymbol{z}|\boldsymbol{n},\boldsymbol{x})$ and the prior $p(\boldsymbol{z}|\boldsymbol{x})$ distribution of latents. The variational distribution $q(\boldsymbol{z}|\boldsymbol{n},\boldsymbol{x})$ approximates the true posterior of latent variables $p(\boldsymbol{z}|\boldsymbol{n},\boldsymbol{x})$[4], and has a similar autoregressive structure to the prior:

$$q(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_L|\boldsymbol{n},\boldsymbol{x}) = q(\boldsymbol{z}_L|\boldsymbol{z}_{<L},\boldsymbol{n},\boldsymbol{x})\cdot\ldots\cdot q(\boldsymbol{z}_2|\boldsymbol{z}_1,\boldsymbol{n},\boldsymbol{x})\cdot q(\boldsymbol{z}_1|\boldsymbol{n},\boldsymbol{x}). \tag{4.30}$$

It is modeled by an auxiliary network, which will shortly be introduced. The KL term assimilates p and q, enriching p with information from q and encouraging smooth and structured latent representations. It is shown in [**18**] that for the HPU-Net, $\mathcal{L}_{\mathrm{KL}}$ decomposes into the sum of the KL divergences between individual latent spaces:

$$D_{\mathrm{KL}}\Big(q(\boldsymbol{z}|\boldsymbol{n},\boldsymbol{x}) \parallel p(\boldsymbol{z}|\boldsymbol{x})\Big) = \sum_{l=1}^{L} \mathbb{E}_{\boldsymbol{z}_{<l}\sim q} D_{\mathrm{KL}}\Big(q(\boldsymbol{z}_l|\boldsymbol{z}_{<l},\boldsymbol{n},\boldsymbol{x}) \parallel p(\boldsymbol{z}_l|\boldsymbol{z}_{<l},\boldsymbol{x})\Big) \tag{4.31}$$

If we choose the latent spaces to represent pixel-wise Gaussian distributions, $\mathcal{L}_{\mathrm{KL}}$ can be evaluated analytically using the pixel-wise means and variances of each (p, q) pair:

$$D_{\mathrm{KL}}(q||p) = \frac{1}{2}\sum_i \left[\ln\left(\frac{\sigma_{p,i}^2}{\sigma_{q,i}^2}\right) + \frac{\sigma_{q,i}^2 + (\mu_{q,i}-\mu_{p,i})^2}{\sigma_{p,i}^2} - 1\right]. \tag{4.32}$$

The summation is performed over the latent spaces' spatial dimensions (i.e., pixels).

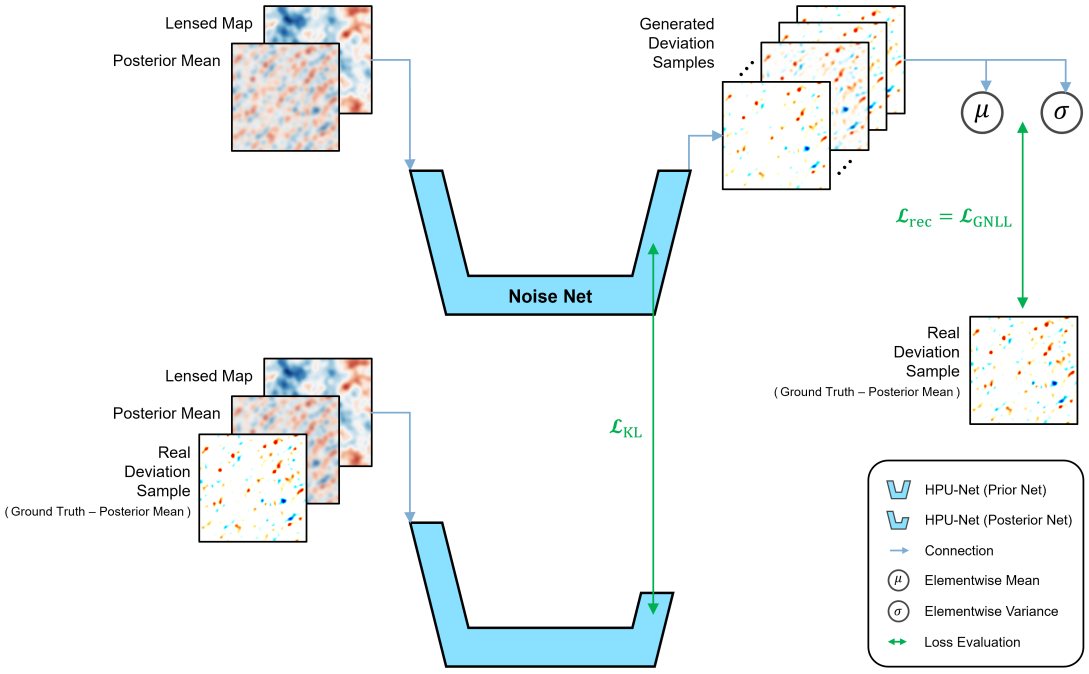

Fig. 4.3. Training process of the Noise Net

Figure 4.3 illustrates the training process of the Noise Net. The auxiliary network (aka the Posterior Net) has the same hierarchical topology as the primary network (aka the Prior

[4]Note that the latents posterior $p(\boldsymbol{z}|\boldsymbol{n},\boldsymbol{x})$ differs from the parameters posterior $p(\boldsymbol{n}|\boldsymbol{x})$.

Net)[5], but with two exceptions: 1) the Posterior Net receives both the observation and the real sample as input, and 2) since the only purpose of the Posterior Net is to draw samples from the variational posterior, it has a truncated decoder, i.e., the layers with the **sole** purpose of decoding latents are not present in the Posterior Net. During training, pixel-wise means and variances are calculated for both the Prior and Posterior Net latents. However, the latent samples used for training are drawn from the Posterior Net and injected into the Prior Net.

3. Experiments and Results

This section presents the experiments carried out with HPU-Net. It starts with introducing the employed performance measurement methods. Subsequently, the details of the experiments and their results are presented.

3.1. Performance Measures

We use the following methods to assess the quality of the learned posterior:

(1) **Comparing Moments:** Since we can only access samples from the learned posterior, with no direct access to the posterior itself, our examination options are limited. One option is to compare the moments of the true posterior with estimated moments from the model's generated samples. The available orders and the accuracy of moment comparison are determined by the amount of information available from the true posterior, as well as the number of generated samples.

(2) **Comparing Power Spectra:** The power spectrum is a fundamental prediction of various cosmological models. By comparing the power spectra of the model's predictions with the GT's power spectrum, we can determine how accurately the model can generate maps with correct spatial features.

(3) **Coverage Probability Test:** We employ the Test of Accuracy with Random Points (TARP) [**19**] to evaluate the accuracy of credible regions predicted by the model. For each credible region $\mathcal{R}$, this test compares its estimated credibility level $l_{\mathcal{R}}$ with the fraction of times it actually covers the true parameter, i.e., its coverage probability $c_{\mathcal{R}}$. For a calibrated model, $l_{\mathcal{R}}$ and $c_{\mathcal{R}}$ should match. However, if the model constantly forms smaller-than-expected credible regions, they will be less likely to cover the true value than their specified credibility level. In this case, $c_{\mathcal{R}} < l_{\mathcal{R}}$, and the model is called *overconfident*. On the other hand, an *underconfident* or *conservative* model forms larger-than-expected credible regions and has $c_{\mathcal{R}} > l_{\mathcal{R}}$. Figure 4.4 summarizes how TARP is conducted in four steps.

[5]Posterior and prior here refer to the distributions of latent variables $\boldsymbol{z}$, which must be differentiated from the posterior and prior distributions of parameters $\boldsymbol{y}$.

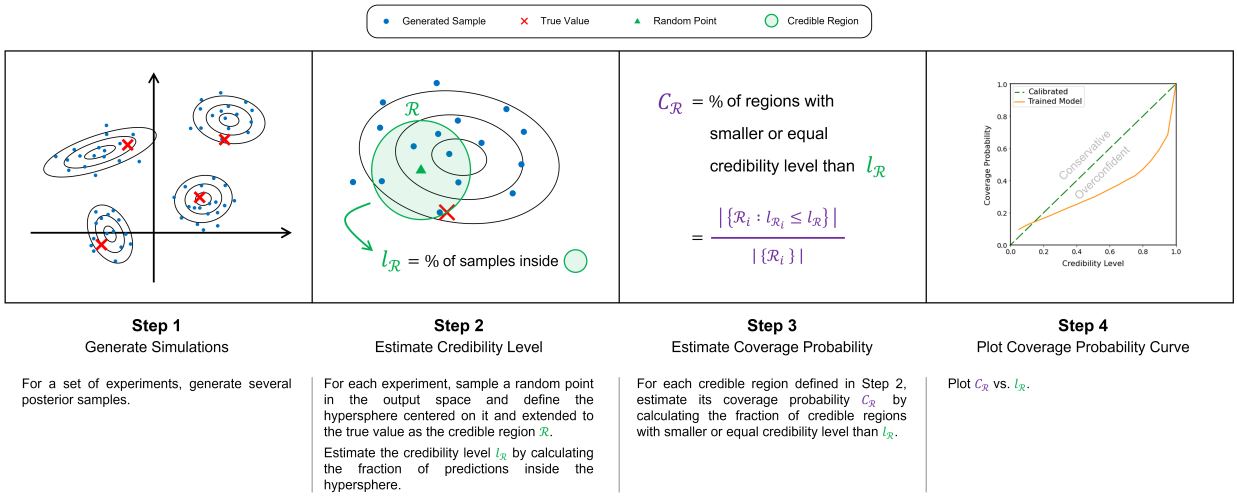

Fig. 4.4. TARP coverage probability test in four steps.

3.2. Problem 1: Rotating Gaussian Random Fields (GRFs)

Motivation and Theoretical Framework

This experiment is aimed to assess the model's performance on a problem where the posterior $p(\boldsymbol{y}|\boldsymbol{x})$ can be analytically derived. In this case, the model was trained to solve a linear inverse problem, where the inputs ($\boldsymbol{x}$) and outputs ($\boldsymbol{y}$) of the model followed the relation

$$\boldsymbol{x} = R\boldsymbol{y} + \boldsymbol{n}, \tag{4.33}$$

where R is a transformation matrix, and $\boldsymbol{n}$ is a random noise vector. It can be shown that if the parameters' prior $p(\boldsymbol{y})$ and the noise $p(\boldsymbol{n})$ follow Gaussian distributions, the posterior distribution $p(\boldsymbol{y}|\boldsymbol{x})$ will also be a Gaussian with the following parameters:

$$\boldsymbol{\mu}_{\text{post}} = (M + M^\top)^{-1}\boldsymbol{d} \qquad \Sigma_{\text{post}}^{-1} = M, \tag{4.34}$$

with M and $\boldsymbol{d}$ defined as:

$$\begin{aligned} M &:= (\Sigma_{\text{pri}}^{-1} + R^\top \Sigma_{\text{n}}^{-1} R) \\ \boldsymbol{d} &:= 2\,(\Sigma_{\text{pri}}^{-1}\,\boldsymbol{\mu}_{\text{pri}} + R^\top \Sigma_{\text{n}}^{-1}\boldsymbol{x}), \end{aligned} \tag{4.35}$$

where the pri, post, and n indices indicate the parameters of the prior, posterior, and noise distributions, respectively. If we choose R to be a 90° rotation, we will have $R^\top R = RR^\top = \mathbb{I}$[6]. If we further assume that $\boldsymbol{\mu}_{\text{pri}} = 0$, $\Sigma_{\text{pri}} = \sigma_{\text{pri}}^2\mathbb{I}$, and $\Sigma_{\text{n}} = \sigma_{\text{n}}^2\mathbb{I}$, Equation 4.35 will reduce to:

$$M = \left(\frac{1}{\sigma_{\text{pri}}^2} + \frac{1}{\sigma_{\text{n}}^2}\right)\mathbb{I} \qquad \boldsymbol{d} = \frac{2}{\sigma_{\text{n}}^2}R^\top\boldsymbol{x}. \tag{4.36}$$

[6]This is true for every orthogonal matrix, and since rotation preserves the norm of its input, we can describe it using an orthogonal matrix.

By defining $\sigma^2_{\text{post}} := 1/(\sigma^{-2}_{\text{pri}} + \sigma^{-2}_{\text{n}})$ and combining it with Equation 4.36 we can simplify Equation 4.34 to:

$$\boldsymbol{\mu}_{\text{post}} = \frac{\sigma^2_{\text{post}}}{\sigma^2_{\text{n}}} R^\top \boldsymbol{x} \qquad \Sigma^{-1}_{\text{post}} = \sigma^2_{\text{post}} \mathbb{I}. \tag{4.37}$$

Once the model is trained, we can compare $\boldsymbol{\mu}_{\text{post}}$ and Σ_{post} with the empirical mean and covariance calculated from the model's predictions.

Data Generation

We train the model on a dataset of 2^{16} examples. The data was generated by sampling from the prior distribution, rotating the samples and adding noise to the rotated samples. In order to increase the effective size of the training set and enhance the model's generalization, different noise realizations were used for each training epoch. Figure 4.5 summarizes the data generation process.

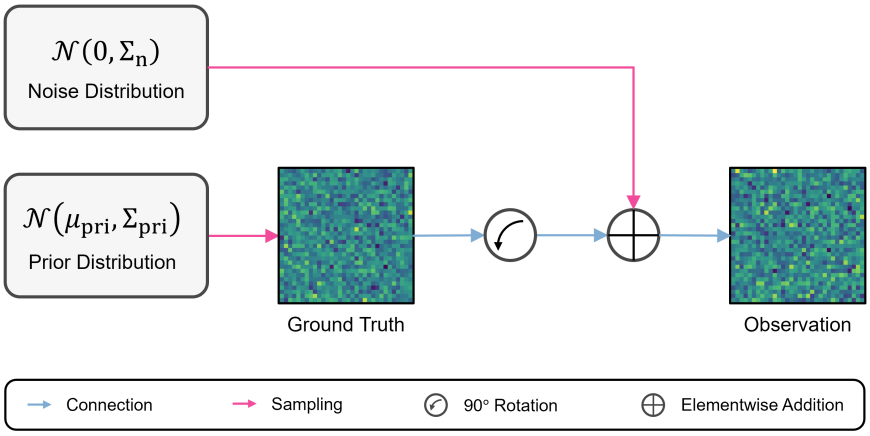

Fig. 4.5. Data generation process for the GRF rotation experiment.

Evaluating Performance

Thanks to having full access to the true posterior, we can directly compare its moments with the learned posterior. For this purpose, we use the model to generate posterior samples, calculate their pixel-wise means $\tilde{\boldsymbol{\mu}}_{\text{post}}$ and standard deviations $\tilde{\boldsymbol{\sigma}}_{\text{post}}$, and compare them to $\boldsymbol{\mu}_{\text{post}}$ and σ_{post} obtained from Equation 4.37. All higher-order pixel-wise moments are expected to be zero. In another approach, we can take the GT as a point estimate of $\boldsymbol{\mu}_{\text{post}}$ and compare it with $\tilde{\boldsymbol{\mu}}_{\text{post}}$. Figure 4.16 displays the empirical moments calculated from posterior samples and how they compare with the GT and theoretical moments for a given test example. Several posterior samples are included in the figure as well. For more examples, see Appendix A. Figure 4.7 shows the result of the TARP test for this experiment.

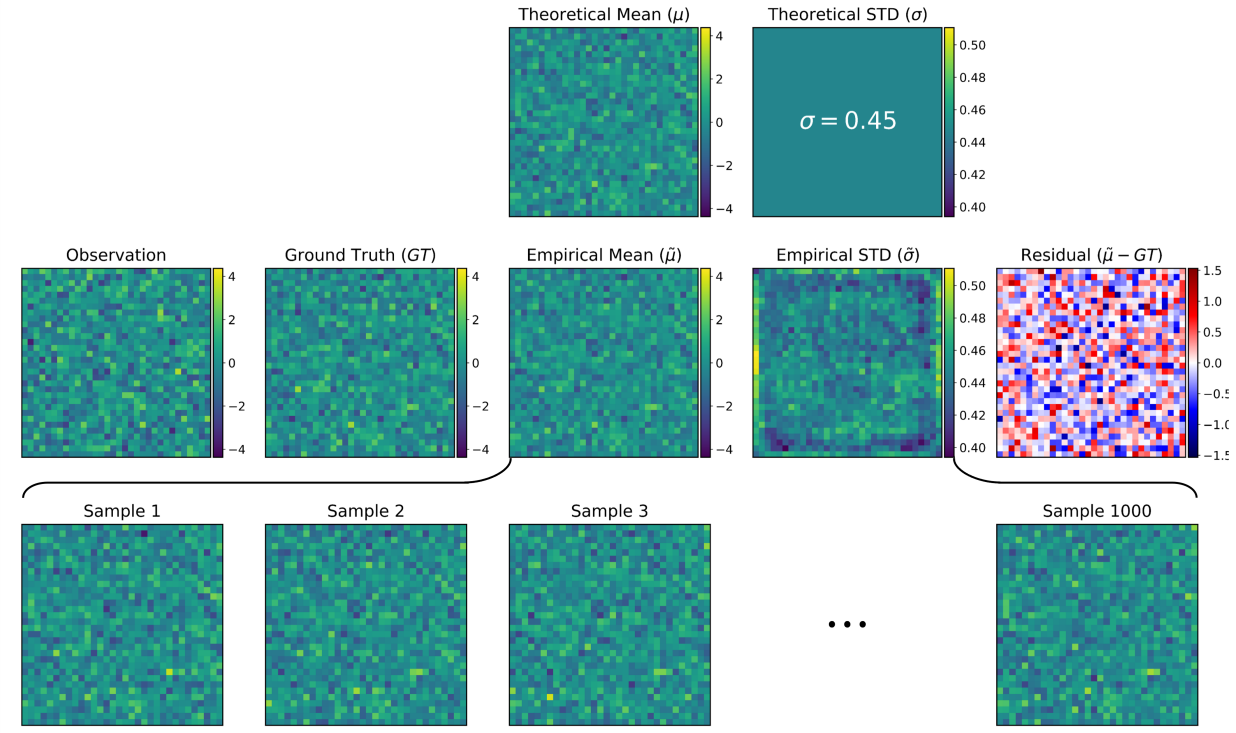

Fig. 4.6. Moment comparison plot for a particular test example of the GRF rotation experiment. 1000 posterior samples were drawn using the model and used to calculate pixel-wise empirical means and standard deviations. For more examples, see Appendix A.

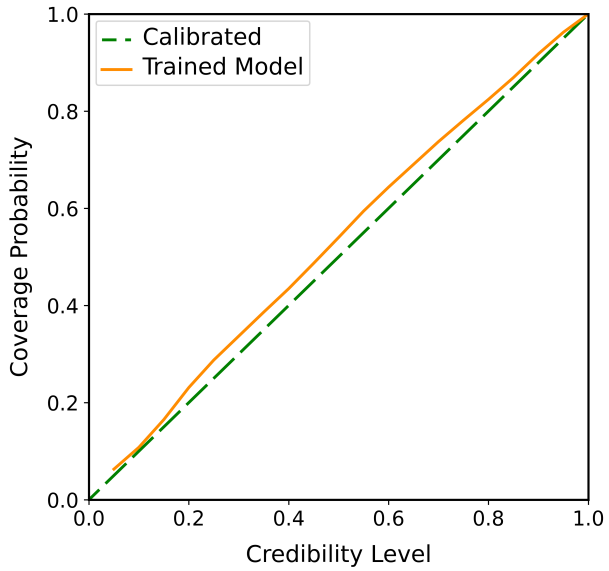

Fig. 4.7. TARP test result for the GRF rotation experiment. The test was conducted on 2048 test examples, with 200 posterior samples generated for each example.

3.3. Problem 2: CMB Delensing

Motivation and Theoretical Framework

CMB lensing is a secondary effect on the cosmic microwave background caused by the weak gravitational lensing effect of intervening cosmic structures between the last scattering surface and us. It alters the observed directions of CMB photons, such that they no longer represent their original emission direction. The root mean square of deflection angles is around 2 arcminutes, and they are correlated over the sky in areas as large as around 2 degrees, which is comparable to the degree-scale primary fluctuations of CMB [**20, 11**]. The lensing also introduces small amounts of non-Gaussianity (non-zero higher order moments) and statistical anisotropy (non-zero off-diagonal covariance elements) into the primordial signal. Using small angle, flat sky, and Born approximations, the deflection angle $\boldsymbol{\alpha}$ caused by lensing is given by

$$\boldsymbol{\alpha} \approx -2 \int_{\text{los}} \frac{S_\kappa(\chi_{\text{CMB}} - \chi)}{S_\kappa(\chi_{\text{CMB}})} \nabla_\perp \Psi \; d\chi, \tag{4.38}$$

where the integration is performed along the line of sight (los), $\nabla_\perp$ are the gradient components perpendicular to the line of sight, Ψ is the Weyl potential, χ is comoving distance, and $S_\kappa(\chi)$ depends on the geometry of the universe and is given by the following relation:

$$S_\kappa(\chi) = \begin{cases} R \sin(\chi/R) & \kappa = +1 \text{ (Closed)} \\ \chi & \kappa = 0 \quad \text{(Flat)} \\ R \sinh(\chi/R) & \kappa = -1 \text{ (Open)} \end{cases}, \tag{4.39}$$

where R is the present-day radius of curvature of the universe. Since we can safely approximate the source CMB radiation to be instantaneously emitted, it is convenient to aggregate all lens information in the projected lensing potential ψ on the sky plane:

$$\psi(\hat{\boldsymbol{n}}) := -2 \int \frac{S_\kappa(\chi_{\text{CMB}} - \chi)}{S_\kappa(\chi_{\text{CMB}})\, S_\kappa(\chi)} \Psi \; d\chi, \tag{4.40}$$

such that the deflection angle is given by

$$\boldsymbol{\alpha}(\hat{\boldsymbol{n}}) = \nabla_{\hat{\boldsymbol{n}}} \psi, \tag{4.41}$$

with $\hat{\boldsymbol{n}}$ being the desired direction on the sky and $\nabla_{\hat{\boldsymbol{n}}}(\cdot) = S_\kappa(\chi)\, \nabla_\perp(\cdot)$ being the angular gradient.

As the lensing potential interacts with CMB, it leads to a mixing of different spatial scales (i.e., modes) of CMB temperature and polarization fluctuations. This mixing (i.e., mode-coupling) introduces correlations between fluctuations on different scales, resulting in off-diagonal elements in the covariance matrix of the observed CMB. The characteristic spacing of these elements is $\delta\ell = 50$, given by the peak of the deflection angle power spectrum. In this work, we concentrate on the observed temperature map, where lensing alters its primordial

power spectrum C_ℓ^{TT} via[7]:

$$C_\ell^{\tilde{T}\tilde{T}} = (1 - \ell^2 R^\psi) C_\ell^{TT} + \int \frac{d^2\boldsymbol{k'}}{(2\pi)^2} \Big[\boldsymbol{k'} \cdot (\boldsymbol{k} - \boldsymbol{k'})\Big]^2 C^{\psi\psi}_{|\boldsymbol{k}-\boldsymbol{k'}|} C^{TT}_{\ell'}, \tag{4.42}$$

where $C_\ell^{\tilde{T}\tilde{T}}$ is the observed power spectrum, $C_\ell^{\psi\psi}$ is the lensing potential power spectrum, $\ell \equiv |\boldsymbol{k}|$, and R^ψ is half of the total mean-squared deflection, defined by:

$$R^\psi := \frac{1}{2}\langle |\nabla\psi|^2 \rangle \sim 3 \times 10^7. \tag{4.43}$$

Correcting the lensing effect is important for obtaining unbiased estimates of cosmological parameters from CMB. Furthermore, analyzing the additional information introduced by lensing enables probing the state of the universe at the moments the deflections took place. Finally, the lensing alters the polarization of CMB photons, most importantly by introducing a B-mode pattern that can act as a source of confusion with any primordial signal from gravitational waves. The latest Planck satellite's lensing results and analysis are discussed in [**3, 7**].

Traditional approaches to CMB delensing are based on reconstructing the projected potential ψ using a quadratic estimator [**15**]. Despite being successful at the present-generation instruments' noise levels, these methods fall short in the high signal-to-noise regime soon to be brought about by next-generation CMB missions. Hence, alternative analytical methods [**29, 10, 6, 24, 23**] have been developed to exploit the full potential of next-generation CMB data. [**8**] summarizes various delensing strategies and compares their performance.

Since the upsurge of interest and application of machine learning in scientific domains, it has emerged as an additional viable option for CMB delensing. Specifically, there have been efforts to utilize convolutional neural networks for this task [**5, 22, 30, 31**]. While the proposed models can remove the lensing effect from CMB with promising accuracy, they are limited to providing point estimates, with no measure of uncertainty. Our model is designed to draw samples from the posterior distribution of delensed maps, conditioned on the observed CMB. This way, we can obtain uncertainty estimates using the generated samples.

Data Generation

The data we used for this experiment was generated using the Python interface for the Code for Anisotropies in the Microwave Background (CAMB) [**21**]. CAMB is a software package used in cosmology to calculate theoretical predictions for CMB and the large-scale structure of the universe. CAMB uses cosmological parameters specified by the user to calculate the evolution of the universe from its early stages to the present time. It considers the primordial

[7]This approximation is obtained using the taylor expansion of CMB temperature field and under the flat sky approximation. The terms were kept up to the first order in $C_\ell^{\psi\psi}$.

fluctuations left by inflation and calculates their statistical properties. These fluctuations serve as initial conditions for the evolution of density perturbations, which determine the growth of structures in the universe. CAMB also computes CMB properties, taking into account how density perturbations affect its temperature and polarization. Finally, it generates observable quantities, such as temperature and polarization power spectrums, with various secondary effects (e.g., galactic emission, reionization, gravitational lensing, SZ effect, etc.) taken into account.

In this work, we use CAMB's temperature angular power spectra to generate synthetic CMB maps. To do so, we apply unlensed and lensed power spectra to the same noise realization in Fourier space and transform the resulting maps back to real space. The training set consists of 2^{13} maps with fixed cosmological parameters that cover $160' \times 160'$ regions in the sky, with a resolution of 32 pixels per dimension. The model receives a lensed map as input and provides posterior samples of the **difference** between the lensed map and its unlensed variant[8]. We add a random noise to the lensed maps to simulate observational noise. Similar to the previous experiment, we apply noise during training and use different realizations for each epoch. Figure 4.8 summarizes the data generation process.

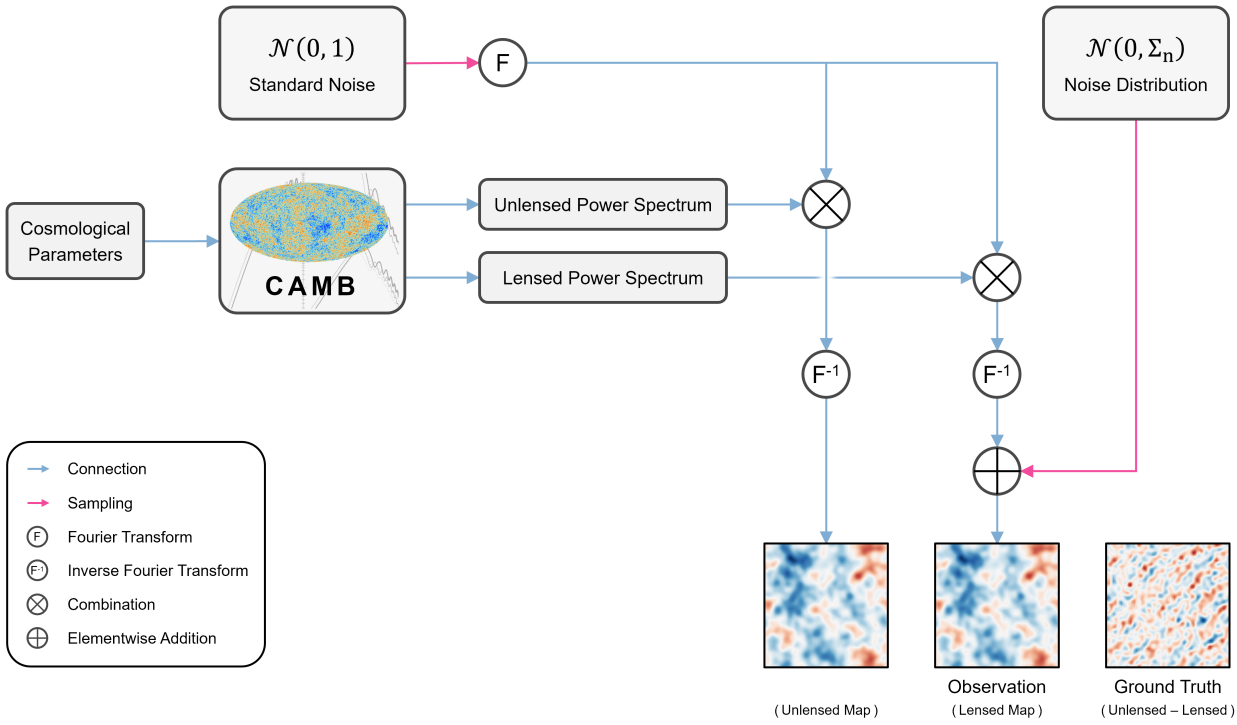

Fig. 4.8. Data generation process for the CMB delensing experiment.

[8]So, the Mean Net predicts the posterior mean of this difference, and the Noise Net predicts the deviation of a difference sample from the posterior mean.

Evaluating Performance

For this experiment, the true posterior distribution is unknown. Hence, when it comes to comparing moments, we can only compare the GT with the empirical mean. Figure 4.19 displays the empirical moments calculated from posterior samples and how the mean compares with the GT for a given test example. Several posterior samples are also included in the figure. For more examples, see Appendix B. Figure 4.10 shows how the mean power spectrum of the posterior samples (delensed maps) compares with the power spectrum of the unlensed map for a given test example. Finally, Figure 4.11 presents the result of the TARP test for this experiment.

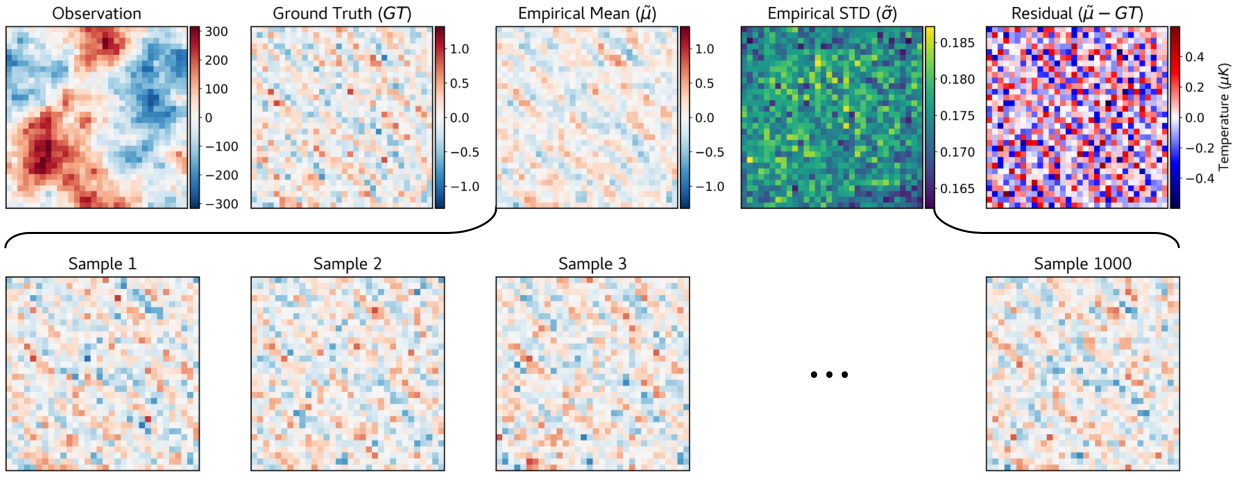

Fig. 4.9. Moment comparison plot for a particular test example of the CMB delensing experiment. 1000 posterior samples were drawn using the model and used to calculate pixel-wise empirical means and standard deviations. For more examples, see Appendix B.

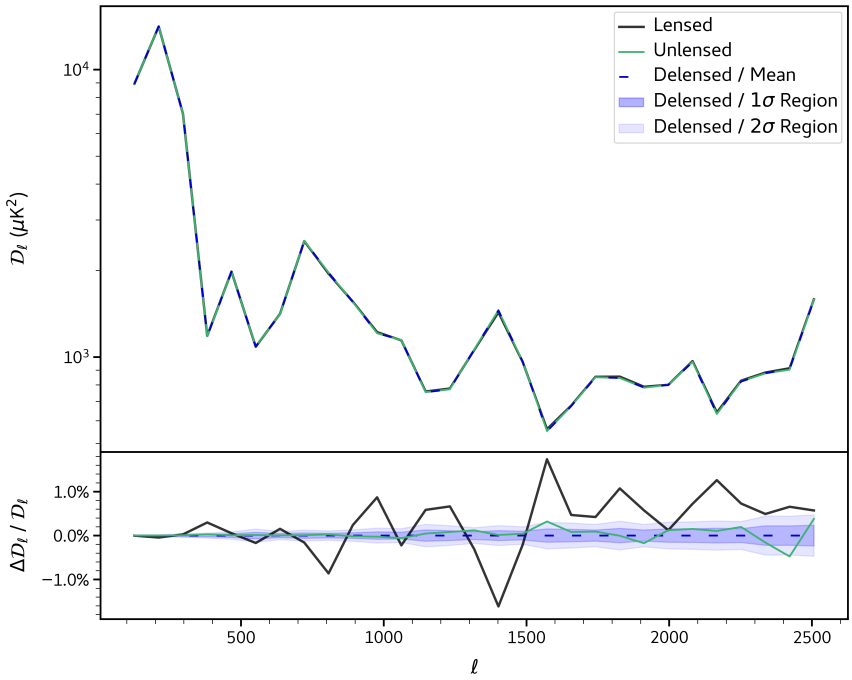

Fig. 4.10. Top panel: Comparison of the lensed (observation), unlensed (target), and mean delensed (model) temperature power spectra, generated from CMB maps for a particular test example. The delensed spectrum was calculated by averaging the power spectra of 1000 posterior samples. Bottom panel: The relative differences of the lensed and unlensed spectra from the mean delensed spectrum. The shaded regions represent 1σ and 2σ uncertainty regions of the delensed spectrum, calculated using the standard deviation of the delensed spectra. The unlensed spectrum lies mostly within the 1σ range and always within the 2σ range. For more examples, see the Base column of Figures 4.12 or 4.13.

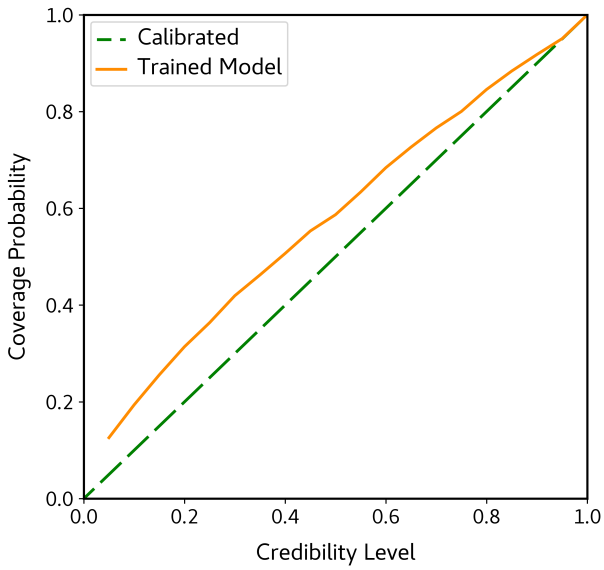

Fig. 4.11. TARP test result for the CMB delensing experiment. The test was conducted on 2048 test examples, with 200 posterior samples generated for each example.

Out-of-distribution Performance

To test our model's robustness, we evaluate its performance on out-of-distribution test examples. To generate out-of-distribution examples, we follow the same procedure as in Figure 4.8, but this time feed CAMB with different values for matter density parameter Ω_m and scalar amplitude A_s. We use Planck satellite's measurement error, $\sigma_{\Omega_m} = 0.0073$ ($\sigma_{A_s} = 1.014 \times 10^{-10}$) as increment to alter Ω_m (A_s), with every other parameter unchanged. We then examine the power spectra of the generated posterior samples. Figures 4.12 and 4.13 present the relative difference between the unlensed spectrum (target) and the mean delensed spectrum for several out-of-distribution test examples.

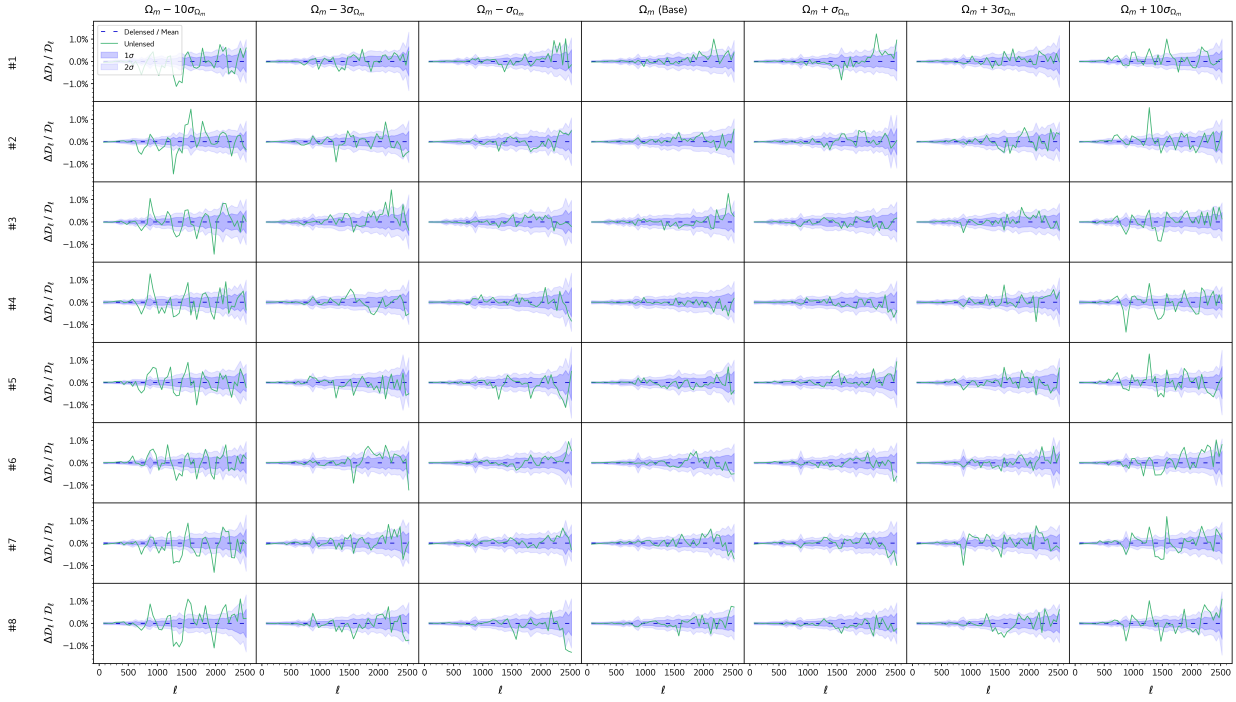

Fig. 4.12. Power spectra of out-of-distribution test examples for different matter density parameter Ω_m values, with every other cosmological parameter unchanged. Each column corresponds to an Ω_m value, which differs from the training value by a factor of the Planck measurement error σ_{Ω_m}. Each row corresponds to the same noise realization applied during data generation to the lensed and unlensed maps in Fourier space. Each panel depicts the relative difference of power spectra, similar to the bottom panel of Figure 4.10. Unless Ω_m is altered by a large factor of σ_{Ω_m}, the unlensed (target) spectrum stays mostly within the 2σ region of the mean delensed spectrum.

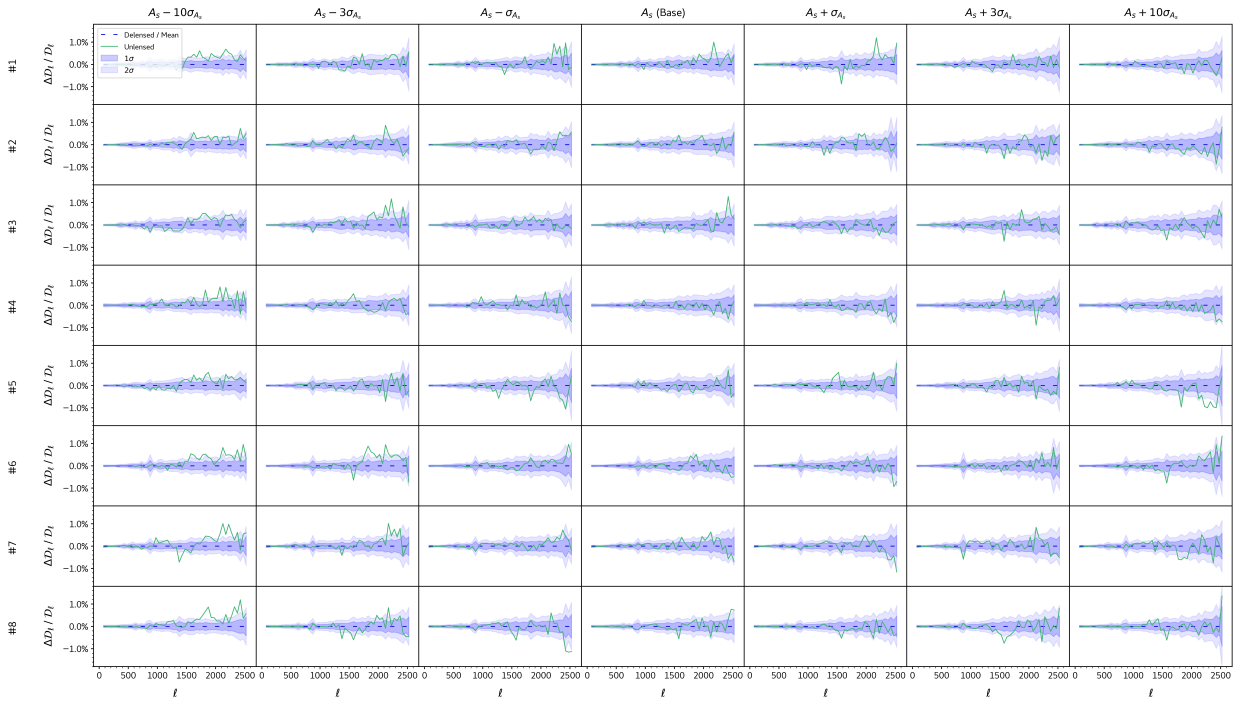

Fig. 4.13. Power spectra of out-of-distribution test examples for different A_s values, with every other cosmological parameter unchanged. Each column corresponds to an A_s value, which differs from the training value by a factor of the Planck measurement error σ_{A_s}. Each row corresponds to the same noise realization applied during data generation to the lensed and unlensed maps in Fourier space. Each panel depicts the relative difference of power spectra, similar to the bottom panel of Figure 4.10. Unless A_s is altered by a large factor of σ_{A_s}, the unlensed (target) spectrum stays mostly within the 2σ region of the mean delensed spectrum.

4. Discussion

As discussed in Section 2, we chose to leverage VAEs' inference model in developing our framework for two reasons. Firstly, VAEs exhibit efficient sampling, offering lower computational cost compared to alternatives such as conventional diffusion models, which often demand extensive iterations to generate a sample. This brings about a significant advantage for large datasets and high dimensions. Furthermore, the well-understood approximate inference framework of VAEs was another crucial factor in our choice. This framework elucidates the probabilistic behavior of VAEs and provides a foundation to address potential limitations.

According to the results presented in Section 3, our model's capability to accurately learn the posterior distribution of rotated GRFs is demonstrated by comparing the empirical and theoretical moments of the posterior (Figure 4.16). The samples exhibit variability and the desired resemblance to the ground truth, both in the GRF rotation and CMB delensing

experiments (Figures 4.16 and Figure 4.19). Furthermore, analyzing power spectra in the CMB delensing experiment reveals that the unlensed power spectrum lies almost always within 2σ of the mean delensed power spectrum (Figure 4.10), all of which verify the accuracy of the model. Our model is also robust against out-of-distribution examples within at least 1σ of Planck measurement errors. It exhibits greater resilience to variations in the scalar amplitude A_s compared to the matter density parameter Ω_m (Figures 4.12 and 4.13).

That being said, our evaluations using the TARP coverage probability test indicate that our model provides conservative uncertainty estimates (Figures 4.7 and 4.11). While this "cautiousness" ensures that our uncertainty estimates contain the true values with higher probability, it could potentially hinder valid physical interpretations. For instance, it can impede the ability to detect anomalies and outliers or introduce a false sense of agreement between inconsistent models. There might be situations where two inconsistent estimates appear to agree with each other within their wide uncertainty bounds. A possible resolution for this behavior is to use a more effective optimization scheme. Furthermore, with more computational cost, one can generate more samples to calculate the GNLL reconstruction loss during the Noise Net's training. Finally, one can think of designing specialized layers that help the model extract statistical information more efficiently from its training data, or to train the model with additional constraints specific to the problem at hand (e.g., statistical isotropy in case of CMB delensing).

5. Conclusion

In this paper, we presented a framework for **high-dimensional posterior sampling** based on the Hierarchical Probabilistic U-Net (HPU-Net) architecture. The task is done using two networks: A vanilla U-Net that learns the mean of the posterior distribution (Mean Net) and an HPU-Net that can generate the deviation of the posterior samples from the posterior mean (Noise Net). Our method provides a **fast** way to learn high-dimensional posteriors and is supported by a **sound theoretical framework** that enables the acquisition of probability distributions in a more principled manner. Additionally, its **robustness** against variabilities in input distribution makes it more applicable to real observational data. Our model suits various **implicit-likelihood** inference scenarios where direct likelihood modeling is not feasible. Furthermore, rather than point estimates with no measure of uncertainty, our model provides samples from the posterior distribution. It enables calculating **uncertainty estimates** for predictions, which is crucial for every physical measurement. This is a step towards making deep learning models more suitable for physical applications, making them even more powerful to facilitate scientific discoveries.

Software

This research made use of PyTorch [**25**], NumPy [**12**], matplotlib [**16**], and TensorBoard [**1**]. We acknowledge the use of the GPT-3.5 model by OpenAI for providing guidance and clarifications during the course of this research.

Acknowledgment

This work is supported by the Simons Collaboration on "Learning the Universe". It was enabled in part by computational resources provided by Calcul Quebec, Compute Canada, and the Digital Research Alliance of Canada. MH.S. was supported by an MSc Excellence Scholarship from IVADO.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: Large-scale machine learning on heterogeneous distributed systems, 2016.

[2] Jonas Adler and Ozan Öktem. Deep bayesian inversion, 2018.

[3] and N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo, S. Basak, K. Benabed, J.-P. Bernard, M. Bersanelli, P. Bielewicz, J. J. Bock, J. R. Bond, J. Borrill, F. R. Bouchet, F. Boulanger, M. Bucher, C. Burigana, E. Calabrese, J.-F. Cardoso, J. Carron, A. Challinor, H. C. Chiang, L. P. L. Colombo, C. Combet, B. P. Crill, F. Cuttaia, P. de Bernardis, G. de Zotti, J. Delabrouille, E. Di Valentino, J. M. Diego, O. Doré , M. Douspis, A. Ducout, X. Dupac, G. Efstathiou, F. Elsner, T. A. Enßlin, H. K. Eriksen, Y. Fantaye, R. Fernandez-Cobos, F. Finelli, F. Forastieri, M. Frailis, A. A. Fraisse, E. Franceschi, A. Frolov, S. Galeotta, S. Galli, K. Ganga, R. T. Génova-Santos, M. Gerbino, T. Ghosh, J. González-Nuevo, K. M. Górski, S. Gratton, A. Gruppuso, J. E. Gudmundsson, J. Hamann, W. Handley, F. K. Hansen, D. Herranz, E. Hivon, Z. Huang, A. H. Jaffe, W. C. Jones, A. Karakci, E. Keihänen, R. Keskitalo, K. Kiiveri, J. Kim, L. Knox, N. Krachmalnicoff, M. Kunz, H. Kurki-Suonio, G. Lagache, J.-M. Lamarre, A. Lasenby, M. Lattanzi, C. R. Lawrence, M. Le Jeune, F. Levrier, A. Lewis, M. Liguori, P. B. Lilje, V. Lindholm, M. López-Caniego, P. M. Lubin, Y.-Z. Ma, J. F. Macías-Pérez, G. Maggio, D. Maino, N. Mandolesi, A. Mangilli, A. Marcos-Caballero, M. Maris, P. G. Martin, E. Martínez-González, S. Matarrese, N. Mauri, J. D. McEwen, A. Melchiorri, A. Mennella, M. Migliaccio, M.-A. Miville-Deschênes, D. Molinari, A. Moneti, L. Montier, G. Morgante, A. Moss, P. Natoli, L. Pagano, D. Paoletti, B. Partridge, G. Patanchon, F. Perrotta, V. Pettorino, F. Piacentini, L. Polastri, G. Polenta, J.-L. Puget, J. P. Rachen, M. Reinecke, M. Remazeilles, A. Renzi, G. Rocha, C. Rosset, G. Roudier, J. A. Rubiño-Martín, B. Ruiz-Granados, L. Salvati, M. Sandri, M. Savelainen, D. Scott, C. Sirignano, R. Sunyaev, A.-S. Suur-Uski, J. A. Tauber, D. Tavagnacco, M. Tenti, L. Toffolatti, M. Tomasi, T. Trombetti, J. Valiviita, B. Van Tent, P. Vielva, F. Villa, N. Vittorio, B. D. Wandelt, I. K. Wehus, M. White, S. D. M. White, A. Zacchei, and A. Zonca. iplanck/i2018 results. *Astronomy &amp Astrophysics*, 641:A8, sep 2020.

[4] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G. Willcocks. Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7327–7347, nov 2022.

[5] J. Caldeira, W.L.K. Wu, B. Nord, C. Avestruz, S. Trivedi, and K.T. Story. DeepCMB: Lensing reconstruction of the cosmic microwave background with deep neural networks. *Astronomy and Computing*, 28:100307, jul 2019.

[6] Julien Carron and Antony Lewis. Maximum a posteriori cmb lensing reconstruction. *Phys. Rev. D*, 96:063510, Sep 2017.

[7] Julien Carron, Mark Mirmelstein, and Antony Lewis. CMB lensing from planck PR4 maps. *Journal of Cosmology and Astroparticle Physics*, 2022(09):039, sep 2022.

[8] P. Diego-Palazuelos, P. Vielva, E. Martí nez-González, and R.B. Barreiro. Comparison of delensing methodologies and assessment of the delensing capabilities of future experiments. *Journal of Cosmology and Astroparticle Physics*, 2020(11):058–058, nov 2020.

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[10] Boryana Hadzhiyska, Blake D. Sherwin, Mathew Madhavacheril, and Simone Ferraro. Improving small-scale cmb lensing reconstruction. *Phys. Rev. D*, 100:023547, Jul 2019.

[11] Duncan Hanson, Anthony Challinor, and Antony Lewis. Weak lensing of the CMB. *General Relativity and Gravitation*, 42(9):2197–2218, jun 2010.

[12] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, sep 2020.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[15] Wayne Hu. Mapping the dark matter through the cosmic microwave background damping tail. *The Astrophysical Journal*, 557(2):L79, jul 2001.

[16] John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science Engineering*, 9(3):90–95, 2007.

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[18] Simon A. A. Kohl, Bernardino Romera-Paredes, Klaus H. Maier-Hein, Danilo Jimenez Rezende, S. M. Ali Eslami, Pushmeet Kohli, Andrew Zisserman, and Olaf Ronneberger. A hierarchical probabilistic u-net for modeling multi-scale ambiguities, 2019.

[19] Pablo Lemos, Adam Coogan, Yashar Hezaveh, and Laurence Perreault-Levasseur. Sampling-based accuracy testing of posterior estimators for general inference, 2023.

[20] A LEWIS and A CHALLINOR. Weak gravitational lensing of the CMB. *Physics Reports*, 429(1):1–65, jun 2006.

[21] Antony Lewis and Sarah Bridle. Cosmological parameters from CMB and other data: A Monte Carlo approach. , 66:103511, 2002.

[22] Peikai Li, Ipek Ilayda Onur, Scott Dodelson, and Shreyas Chaudhari. High-resolution cmb lensing reconstruction with deep learning, 2022.

[23] Abhishek S. Maniyar, Yacine Ali-Haïmoud, Julien Carron, Antony Lewis, and Mathew S. Madhavacheril. Quadratic estimators for cmb weak lensing. *Phys. Rev. D*, 103:083524, Apr 2021.

[24] Marius Millea, Ethan Anderes, and Benjamin D. Wandelt. Bayesian delensing of CMB temperature and polarization. *Physical Review D*, 100(2), jul 2019.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[26] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows, 2016.

[27] Danilo Jimenez Rezende and Fabio Viola. Taming vaes, 2018.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[29] Kendrick M. Smith, Duncan Hanson, Marilena LoVerde, Christopher M. Hirata, and Oliver Zahn. Delensing cmb polarization with external datasets. *Journal of Cosmology and Astroparticle Physics*, 2012(06):014, jun 2012.

[30] Ye-Peng Yan, Guo-Jian Wang, Si-Yu Li, and Jun-Qing Xia. Delensing of cosmic microwave background polarization with machine learning. *The Astrophysical Journal Supplement Series*, 267(1):2, jun 2023.

[31] Ye-Peng Yan, Guo-Jian Wang, Si-Yu Li, Yang-Jie Yan, and Jun-Qing Xia. Lensing reconstruction from the cosmic microwave background polarization with machine learning. *The Astrophysical Journal*, 952(1):15, jul 2023.

Appendices

A: Additional Plots for the GRF Rotation Experiment

This appendix presents moment comparison plots for three additional test examples of the GRF rotation experiment. For each example, 1000 posterior samples were drawn using the model and used to calculate pixel-wise empirical means and standard deviations.

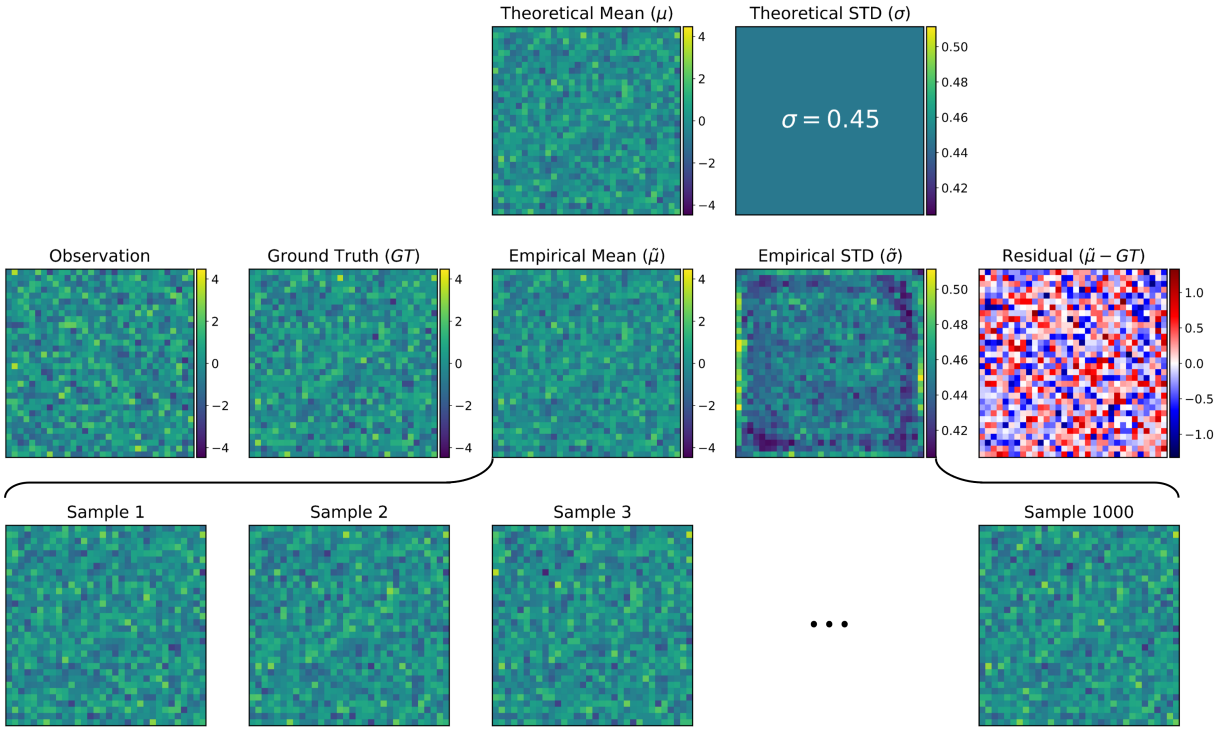

Fig. 4.14

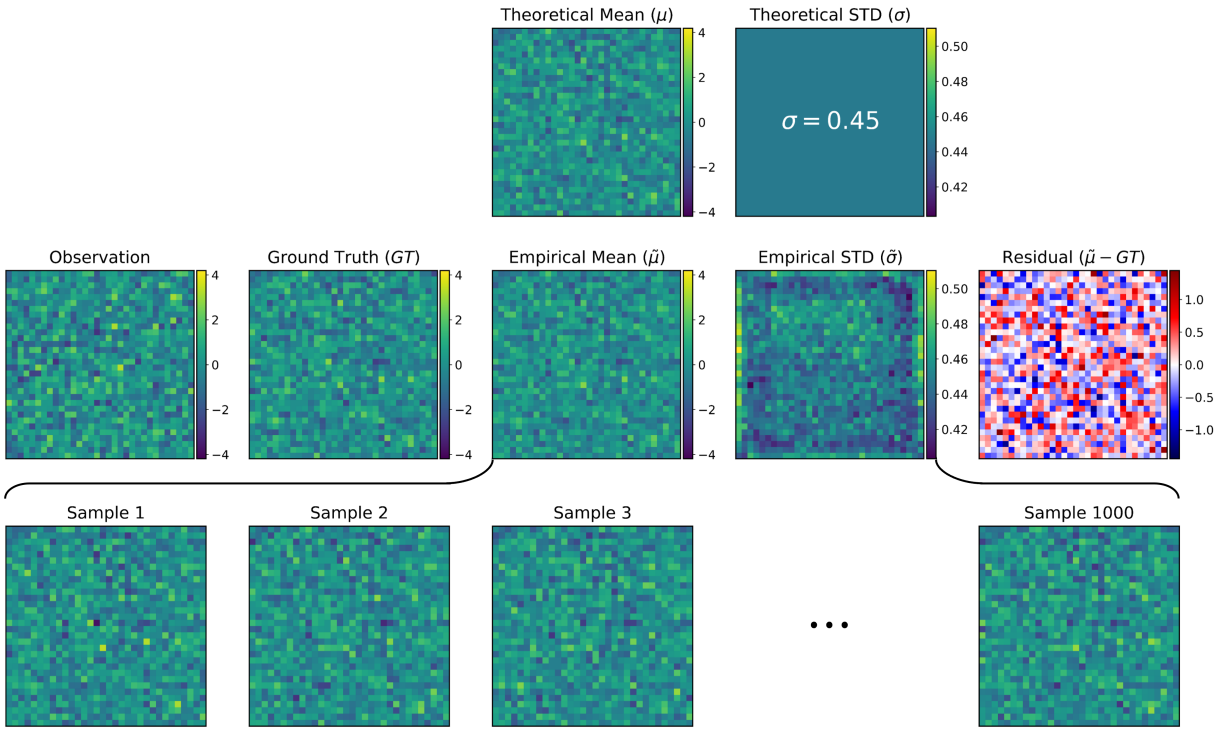

Fig. 4.15

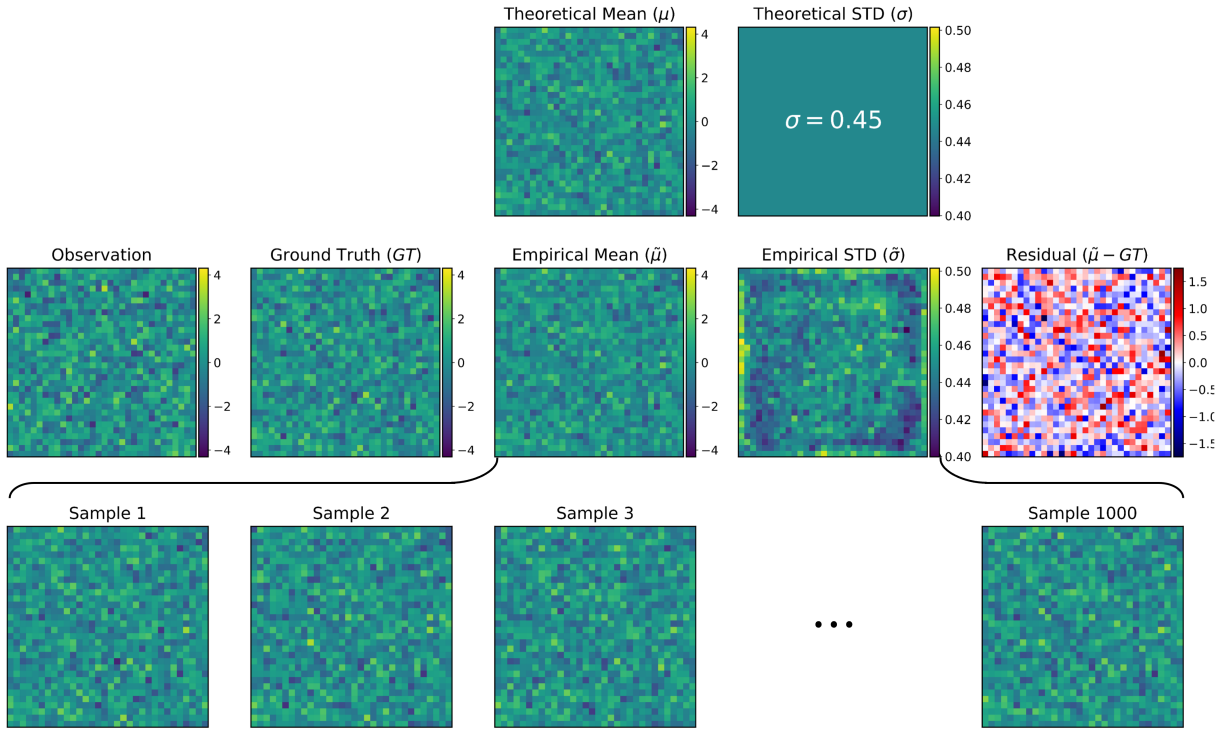

Fig. 4.16

B: Additional Plots for the CMB Delensing Experiment

This appendix presents moment comparison plots for three additional test examples of the CMB delensing experiment. For each example, 1000 posterior samples were drawn using the model and used to calculate pixel-wise empirical means and standard deviations.

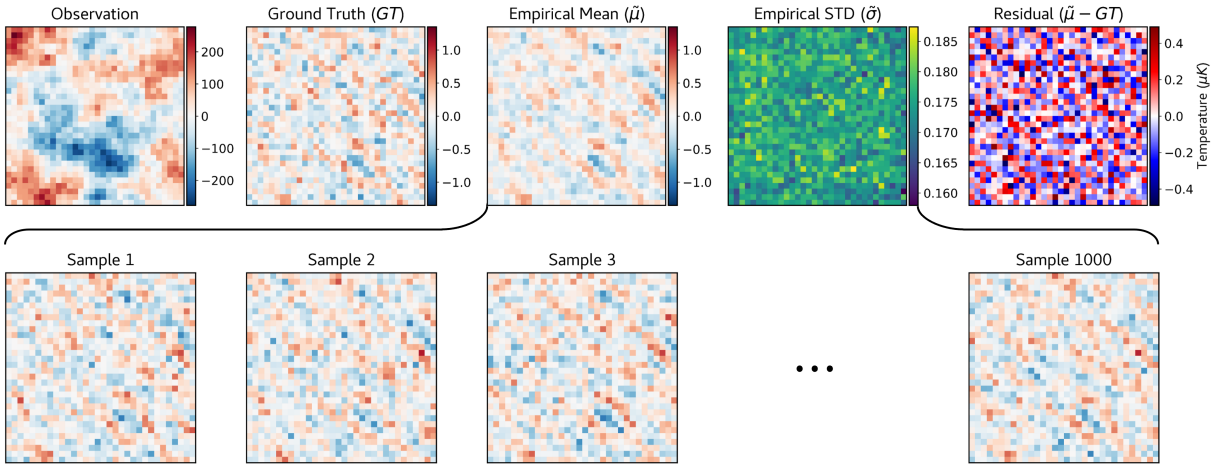

Fig. 4.17

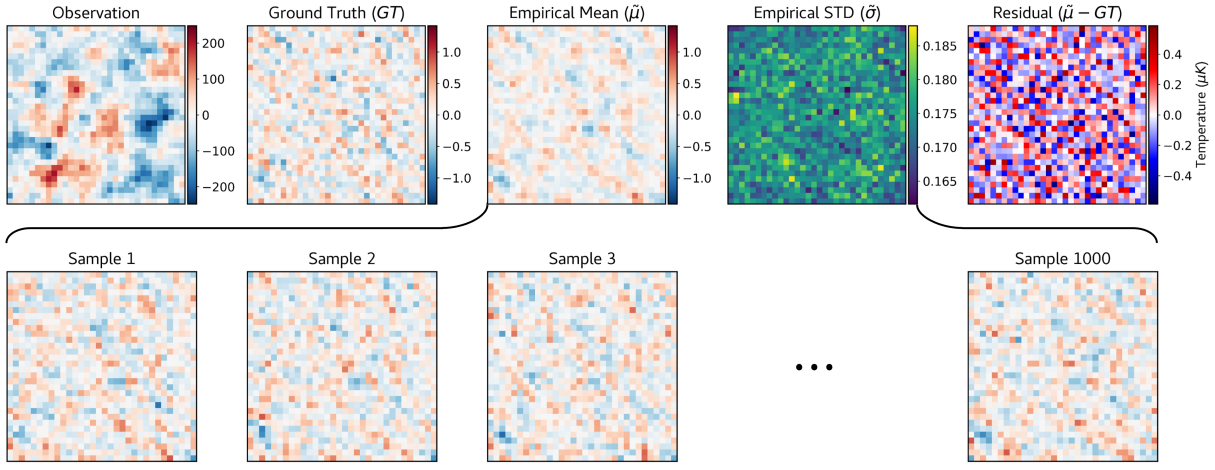

Fig. 4.18

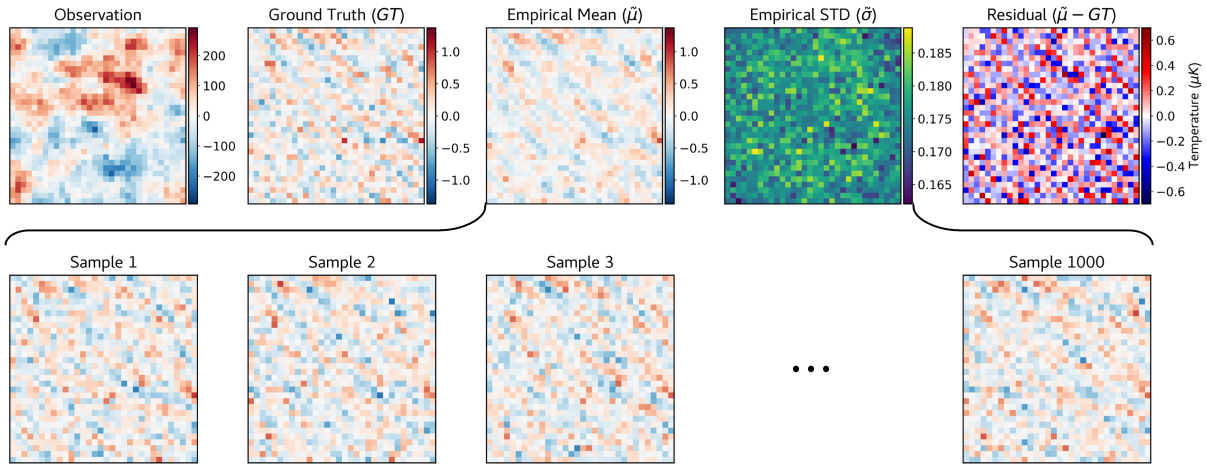
Observation
Ground Truth (GT)
Empirical Mean ($\tilde{\mu}$)
Empirical STD ($\tilde{\sigma}$)
Residual ($\tilde{\mu} - GT$)
200
100
0
−100
−200
1.0
0.5
0.0
−0.5
−1.0
1.0
0.5
0.0
−0.5
−1.0
0.185
0.180
0.175
0.170
0.165
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
Temperature (μK)
Sample 1
Sample 2
Sample 3
Sample 1000

Fig. 4.19

Chapter 5

Conclusion

In this thesis, we explored the development of a deep learning framework for fast high-dimensional posterior sampling at pixel level and its application to remove the weak lensing effect from the CMB. Our model does not require direct modeling or evaluation of likelihood. Hence, it does not rely on simplifying assumptions often employed by explicit likelihood inference methods, which enables it to exploit the full potential of the observed data. Contrary to the previous deep learning approaches to CMB delensing that only provide point estimates with no measure of uncertainty, our model generates posterior samples that can be used to derive uncertainty estimates for the inferred parameters. This, together with the model's robust performance against out-of-distribtion examples, make it suitable for physical applications, especially for problems dealing with real observational data.

Our evaluations indicate that our model can accurately learn the posterior in a problem that the actual posterior is analytically accessible. Furthermore, for the CMB delensing problem, the power spectra of its posterior samples contain the unlensed (target) power spectrum mostly within their 2σ confidence region. However, an analysis of its credible regions using coverage probability test reveals that our model's uncertainty estimates are yet to be perfectly calibrated. Instead, it provides conservative (i.e., larger-than-expected) uncertainty estimates, which can possibly lead to invalid physical interpretations. For instance, it can prevent the detection of anomalies and outliers or incorrectly indicate agreement among inconsistent models. Possible workarounds for this issue are improving the model's optimization scheme, generating more samples to calculate the reconstruction loss, and using more informative learning objectives.

Having the model's underconfidence resolved, it can be applied to numerous physical and astrophysical problems including and beyond CMB delensing, e.g., source reconstruction in strong gravitational lensing, reconstructing the initial conditions of the universe using past light cone observations, and evolution of galaxies and galaxy clusters. All mentioned

areas require novel data analysis methods to benefit from the wealth of high-resolution data provided by next-generation observational missions. Our work brings an important contribution to high-dimensional Bayesian inference in physical problems, facilitating data-driven discoveries and improving our knowledge of fundamental physics in the *data-intensive science* era.

References

[1] *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC, Sep 2018.

[2] Kevork N. Abazajian, Peter Adshead, Zeeshan Ahmed, Steven W. Allen, David Alonso, Kam S. Arnold, Carlo Baccigalupi, James G. Bartlett, Nicholas Battaglia, Bradford A. Benson, Colin A. Bischoff, Julian Borrill, Victor Buza, Erminia Calabrese, Robert Caldwell, John E. Carlstrom, Clarence L. Chang, Thomas M. Crawford, Francis-Yan Cyr-Racine, Francesco De Bernardis, Tijmen de Haan, Sperello di Serego Alighieri, Joanna Dunkley, Cora Dvorkin, Josquin Errard, Giulio Fabbian, Stephen Feeney, Simone Ferraro, Jeffrey P. Filippini, Raphael Flauger, George M. Fuller, Vera Gluscevic, Daniel Green, Daniel Grin, Evan Grohs, Jason W. Henning, J. Colin Hill, Renee Hlozek, Gilbert Holder, William Holzapfel, Wayne Hu, Kevin M. Huffenberger, Reijo Keskitalo, Lloyd Knox, Arthur Kosowsky, John Kovac, Ely D. Kovetz, Chao-Lin Kuo, Akito Kusaka, Maude Le Jeune, Adrian T. Lee, Marc Lilley, Marilena Loverde, Mathew S. Madhavacheril, Adam Mantz, David J. E. Marsh, Jeffrey McMahon, Pieter Daniel Meerburg, Joel Meyers, Amber D. Miller, Julian B. Munoz, Ho Nam Nguyen, Michael D. Niemack, Marco Peloso, Julien Peloton, Levon Pogosian, Clement Pryke, Marco Raveri, Christian L. Reichardt, Graca Rocha, Aditya Rotti, Emmanuel Schaan, Marcel M. Schmittfull, Douglas Scott, Neelima Sehgal, Sarah Shandera, Blake D. Sherwin, Tristan L. Smith, Lorenzo Sorbo, Glenn D. Starkman, Kyle T. Story, Alexander van Engelen, Joaquin D. Vieira, Scott Watson, Nathan Whitehorn, and W. L. Kimmy Wu. Cmb-s4 science book, first edition, 2016.

[3] Peter Ade, James Aguirre, Zeeshan Ahmed, Simone Aiola, Aamir Ali, David Alonso, Marcelo A. Alvarez, Kam Arnold, Peter Ashton, Jason Austermann, Humna Awan, Carlo Baccigalupi, Taylor Baildon, Darcy Barron, Nick Battaglia, Richard Battye, Eric Baxter, Andrew Bazarko, James A. Beall, Rachel Bean, Dominic Beck, Shawn Beckman, Benjamin Beringue, Federico Bianchini, Steven Boada, David Boettger, J. Richard Bond, Julian Borrill, Michael L. Brown, Sarah Marie Bruno, Sean Bryan, Erminia Calabrese, Victoria Calafut, Paolo Calisse, Julien Carron, Anthony Challinor, Grace Chesmore, Yuji Chinone, Jens Chluba, Hsiao-Mei Sherry Cho, Steve Choi, Gabriele Coppi, Nicholas F. Cothard, Kevin Coughlin, Devin Crichton, Kevin D. Crowley, Kevin T. Crowley, Ari Cukierman, John M. D'Ewart, Rolando Dünner, Tijmen de Haan, Mark Devlin, Simon Dicker, Joy Didier, Matt Dobbs, Bradley Dober, Cody J. Duell, Shannon Duff, Adri Duivenvoorden, Jo Dunkley, John Dusatko, Josquin Errard, Giulio Fabbian, Stephen Feeney, Simone Ferraro, Pedro Fluxà , Katherine Freese, Josef C. Frisch, Andrei Frolov, George Fuller, Brittany Fuzia, Nicholas Galitzki, Patricio A. Gallardo, Jose Tomas Galvez Ghersi, Jiansong Gao, Eric Gawiser, Martina Gerbino, Vera Gluscevic, Neil Goeckner-Wald, Joseph Golec, Sam Gordon, Megan Gralla, Daniel Green, Arpi Grigorian, John Groh, Chris Groppi, Yilun Guan, Jon E. Gudmundsson, Dongwon Han, Peter Hargrave, Masaya Hasegawa, Matthew Hasselfield, Makoto Hattori, Victor Haynes, Masashi Hazumi, Yizhou He, Erin Healy, Shawn W. Henderson, Carlos Hervias-Caimapo, Charles A. Hill, J. Colin Hill, Gene Hilton, Matt Hilton, Adam D. Hincks, Gary Hinshaw, Renée Hložek, Shirley

Ho, Shuay-Pwu Patty Ho, Logan Howe, Zhiqi Huang, Johannes Hubmayr, Kevin Huffenberger, John P. Hughes, Anna Ijjas, Margaret Ikape, Kent Irwin, Andrew H. Jaffe, Bhuvnesh Jain, Oliver Jeong, Daisuke Kaneko, Ethan D. Karpel, Nobuhiko Katayama, Brian Keating, Sarah S. Kernasovskiy, Reijo Keskitalo, Theodore Kisner, Kenji Kiuchi, Jeff Klein, Kenda Knowles, Brian Koopman, Arthur Kosowsky, Nicoletta Krachmalnicoff, Stephen E. Kuenstner, Chao-Lin Kuo, Akito Kusaka, Jacob Lashner, Adrian Lee, Eunseong Lee, David Leon, Jason S.-Y. Leung, Antony Lewis, Yaqiong Li, Zack Li, Michele Limon, Eric Linder, Carlos Lopez-Caraballo, Thibaut Louis, Lindsay Lowry, Marius Lungu, Mathew Madhavacheril, Daisy Mak, Felipe Maldonado, Hamdi Mani, Ben Mates, Frederick Matsuda, Loïc Maurin, Phil Mauskopf, Andrew May, Nialh McCallum, Chris McKenney, Jeff McMahon, P. Daniel Meerburg, Joel Meyers, Amber Miller, Mark Mirmelstein, Kavilan Moodley, Moritz Munchmeyer, Charles Munson, Sigurd Naess, Federico Nati, Martin Navaroli, Laura Newburgh, Ho Nam Nguyen, Michael Niemack, Haruki Nishino, John Orlowski-Scherer, Lyman Page, Bruce Partridge, Julien Peloton, Francesca Perrotta, Lucio Piccirillo, Giampaolo Pisano, Davide Poletti, Roberto Puddu, Giuseppe Puglisi, Chris Raum, Christian L. Reichardt, Mathieu Remazeilles, Yoel Rephaeli, Dominik Riechers, Felipe Rojas, Anirban Roy, Sharon Sadeh, Yuki Sakurai, Maria Salatino, Mayuri Sathyanarayana Rao, Emmanuel Schaan, Marcel Schmittfull, Neelima Sehgal, Joseph Seibert, Uros Seljak, Blake Sherwin, Meir Shimon, Carlos Sierra, Jonathan Sievers, Precious Sikhosana, Maximiliano Silva-Feaver, Sara M. Simon, Adrian Sinclair, Praween Siritanasak, Kendrick Smith, Stephen R. Smith, David Spergel, Suzanne T. Staggs, George Stein, Jason R. Stevens, Radek Stompor, Aritoki Suzuki, Osamu Tajima, Satoru Takakura, Grant Teply, Daniel B. Thomas, Ben Thorne, Robert Thornton, Hy Trac, Calvin Tsai, Carole Tucker, Joel Ullom, Sunny Vagnozzi, Alexander van Engelen, Jeff Van Lanen, Daniel D. Van Winkle, Eve M. Vavagiakis, Clara Vergès, Michael Vissers, Kasey Wagoner, Samantha Walker, Jon Ward, Ben Westbrook, Nathan Whitehorn, Jason Williams, Joel Williams, Edward J. Wollack, Zhilei Xu, Byeonghee Yu, Cyndia Yu, Fernando Zago, Hezi Zhang, and Ningfeng Zhu. The simons observatory: science goals and forecasts. *Journal of Cosmology and Astroparticle Physics*, 2019(02):056–056, feb 2019.

[4] and N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo, S. Basak, K. Benabed, J.-P. Bernard, M. Bersanelli, P. Bielewicz, J. J. Bock, J. R. Bond, J. Borrill, F. R. Bouchet, F. Boulanger, M. Bucher, C. Burigana, R. C. Butler, E. Calabrese, J.-F. Cardoso, J. Carron, B. Casaponsa, A. Challinor, H. C. Chiang, L. P. L. Colombo, C. Combet, B. P. Crill, F. Cuttaia, P. de Bernardis, A. de Rosa, G. de Zotti, J. Delabrouille, J.-M. Delouis, E. Di Valentino, J. M. Diego, O. Doré , M. Douspis, A. Ducout, X. Dupac, S. Dusini, G. Efstathiou, F. Elsner, T. A. Enßlin, H. K. Eriksen, Y. Fantaye, R. Fernandez-Cobos, F. Finelli, M. Frailis, A. A. Fraisse, E. Franceschi, A. Frolov, S. Galeotta, S. Galli, K. Ganga, R. T. Génova-Santos, M. Gerbino, T. Ghosh, Y. Giraud-Héraud, J. González-Nuevo, K. M. Górski, S. Gratton, A. Gruppuso, J. E. Gudmundsson, J. Hamann, W. Handley, F. K. Hansen, D. Herranz, E. Hivon, Z. Huang, A. H. Jaffe, W. C. Jones, E. Keihänen, R. Keskitalo, K. Kiiveri, J. Kim, T. S. Kisner, N. Krachmalnicoff, M. Kunz, H. Kurki-Suonio, G. Lagache, J.-M. Lamarre, A. Lasenby, M. Lattanzi, C. R. Lawrence, M. Le Jeune, F. Levrier, A. Lewis, M. Liguori, P. B. Lilje, M. Lilley, V. Lindholm, M. López-Caniego, P. M. Lubin, Y.-Z. Ma, J. F. Macías-Pérez, G. Maggio, D. Maino, N. Mandolesi, A. Mangilli, A. Marcos-Caballero, M. Maris, P. G. Martin, E. Martínez-González, S. Matarrese, N. Mauri, J. D. McEwen, P. R. Meinhold, A. Melchiorri, A. Mennella, M. Migliaccio, M. Millea, M.-A. Miville-Deschênes, D. Molinari, A. Moneti, L. Montier, G. Morgante, A. Moss, P. Natoli, H. U. Nørgaard-Nielsen, L. Pagano, D. Paoletti, B. Partridge, G. Patanchon, H. V. Peiris, F. Perrotta, V. Pettorino, F. Piacentini, G. Polenta, J.-L. Puget, J. P. Rachen, M. Reinecke, M. Remazeilles, A. Renzi, G. Rocha, C. Rosset, G. Roudier, J. A.

Rubiño-Martín, B. Ruiz-Granados, L. Salvati, M. Sandri, M. Savelainen, D. Scott, E. P. S. Shellard, C. Sirignano, G. Sirri, L. D. Spencer, R. Sunyaev, A.-S. Suur-Uski, J. A. Tauber, D. Tavagnacco, M. Tenti, L. Toffolatti, M. Tomasi, T. Trombetti, J. Valiviita, B. Van Tent, P. Vielva, F. Villa, N. Vittorio, B. D. Wandelt, I. K. Wehus, A. Zacchei, and A. Zonca. iplanck/i 2018 results. *Astronomy &amp Astrophysics*, 641:A5, sep 2020.

[5] J. E. Austermann, K. A. Aird, J. A. Beall, D. Becker, A. Bender, B. A. Benson, L. E. Bleem, J. Britton, J. E. Carlstrom, C. L. Chang, H. C. Chiang, H.-M. Cho, T. M. Crawford, A. T. Crites, A. Datesman, T. de Haan, M. A. Dobbs, E. M. George, N. W. Halverson, N. Harrington, J. W. Henning, G. C. Hilton, G. P. Holder, W. L. Holzapfel, S. Hoover, N. Huang, J. Hubmayr, K. D. Irwin, R. Keisler, J. Kennedy, L. Knox, A. T. Lee, E. Leitch, D. Li, M. Lueker, D. P. Marrone, J. J. McMahon, J. Mehl, S. S. Meyer, T. E. Montroy, T. Natoli, J. P. Nibarger, M. D. Niemack, V. Novosad, S. Padin, C. Pryke, C. L. Reichardt, J. E. Ruhl, B. R. Saliwanchik, J. T. Sayre, K. K. Schaffer, E. Shirokoff, A. A. Stark, K. Story, K. Vanderlinde, J. D. Vieira, G. Wang, R. Williamson, V. Yefremenko, K. W. Yoon, and O. Zahn. SPTpol: an instrument for CMB polarization measurements with the south pole telescope. In Wayne S. Holland, editor, *SPIE Proceedings*. SPIE, sep 2012.

[6] Claudine Badue, Rânik Guidolini, Raphael Vivacqua Carneiro, Pedro Azevedo, Vinicius B. Cardoso, Avelino Forechi, Luan Jesus, Rodrigo Berriel, Thiago M. Paixão, Filipe Mutz, Lucas de Paula Veronese, Thiago Oliveira-Santos, and Alberto F. De Souza. Self-driving cars: A survey. *Expert Systems with Applications*, 165:113816, 2021.

[7] C. L. Bennett, M. Bay, M. Halpern, G. Hinshaw, C. Jackson, N. Jarosik, A. Kogut, M. Limon, S. S. Meyer, L. Page, D. N. Spergel, G. S. Tucker, D. T. Wilkinson, E. Wollack, and E. L. Wright. Theimicrowave anisotropy probe/imission. *The Astrophysical Journal*, 583(1):1–23, jan 2003.

[8] Michael Betancourt. A conceptual introduction to hamiltonian monte carlo, 2018.

[9] S. Birrer, M. Millon, D. Sluse, A. J. Shajib, F. Courbin, L. V. E. Koopmans, S. H. Suyu, and T. Treu. Time-delay cosmography: Measuring the hubble constant and other cosmological parameters with strong gravitational lensing, 2023.

[10] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, apr 2017.

[11] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[12] J. E. Carlstrom, P. A. R. Ade, K. A. Aird, B. A. Benson, L. E. Bleem, S. Busetti, C. L. Chang, E. Chauvin, H.-M. Cho, T. M. Crawford, A. T. Crites, M. A. Dobbs, N. W. Halverson, S. Heimsath, W. L. Holzapfel, J. D. Hrubes, M. Joy, R. Keisler, T. M. Lanting, A. T. Lee, E. M. Leitch, J. Leong, W. Lu, M. Lueker, D. Luong-Van, J. J. McMahon, J. Mehl, S. S. Meyer, J. J. Mohr, T. E. Montroy, S. Padin, T. Plagge, C. Pryke, J. E. Ruhl, K. K. Schaffer, D. Schwan, E. Shirokoff, H. G. Spieler, Z. Staniszewski, A. A. Stark, C. Tucker, K. Vanderlinde, J. D. Vieira, and R. Williamson. The 10 meter south pole telescope. *Publications of the Astronomical Society of the Pacific*, 123(903):568–581, may 2011.

[13] J. H. H. Chan, C. Lemon, F. Courbin, R. Gavazzi, B. Clé ment, M. Millon, E. Paic, K. Rojas, E. Savary, G. Vernardos, J.-C. Cuillandre, S. Fabbro, S. Gwyn, M. J. Hudson, M. Kilbinger, and A. McConnachie.

Discovery of strongly lensed quasars in the ultraviolet near infrared optical northern survey (UNIONS). *Astronomy &amp Astrophysics*, 659:A140, mar 2022.

[14] Wenlei Chen, Patrick L. Kelly, Masamune Oguri, Thomas J. Broadhurst, Jose M. Diego, Najmeh Emami, Alexei V. Filippenko, Tommaso L. Treu, and Adi Zitrin. Shock cooling of a red-supergiant supernova at redshift 3 in lensed images. *Nature*, 611(7935):256–259, nov 2022.

[15] Douglas Clowe, Maruš a Bradač, Anthony H. Gonzalez, Maxim Markevitch, Scott W. Randall, Christine Jones, and Dennis Zaritsky. A direct empirical proof of the existence of dark matter. *The Astrophysical Journal*, 648(2):L109–L113, aug 2006.

[16] The Planck Collaboration. The scientific programme of planck, 2006.

[17] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

[18] Christopher T Davies, Marius Cautun, Benjamin Giblin, Baojiu Li, Joachim Harnois-Déraps, and Yan-Chuan Cai. Constraining cosmology with weak lensing voids. *Monthly Notices of the Royal Astronomical Society*, 507(2):2267–2282, 08 2021.

[19] Sander Dieleman, Kyle W. Willett, and Joni Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450(2):1441–1459, apr 2015.

[20] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning, 2018.

[21] Abdelkarim Erradi, Waheed Iqbal, Arif Mahmood, and Athman Bouguettaya. Web application resource requirements estimation based on the workload latent features. *IEEE Transactions on Services Computing*, 14(6):1638–1649, 2021.

[22] J. W. Fowler, V. Acquaviva, P. A. R. Ade, P. Aguirre, M. Amiri, J. W. Appel, L. F. Barrientos, E. S. Battistelli, J. R. Bond, B. Brown, B. Burger, J. Chervenak, S. Das, M. J. Devlin, S. R. Dicker, W. B. Doriese, J. Dunkley, R. Dünner, T. Essinger-Hileman, R. P. Fisher, A. Hajian, M. Halpern, M. Hasselfield, C. Herná ndez-Monteagudo, G. C. Hilton, M. Hilton, A. D. Hincks, R. Hlozek, K. M. Huffenberger, D. H. Hughes, J. P. Hughes, L. Infante, K. D. Irwin, R. Jimenez, J. B. Juin, M. Kaul, J. Klein, A. Kosowsky, J. M. Lau, M. Limon, Y.-T. Lin, R. H. Lupton, T. A. Marriage, D. Marsden, K. Martocci, P. Mauskopf, F. Menanteau, K. Moodley, H. Moseley, C. B. Netterfield, M. D. Niemack, M. R. Nolta, L. A. Page, L. Parker, B. Partridge, H. Quintana, B. Reid, N. Sehgal, J. Sievers, D. N. Spergel, S. T. Staggs, D. S. Swetz, E. R. Switzer, R. Thornton, H. Trac, C. Tucker, L. Verde, R. Warne, G. Wilson, E. Wollack, and Y. Zhao. THE ATACAMA COSMOLOGY TELESCOPE: A MEASUREMENT OF THE 600 &lt ℓ &lt 8000 COSMIC MICROWAVE BACKGROUND POWER SPECTRUM AT 148 GHz. *The Astrophysical Journal*, 722(2):1148–1161, sep 2010.

[23] Wendy L. Freedman. Measuring cosmologicalparameters. *Proceedings of the National Academy of Sciences*, 95(1):2–7, 1998.

[24] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 1980.

[25] T. Futamase and S. Yoshida. Possible measurement of quintessence and density parameter using strong gravitational lensing events. *Progress of Theoretical Physics*, 105(5):887–891, may 2001.

[26] A. Goobar, R. Amanullah, S. R. Kulkarni, P. E. Nugent, J. Johansson, C. Steidel, D. Law, E. Mörtsell, R. Quimby, N. Blagorodnova, A. Brandeker, Y. Cao, A. Cooray, R. Ferretti, C. Fremling, L. Hangard, M. Kasliwal, T. Kupfer, R. Lunnan, F. Masci, A. A. Miller, H. Nayyeri, J. D. Neill, E. O. Ofek, S. Papadogiannakis, T. Petrushevska, V. Ravi, J. Sollerman, M. Sullivan, F. Taddia, R. Walters, D. Wilson,

L. Yan, and O. Yaron. iPTF16geu: A multiply imaged, gravitationally lensed type ia supernova. *Science*, 356(6335):291–295, apr 2017.

[27] Ariel Goobar, Joel Johansson, Steve Schulze, Nikki Arendse, Ana Sagués Carracedo, Suhail Dhawan, Edvard Mörtsell, Christoffer Fremling, Lin Yan, Daniel Perley, Jesper Sollerman, Rémy Joseph, K-Ryan Hinds, William Meynardie, Igor Andreoni, Eric Bellm, Josh Bloom, Thomas E. Collett, Andrew Drake, Matthew Graham, Mansi Kasliwal, Shri R. Kulkarni, Cameron Lemon, Adam A. Miller, James D. Neill, Jakob Nordin, Justin Pierel, Johan Richard, Reed Riddle, Mickael Rigault, Ben Rusholme, Yashvi Sharma, Robert Stein, Gabrielle Stewart, Alice Townsend, Jozsef Vinko, J. Craig Wheeler, and Avery Wold. Uncovering a population of gravitational lens galaxies with magnified standard candle SN zwicky. *Nature Astronomy*, jun 2023.

[28] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[29] C. Grillo, M. Lombardi, and G. Bertin. Cosmological parameters from strong gravitational lensing and stellar dynamics in elliptical galaxies. *Astronomy &amp Astrophysics*, 477(2):397–406, nov 2007.

[30] Duncan Hanson, Anthony Challinor, and Antony Lewis. Weak lensing of the CMB. *General Relativity and Gravitation*, 42(9):2197–2218, jun 2010.

[31] M R S Hawkins. The double quasar Q2138-431: detection of a lensing galaxy. *Monthly Notices of the Royal Astronomical Society*, 503(3):3848–3855, 03 2021.

[32] A. F. Heavens, T. D. Kitching, and A. N. Taylor. Measuring dark energy properties with 3D cosmic shear. *Monthly Notices of the Royal Astronomical Society*, 373(1):105–120, 10 2006.

[33] Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, Volodimir Begy, and Gilles Louppe. A trust crisis in simulation-based inference? your posterior approximations can be unfaithful, 2022.

[34] I. Heywood, E. J. Murphy, E. F. Jiménez-Andrade, L. Armus, W. D. Cotton, C. DeCoursey, M. Dickinson, T. J. W. Lazio, E. Momjian, K. Penner, I. Smail, and O. M. Smirnov. The vla frontier fields survey: Deep, high-resolution radio imaging of the macs lensing clusters at 3 and 6 ghz. *The Astrophysical Journal*, 910(2):105, apr 2021.

[35] Yashar D. Hezaveh, Neal Dalal, Daniel P. Marrone, Yao-Yuan Mao, Warren Morningstar, Di Wen, Roger D. Blandford, John E. Carlstrom, Christopher D. Fassnacht, Gilbert P. Holder, Athol Kemball, Philip J. Marshall, Norman Murray, Laurence Perreault Levasseur, Joaquin D. Vieira, and Risa H. Wechsler. DETECTION OF LENSING SUBSTRUCTURE USING ALMA OBSERVATIONS OF THE DUSTY GALAXY SDP.81. *The Astrophysical Journal*, 823(1):37, may 2016.

[36] J. R. Hinderks, P. Ade, J. Bock, M. Bowden, M. L. Brown, G. Cahill, J. E. Carlstrom, P. G. Castro, S. Church, T. Culverhouse, R. Friedman, K. Ganga, W. K. Gear, S. Gupta, J. Harris, V. Haynes, B. G. Keating, J. Kovac, E. Kirby, A. E. Lange, E. Leitch, O. E. Mallie, S. Melhuish, Y. Memari, A. Murphy, A. Orlando, R. Schwarz, C. O' Sullivan, L. Piccirillo, C. Pryke, N. Rajguru, B. Rusholme, A. N. Taylor, K. L. Thompson, C. Tucker, A. H. Turner, E. Y. S. Wu, and M. Zemcov. QUaD: A High-Resolution Cosmic Microwave Background Polarimeter. , 692(2):1221–1246, February 2009.

[37] Austin Hoag, Maruša Bradač, Michele Trenti, Tommaso Treu, Kasper B. Schmidt, Kuang-Han Huang, Brian C. Lemaux, Julie He, Stephanie R. Bernard, Louis E. Abramson, Charlotte A. Mason, Takahiro Morishita, Laura Pentericci, and Tim Schrabback. Spectroscopic confirmation of an ultra-faint galaxy at the epoch of reionization. *Nature Astronomy*, 1(5), apr 2017.

[38] Henk Hoekstra. Mapping the dark matter using weak lensing. *Symposium - International Astronomical Union*, 216:140–151, 2005.

[39] Lei Huang, Jie Qin, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. Normalization techniques in training dnns: Methodology, analysis and application, 2020.

[40] Pavel Hála. Spectral classification using convolutional neural networks, 2014.

[41] M. J. Jee, H. C. Ford, G. D. Illingworth, R. L. White, T. J. Broadhurst, D. A. Coe, G. R. Meurer, A. van der Wel, N. Benitez, J. P. Blakeslee, R. J. Bouwens, L. D. Bradley, R. Demarco, N. L. Homeier, A. R. Martel, and S. Mei. Discovery of a ringlike dark matter structure in the core of the galaxy cluster cl 002417. *The Astrophysical Journal*, 661(2):728–749, jun 2007.

[42] N Jeffrey, M Gatti, C Chang, L Whiteway, U Demirbozan, A Kovacs, G Pollina, D Bacon, N Hamaus, T Kacprzak, O Lahav, F Lanusse, B Mawdsley, S Nadathur, J L Starck, P Vielzeuf, D Zeurcher, A Alarcon, A Amon, K Bechtol, G M Bernstein, A Campos, A Carnero Rosell, M Carrasco Kind, R Cawthon, R Chen, A Choi, J Cordero, C Davis, J DeRose, C Doux, A Drlica-Wagner, K Eckert, F Elsner, J Elvin-Poole, S Everett, A Ferté , G Giannini, D Gruen, R A Gruendl, I Harrison, W G Hartley, K Herner, E M Huff, D Huterer, N Kuropatkin, M Jarvis, P F Leget, N MacCrann, J McCullough, J Muir, J Myles, A Navarro-Alsina, S Pandey, J Prat, M Raveri, R P Rollins, A J Ross, E S Rykoff, C Sánchez, L F Secco, I Sevilla-Noarbe, E Sheldon, T Shin, M A Troxel, I Tutusaus, T N Varga, B Yanny, B Yin, Y Zhang, J Zuntz, T M C Abbott, M Aguena, S Allam, F Andrade-Oliveira, M R Becker, E Bertin, S Bhargava, D Brooks, D L Burke, J Carretero, F J Castander, C Conselice, M Costanzi, M Crocce, L N da Costa, M E S Pereira, J De Vicente, S Desai, H T Diehl, J P Dietrich, P Doel, I Ferrero, B Flaugher, P Fosalba, J García-Bellido, E Gaztanaga, D W Gerdes, T Giannantonio, J Gschwend, G Gutierrez, S R Hinton, D L Hollowood, B Hoyle, B Jain, D J James, M Lima, M A G Maia, M March, J L Marshall, P Melchior, F Menanteau, R Miquel, J J Mohr, R Morgan, R L C Ogando, A Palmese, F Paz-Chinchón, A A Plazas, M Rodriguez-Monroy, A Roodman, E Sanchez, V Scarpine, S Serrano, M Smith, M Soares-Santos, E Suchyta, G Tarle, D Thomas, C To, and J Weller and. Dark energy survey year 3 results: Curved-sky weak lensing mass map reconstruction. *Monthly Notices of the Royal Astronomical Society*, 505(3):4626–4645, may 2021.

[43] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. *An Introduction to Variational Methods for Graphical Models*, page 105–161. MIT Press, Cambridge, MA, USA, 1999.

[44] John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andy Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David A. Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021.

[45] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019.

[46] Brian G. Keating, Peter A. R. Ade, James J. Bock, Eric Hivon, William L. Holzapfel, Andrew E. Lange, Hien Nguyen, and Ki Won Yoon. BICEP: a large angular scale CMB polarimeter. In Silvano Fineschi, editor, *Polarimetry in Astronomy*, volume 4843 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 284–295, February 2003.

[47] Patrick L. Kelly, Steven Rodney, Tommaso Treu, Masamune Oguri, Wenlei Chen, Adi Zitrin, Simon Birrer, Vivien Bonvin, Luc Dessart, Jose M. Diego, Alexei V. Filippenko, Ryan J. Foley, Daniel Gilman, Jens Hjorth, Mathilde Jauzac, Kaisey Mandel, Martin Millon, Justin Pierel, Keren Sharon, Stephen Thorp, Liliya Williams, Tom Broadhurst, Alan Dressler, Or Graur, Saurabh Jha, Curtis McCully, Marc

Postman, Kasper Borello Schmidt, Brad E. Tucker, and Anja von der Linden. Constraints on the hubble constant from supernova refsdal's reappearance. *Science*, 380(6649), jun 2023.

[48] Arash Keshavarzi Arshadi, Julia Webb, Milad Salem, Emmanuel Cruz, Stacie Calad-Thomson, Niloofar Ghadirian, Jennifer Collins, Elena Diez-Cecilia, Brendan Kelly, Hani Goodarzi, and Jiann Shiun Yuan. Artificial intelligence for covid-19 drug discovery and vaccine development. *Frontiers in Artificial Intelligence*, 3, 2020.

[49] Simon A. A. Kohl, Bernardino Romera-Paredes, Klaus H. Maier-Hein, Danilo Jimenez Rezende, S. M. Ali Eslami, Pushmeet Kohli, Andrew Zisserman, and Olaf Ronneberger. A hierarchical probabilistic u-net for modeling multi-scale ambiguities, 2019.

[50] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[51] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[52] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[53] E. M. Leitch, J. M. Kovac, N. W. Halverson, J. E. Carlstrom, C. Pryke, and M. W. E. Smith. Degree angular scale interferometer 3 year cosmic microwave background polarization results. *The Astrophysical Journal*, 624(1):10, may 2005.

[54] Pablo Lemos, Adam Coogan, Yashar Hezaveh, and Laurence Perreault-Levasseur. Sampling-based accuracy testing of posterior estimators for general inference, 2023.

[55] A LEWIS and A CHALLINOR. Weak gravitational lensing of the CMB. *Physics Reports*, 429(1):1–65, jun 2006.

[56] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets, 2018.

[57] Robert Link and Michael J. Pierce. Cosmological parameters from multiple-arc gravitational lensing systems. i. smooth lensing potentials. *The Astrophysical Journal*, 502(1):63–74, jul 1998.

[58] Hao Lv, Lei Shi, Joshua William Berkenpas, Fu-Ying Dao, Hasan Zulfiqar, Hui Ding, Yang Zhang, Liming Yang, and Renzhi Cao. Application of artificial intelligence and machine learning for COVID-19 drug discovery and vaccine design. *Briefings in Bioinformatics*, 22(6):bbab320, 08 2021.

[59] Rachel Mandelbaum, Barnaby Rowe, James Bosch, Chihway Chang, Frederic Courbin, Mandeep Gill, Mike Jarvis, Arun Kannawadi, Tomasz Kacprzak, Claire Lackner, Alexie Leauthaud, Hironao Miyatake, Reiko Nakajima, Jason Rhodes, Melanie Simet, Joe Zuntz, Bob Armstrong, Sarah Bridle, Jean Coupon, Jörg P. Dietrich, Marc Gentile, Catherine Heymans, Alden S. Jurling, Stephen M. Kent, David Kirkby, Daniel Margala, Richard Massey, Peter Melchior, John Peterson, Aaron Roodman, and Tim Schrabback. The Third Gravitational Lensing Accuracy Testing (GREAT3) Challenge Handbook. , 212(1):5, May 2014.

[60] Jean-Michel Marin, Pierre Pudlo, Christian P. Robert, and Robin Ryder. Approximate bayesian computational methods, 2011.

[61] J. C. Mather, E. S. Cheng, Jr. Eplee, R. E., R. B. Isaacman, S. S. Meyer, R. A. Shafer, R. Weiss, E. L. Wright, C. L. Bennett, N. W. Boggess, E. Dwek, S. Gulkis, M. G. Hauser, M. Janssen, T. Kelsall, P. M. Lubin, Jr. Moseley, S. H., T. L. Murdock, R. F. Silverberg, G. F. Smoot, and D. T. Wilkinson. A Preliminary Measurement of the Cosmic Microwave Background Spectrum by the Cosmic Background Explorer (COBE) Satellite. , 354:L37, May 1990.

[62] J P McKean, R Luichies, A Drabent, G Gürkan, P Hartley, A Lafontaine, I Prandoni, H J A Röttgering, T W Shimwell, H R Stacey, and C Tasse. Gravitational lensing in LoTSS DR2: extremely faint 144-MHz radio emission from two highly magnified quasars. *Monthly Notices of the Royal Astronomical Society: Letters*, 505(1):L36–L40, apr 2021.

[63] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 12 2004.

[64] J W Nightingale, Russell J Smith, Qiuhan He, Conor M O'Riordan, Jacob A Kegerreis, Aristeidis Amvrosiadis, Alastair C Edge, Amy Etherington, Richard G Hayes, Ash Kelly, John R Lucey, and Richard J Massey. Abell 1201: detection of an ultramassive black hole in a strong gravitational lens. *Monthly Notices of the Royal Astronomical Society*, 521(3):3298–3322, 03 2023.

[65] IV Ogburn, R. W., P. A. R. Ade, R. W. Aikin, M. Amiri, S. J. Benton, J. J. Bock, J. A. Bonetti, J. A. Brevik, B. Burger, C. D. Dowell, L. Duband, J. P. Filippini, S. R. Golwala, M. Halpern, M. Hasselfield, G. Hilton, V. V. Hristov, K. Irwin, J. P. Kaufman, B. G. Keating, J. M. Kovac, C. L. Kuo, A. E. Lange, E. M. Leitch, C. B. Netterfield, H. T. Nguyen, A. Orlando, C. L. Pryke, C. Reintsema, S. Richter, J. E. Ruhl, M. C. Runyan, C. D. Sheehy, Z. K. Staniszewski, S. A. Stokes, R. V. Sudiwala, G. P. Teply, J. E. Tolan, A. D. Turner, P. Wilson, and C. L. Wong. The BICEP2 CMB polarization experiment. In Wayne S. Holland and Jonas Zmuidzinas, editors, *Millimeter, Submillimeter, and Far-Infrared Detectors and Instrumentation for Astronomy V*, volume 7741 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, page 77411G, July 2010.

[66] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

[67] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[68] R. C. Reis, M. T. Reynolds, J. M. Miller, and D. J. Walton. Reflection from the strong gravity regime in a lensed quasar at redshift z = 0.658. *Nature*, 507(7491):207–209, mar 2014.

[69] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[70] Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[71] Kailash C. Sahu, Jay Anderson, Stefano Casertano, Howard E. Bond, Andrzej Udalski, Martin Dominik, Annalisa Calamida, Andrea Bellini, Thomas M. Brown, Marina Rejkuba, Varun Bajaj, Noé Kains, Henry C. Ferguson, Chris L. Fryer, Philip Yock, Przemek Mróz, Szymon Kozłowski, Paweł Pietrukowicz, Radek Poleski, Jan Skowron, Igor Soszyński, Michał K. Szymański, Krzysztof Ulaczyk, Łukasz Wyrzykowski, Richard K. Barry, David P. Bennett, Ian A. Bond, Yuki Hirao, Stela Ishitani Silva, Iona Kondo, Naoki Koshimoto, Clément Ranc, Nicholas J. Rattenbury, Takahiro Sumi, Daisuke Suzuki, Paul J. Tristram, Aikaterini Vandorou, Jean-Philippe Beaulieu, Jean-Baptiste Marquette, Andrew Cole, Pascal Fouqué, Kym Hill, Stefan Dieters, Christian Coutures, Dijana Dominis-Prester, Clara Bennett, Etienne Bachelet, John Menzies, Michael Albrow, Karen Pollard, Andrew Gould, Jennifer C. Yee, William Allen, Leonardo A. Almeida, Grant Christie, John Drummond, Avishay Gal-Yam, Evgeny Gorbikov, Francisco Jablonski, Chung-Uk Lee, Dan Maoz, Ilan Manulis, Jennie McCormick, Tim Natusch, Richard W. Pogge, Yossi Shvartzvald, Uffe G. Jørgensen, Khalid A. Alsubai, Michael I. Andersen, Valerio Bozza, Sebastiano Calchi Novati, Martin Burgdorf, Tobias C. Hinse, Markus Hundertmark, Tim-Oliver Husser, Eamonn Kerins, Penelope Longa-Peña, Luigi Mancini, Matthew Penny, Sohrab

Rahvar, Davide Ricci, Sedighe Sajadian, Jesper Skottfelt, Colin Snodgrass, John Southworth, Jeremy Tregloan-Reed, Joachim Wambsganss, Olivier Wertz, Yiannis Tsapras, Rachel A. Street, D. M. Bramich, Keith Horne, and Iain A. Steele. An isolated stellar-mass black hole detected through astrometric microlensing∗.$The Astrophysical Journal$, 933(1) : 83, jul2022.

[72] M. Sereno and G. Longo. Determining cosmological parameters from x-ray measurements of strong lensing clusters. *Monthly Notices of the Royal Astronomical Society*, 354(4):1255–1262, nov 2004.

[73] Lior Shamir. Automatic morphological classification of galaxy images. *Monthly Notices of the Royal Astronomical Society*, 399(3):1367–1372, nov 2009.

[74] S. A. Sisson, Y. Fan, and M. A. Beaumont. Overview of approximate bayesian computation, 2018.

[75] John Skilling. Nested Sampling. In Rainer Fischer, Roland Preuss, and Udo Von Toussaint, editors, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 735 of *American Institute of Physics Conference Series*, pages 395–405, November 2004.

[76] Martin A. Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

[77] Simon Tavaré, David J Balding, R C Griffiths, and Peter Donnelly. Inferring Coalescence Times From DNA Sequence Data. *Genetics*, 145(2):505–518, 02 1997.

[78] R. J. Thornton, P. A. R. Ade, S. Aiola, F. E. Angilè , M. Amiri, J. A. Beall, D. T. Becker, H-M. Cho, S. K. Choi, P. Corlies, K. P. Coughlin, R. Datta, M. J. Devlin, S. R. Dicker, R. Dünner, J. W. Fowler, A. E. Fox, P. A. Gallardo, J. Gao, E. Grace, M. Halpern, M. Hasselfield, S. W. Henderson, G. C. Hilton, A. D. Hincks, S. P. Ho, J. Hubmayr, K. D. Irwin, J. Klein, B. Koopman, Dale Li, T. Louis, M. Lungu, L. Maurin, J. McMahon, C. D. Munson, S. Naess, F. Nati, L. Newburgh, J. Nibarger, M. D. Niemack, P. Niraula, M. R. Nolta, L. A. Page, C. G. Pappas, A. Schillaci, B. L. Schmitt, N. Sehgal, J. L. Sievers, S. M. Simon, S. T. Staggs, C. Tucker, M. Uehara, J. van Lanen, J. T. Ward, and E. J. Wollack. THE ATACAMA COSMOLOGY TELESCOPE: THE POLARIZATION-SENSITIVE ACTPol INSTRUMENT. *The Astrophysical Journal Supplement Series*, 227(2):21, dec 2016.

[79] E. Vanzella, S. Cristiani, A. Fontana, M. Nonino, S. Arnouts, E. Giallongo, A. Grazian, G. Fasano, P. Popesso, P. Saracco, and S. Zaggia. Photometric redshifts with the multilayer perceptron neural network: Application to the HDF-s and SDSS. *Astronomy & Astrophysics*, 423(2):761–776, aug 2004.

[80] FH Wagner, A Sanchez, MPM Aidar, ALC Rochelle, Y Tarabalka, MG Fonseca, OL Phillips, E Gloor, and LEOC Aragão. Mapping atlantic rainforest degradation and regeneration history with indicator species using convolutional network. *PLOS ONE*, 15(2):e0229448–e0229448, February 2020. © 2020 Wagner et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

[81] Qi Wang, Yue Ma, Kun Zhao, and Yingjie Tian. A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, 9, 04 2022.

[82] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: icommon practices and emerging technologies/i. *IEEE Access*, 8:58443–58469, 2020.

[83] W. W. Zhu, A. Berndsen, E. C. Madsen, M. Tan, I. H. Stairs, A. Brazier, P. Lazarus, R. Lynch, P. Scholz, K. Stovall, S. M. Ransom, S. Banaszak, C. M. Biwer, S. Cohen, L. P. Dartez, J. Flanigan, G. Lunsford, J. G. Martinez, A. Mata, M. Rohr, A. Walker, B. Allen, N. D. R. Bhat, S. Bogdanov, F. Camilo, S. Chatterjee, J. M. Cordes, F. Crawford, J. S. Deneva, G. Desvignes, R. D. Ferdman, P. C. C. Freire, J. W. T. Hessels, F. A. Jenet, D. L. Kaplan, V. M. Kaspi, B. Knispel, K. J. Lee, J. van Leeuwen,

A. G. Lyne, M. A. McLaughlin, X. Siemens, L. G. Spitler, and A. Venkataraman. SEARCHING FOR PULSARS USING IMAGE PATTERN RECOGNITION. *The Astrophysical Journal*, 781(2):117, jan 2014.