A&A 675, A125 (2023)
https://doi.org/10.1051/0004-6361/202245126
© The Authors 2023

Astronomy & Astrophysics

Modeling lens potentials with continuous neural fields in galaxy-scale strong lenses

L. Biggio[1], G. Vernardos[2], A. Galan[2], A. Peel[2], and F. Courbin[2]

[1] Data Analytics Lab, Department of Computer Science, ETHZ, Universitätstrasse 6, 8006 Zürich, Switzerland
e-mail: `luca.biggio@inf.ethz.ch`
[2] Institute of Physics, Laboratory of Astrophysics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, 1290 Versoix, Switzerland

Received 4 October 2022 / Accepted 9 February 2023

ABSTRACT

Strong gravitational lensing is a unique observational tool for studying the dark and luminous mass distribution both within and between galaxies. Given the presence of substructures, current strong lensing observations demand more complex mass models than smooth analytical profiles, such as power-law ellipsoids. In this work, we introduce a continuous neural field to predict the lensing potential at any position throughout the image plane, allowing for a nearly model-independent description of the lensing mass. We applied our method to simulated *Hubble* Space Telescope imaging data containing different types of perturbations to a smooth mass distribution: a localized dark subhalo, a population of subhalos, and an external shear perturbation. Assuming knowledge of the source surface brightness, we used the continuous neural field to model either the perturbations alone or the full lensing potential. In both cases, the resulting model was able to fit the imaging data, and we were able to accurately recover the properties of both the smooth potential and the perturbations. Unlike many other deep-learning methods, ours explicitly retains lensing physics (i.e., the lens equation) and introduces high flexibility in the model only where required, namely, in the lens potential. Moreover, the neural network does not require pretraining on large sets of labeled data and predicts the potential from the single observed lensing image. Our model is implemented in the fully differentiable lens modeling code HERCULENS.

Key words. gravitational lensing: strong – galaxies: structure – methods: data analysis – gravitation – dark matter

1. Introduction

On large scales of the Universe, typically beyond 10 Mpc, the Lambda cold dark matter (ΛCDM) model, and in particular its dark matter (DM) component, successfully explains the formation of galaxies through the hierarchical gravitational collapse of matter (e.g., Toomre & Toomre 1972; Dubinski 1994; Springel et al. 2006). Below 10 Mpc, the gravitational interactions between cold DM and baryons perpetually reshape the mass distribution in between and within galaxies. Many complex mechanisms are thought to affect the mass content of galaxies, including gas outflows from active galactic nuclei (e.g., Zubovas & King 2012), supernova feedback (e.g., Scannapieco et al. 2008), and tidal stripping from ongoing mergers (e.g., Barnes & Hernquist 1996). As certain challenges remain unresolved within the cold DM paradigm (e.g., the “cusp-core problem”, Moore 1994; de Blok 2010; the “missing satellites problem”, Moore et al. 1999; Klypin et al. 1999), and the “too-big-to-fail problem” (Boylan-Kolchin et al. 2011; Papastergis et al. 2015), it is crucial to improve techniques that characterize the DM distribution in galaxies, and to compare it to theoretical predictions.

Strong gravitational lensing offers a direct probe of the total mass distribution within galaxies. In the galaxy-galaxy regime, this phenomenon arises when two galaxies located at different redshifts are aligned along our line of sight: the mass of the foreground lens galaxy causes deflection in the light emitted by the background source galaxy. From the observer’s point of view, multiple images of the source galaxy are visible and potentially form lensed arcs. Although magnified and highly distorted, these arcs encode significant information about the underlying mass distribution of the lens galaxy. The mass distribution is usually described through a lens potential for which simple functions, such as elliptical power-law profiles, often lead to a sufficiently good fit to the observation (e.g., Shajib et al. 2021). However, the true lens potential of galaxies is known to be more complex and can feature ellipticity twists or gradients, a bar component, frozen shocks, or include faint satellite galaxies. Moreover, populations of DM subhalos along the line of sight also contribute to the observed lensing distortions. Therefore, models that go beyond a simple smooth component are required to capture the complexity of the lens galaxy mass distribution, including its dark component.

Strong lens modeling requires an under-constrained problem to be solved, notably because the unlensed source light distribution is unknown and its multiple images only span a limited fraction of the lens extent. To address this difficulty, smooth analytical profiles have been widely used to regularize the model of the lens potential. If one observes clear signs of structures perturbing the smooth lens potential, such as luminous satellite galaxies or shear effects from a nearby galaxy cluster, it is possible to inject prior knowledge in the model by using well-motivated analytical profiles. However, in most cases, the signatures of complex mass distributions are less obvious, and can range from dark subhalos to high-order moments in the lens potential. Direct reconstruction methods are preferable in such cases, as they relax the strong assumptions on the underlying mass distribution that fully analytic strategies impose. An example is the gravitational imaging technique developed by

Open Access article, published by EDP Sciences, under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0),
which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
This article is published in open access under the Subscribe to Open model. Subscribe to A&A to support open access publication.

Koopmans (2005), which reconstructs both perturbations of the smooth lens potential and of the source light on pixelated grids, based on a first-order expansion of the lens equation. However, the success of the technique can strongly depend on the choice of regularization used to reconstruct the pixelated perturbations (Vernardos & Koopmans 2022). Recently, Galan et al. (2022) introduced a novel multiscale regularization technique based on wavelet transforms and sparsity priors to reconstruct different types of perturbations on a pixelated grid.

Aside from analytical and pixelated models, there has been growing interest in using deep learning to model the lens potential. Such approaches can offer much lower computation times compared to traditional methods, as well as the ability to learn complex features in the data. Most studies so far have focused on training deep convolutional neural networks (CNNs) to predict parameter values of usual analytical lens models (e.g., Hezaveh et al. 2017; Pearson et al. 2019), including the estimation of parameter uncertainties with approximate Bayesian neural networks (BNNs, e.g., Perreault Levasseur et al. 2017; Wagner-Carena et al. 2021). The generation of realistic training sets is very challenging, but progress has been made recently with the use of real images of source and lens galaxies (Schuldt et al. 2021). A common feature of these approaches is that they do not explicitly rely on the well-understood physics of gravitational lensing, namely the lens equation. Instead, they implicitly learn a representation of lensed features through hundreds of thousands of example realizations from a training set.

Deep neural networks have also been used for the detection of DM subhalos (Alexander et al. 2020b), the recovery of their angular positions (Coogan et al. 2020; Yao-Yu Lin et al. 2020; Ostdiek et al. 2022), and distinguishing among different types of DM (e.g., Alexander et al. 2020a). Since considering the effect of populations of subhalos allows for lower subhalo masses to be probed (Hezaveh et al. 2016), similar techniques have been used to frame the weighing of subhalos as a classification problem (Varma et al. 2020), to constrain the power spectrum of lens potential perturbations with statistical uncertainties (Vernardos et al. 2020), and to constrain the DM subhalo mass function (Brehmer et al. 2019; Diaz Rivero & Dvorkin 2020; Wagner-Carena et al. 2023). While these deep-learning approaches probe many facets of the lens complexity problem, they remain limited by the need for large and realistic training sets (in particular for CNNs), which can be challenging to design and generate. Furthermore, as deep-learning-based techniques have not been applied to real strong lensing data yet, it is still unclear whether assumptions as to the training sets might limit their application to more complex systems.

Recently, a novel class of approaches has been explored by several authors that combine deep learning with physics-based modeling. The goal of such hybrid approaches is to retain known physics explicitly while still achieving high model complexity and low computation time. This new paradigm has resulted in a number of successful methods across a wide range of fields, such as robotics (Coros et al. 2021), quantum mechanics (Hermann et al. 2020), molecular dynamics (Doerr et al. 2021), and fluid dynamics (Thuerey et al. 2021). Similar in spirit is the concept of physics-informed neural networks (Raissi et al. 2019; Cuomo et al. 2022), where machine learning is used to fit the data while satisfying constraints imposed by complex partial differential equations from physics.

In the context of strong lensing, Chianese et al. (2020) introduced a modeling method where the source light is generated on a pixelated grid by a pretrained variational auto-encoder (VAE) and then combined through the lens equation with an analytic description of the lens potential to produce an image of the lensed source. This is made possible using differentiable programming, where all model components, namely the VAE (by construction), the analytic profiles, and the lens equation, are written as automatically differentiable functions, allowing for the efficient optimization of parameters via gradient descent. In a similar spirit, Adam et al. (2022) trained neural networks on real galaxy images and cosmological simulations and used them as custom priors to reconstruct the source light and the lens potential, respectively. Differentiable programming was also used in Karchev et al. (2022) to model lensed sources with Gaussian processes and variational inference, and Gu et al. (2022) constructed a GPU-accelerated Bayesian framework for fast modeling and parameter inference applied to large samples of lenses.

Another recent effort in fully differentiable strong lens modeling is HERCULENS[1] (Galan et al. 2022), which combines analytical and pixelated models together with deep-learning optimization methods into a single modular framework. HERCULENS allows for fast gradient-informed optimization over thousands of parameters, in particular the reconstruction of pixelated source light distribution and lens potential perturbations regularized with wavelets. It is based on JAX (Bradbury et al. 2018), a recently introduced Python library designed for performing high-performance automatic differentiation tasks.

Following up the proof-of-concept work of Biggio et al. (2021), we introduced a new fully differentiable method to model the lens potential using a deep neural network that leverages the auto-differentiable capabilities of HERCULENS. We note that recently, a similar technique has been used to reconstruct the lensed source light distribution (Mishra-Sharma & Yang 2022).

We tested our method in various realistic contexts, ranging from a fully smooth lens potential to perturbations on top of a smooth profile caused by a localized DM subhalo, populations of DM subhalos, or higher-order moments within the lens mass distribution. Despite the large number of parameters, automatic differentiation enables efficient (and exact) loss function gradient and higher-order derivative computations, facilitating rapid training and convergence to the best-fit solution. We emphasize that training in our case was performed with only the target lensing image, avoiding the typical requirement of large sets of labeled examples. Furthermore, our neural network implementation of the potential allows spatial derivatives to be computed automatically, meaning deflection angles and other lensing quantities (e.g., convergence and shear) can be obtained with high accuracy.

The paper is organized as follows. We first discuss different image representation methods in Sect. 2 to introduce the continuous neural field we consider in this work. We then present our methodology in Sect. 3, which includes the elements of the strong lensing formalism required for our work. We also describe the framework of implicit representation learning as applied to modeling the lensing potential, along with details of the network architecture and training. In Sect. 4, we describe the experiments carried out on mock data to test our method. We present our experimental analysis in Sect. 5 and assess the performance of the proposed approach. We conclude and discuss future extensions and research directions in Sect. 6.

2. Implicit image representation

In this work, we propose a strategy to represent, in a continuous way, the gravitational potential of a lens at any point within the

---

[1] `https://github.com/austinpeel/herculens`

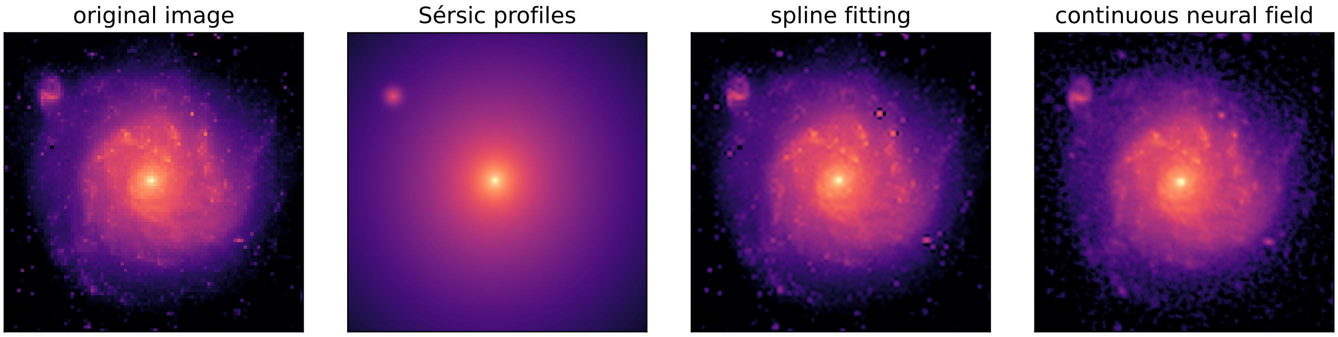

Fig. 1. Examples of image representation methods, including the one used in this work. The left panel shows an HST observation of NGC 1309 featuring complex features such as spiral arms and a companion galaxy. The remaining panels show three representations of the original image, from the smoothest to the most complex, obtained by fitting the input image with an analytical model. The resolution of the models is in practice infinite, but here we simply choose one higher than the resolution of the original image. The second and third panels show a representation with two Sérsic profiles and a spline fit respectively. The fourth panel is the representation of the original image using a continuous neural field. This is obtained after optimizing (or 'training', but not in the conventional machine learning meaning) the weights of the neural network based on the pixel 2D coordinates and values of the input image. In this work, we are interested in spatial derivatives of the image (see Sect. 3.1), which we can compute analytically and efficiently from the neural field representation.

field of view. Since the lens potential corresponds to mass distributions that can span a wide range of spatial scales, our method must be able to represent potentially complex fields. We must also be able to compute derivatives of our model because the spatial derivative of the lens potential is equal to the deflection angle (introduced in the next section), which is the main observable quantity. To this end, we discuss here different methods to represent two-dimensional signals (i.e., images) and to predict their values and derivatives continuously, which motivates our approach with continuous neural fields.

While we are ultimately interested in modeling the gravitational potential of a galaxy, we first illustrate image representation techniques with an example image of the light distribution of a galaxy. The first panel of Fig. 1 shows an example of a relatively complex galaxy that contains features at different spatial scales, such as spiral arms, star-forming regions, and a satellite galaxy. The other panels show three different models representing the galaxy's light distribution continuously throughout the field of view, ordered in increasing complexity.

The first model is analytic and is composed of one elliptical and one circular Sérsic profile (14 total parameters). This representation only captures large-scale features of the main galaxy and its satellite. All Sérsic parameters have been optimized using gradient descent based on a χ^2 likelihood[2]. The second model is a bivariate spline representation of degree 3 with fixed knots (8100 parameters) and whose coefficients are obtained through linear inversion[3]. Compared to the Sérsic model, the spline functions are able to capture many more features including the spiral arms. However, due to these functions being polynomials, this representation is not accurate for sharp, small-scale variations.

In this work, we go beyond these standard methods and consider a continuous neural field model, inspired by the growing literature on implicit representation learning (Mildenhall et al. 2020; Sitzmann et al. 2020; Tancik et al. 2020). The goal of implicit representation is to parameterize a generic signal in a continuous and differentiable fashion by means of a neural network. As an example, we show on the rightmost panel of Fig. 1 the capacity of a simple fully connected network ($\sim 10^5$ parameters) to accurately reproduce the details of the original input image of a galaxy. To obtain such a representation, we trained the network – that is, we optimized all of the network's parameters – to map pixel positions into pixel values by minimizing the squared error between the network output and the input image. Since such a neural network is fully differentiable by construction, we can employ gradient descent techniques (ADAM in this case, Kingma & Ba 2014) to efficiently optimize a large number of parameters simultaneously.

The network we introduce to model the lens potential operates similarly, but with the difference that the true target potential is unknown. It acts as a modular building block within the HERCULENS lens model and interacts with the other model components driven by the physics of strong lensing, which we describe in Sect. 3. In other words, our model is just an "image" of the lens potential that we then use in the lens equation (with the bonus of getting its derivatives automatically, i.e., the deflection angles) in the traditional way. In contrast to pixelated methods, our network is able to model the lens potential continuously at any position in the observed field of view and can equally be used to represent perturbations of a smooth profile or the full potential itself.

3. Methodology

We propose to use a continuous neural field to predict the lens potential based on an image of a strong lens. Below we introduce the necessary formalism of strong gravitational lensing and justify the need for accurate computation of spatial derivatives, which have a very specific meaning in lensing physics.

3.1. Strong gravitational lensing

Apparent distortions of a distant luminous source caused by the gravitational lensing of a foreground mass distribution are captured by the lens equation:

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\alpha}(\boldsymbol{\theta}). \tag{1}$$

It is a mapping from lens plane coordinates $\boldsymbol{\theta}$ to (unlensed) source plane coordinates $\boldsymbol{\beta}$, both two-dimensional angular position vectors on the sky. The reduced deflection angle $\boldsymbol{\alpha}$ is in general a nonlinear function of the lens plane position and can be derived from the projected lens mass density, namely

$$\boldsymbol{\alpha}(\theta) = \nabla\psi(\boldsymbol{\theta}), \tag{2}$$

---

[2] Using HERCULENS.

[3] Using SCIPY's routine `RectBivariateSpline`.

where ψ is the lens potential in the limit of the thin lens approximation. Equation (2) implies that, in order to reconstruct the underlying lens potential from the data, we first need to model deflection angles, which are first-order derivatives of the potential.

As gravitational lensing preserves surface brightness, a source plane image described by $\mathcal{I}(\boldsymbol{\beta})$, for example the unlensed image of a galaxy, is mapped to the lens plane through

$$\mathcal{I}(\boldsymbol{\theta}) = \mathcal{I}\Big(\boldsymbol{\beta}(\boldsymbol{\theta})\Big). \tag{3}$$

This relation is continuous and holds for any light ray emanating from the source and arriving at the observer. In practice, however, our telescopes capture only discretized images of the sky and are subject to instrumental noise. We can thus postulate the problem in a discretized way so that the data (i.e., image) we observe, $\boldsymbol{d}$, is given by

$$\boldsymbol{d} = \mathbf{B}\mathbf{L}_{\psi}\boldsymbol{s} + \mathbf{B}\boldsymbol{\ell} + \boldsymbol{n}, \tag{4}$$

where $\boldsymbol{s}$ is the source light profile, $\mathbf{L}_{\psi}$ is a discretized version of Eq. (1) encoding the lens mapping, and $\boldsymbol{\ell}$ is the light profile of the lens. We use the subscript ψ on $\mathbf{L}_{\psi}$ to emphasize that the precise form the lensing operator takes depends (nonlinearly) on the lens potential. $\mathbf{B}$ is a blurring operator modeling the point spread function (PSF) of the instrument and representing the seeing conditions. The final term, $\boldsymbol{n}$, is additive noise, typically a combination of instrumental read-out noise and signal-dependent shot noise.

Our goal in this work is to recover the lens potential ψ and source light $\boldsymbol{s}$ in Eq. (4) given an observed image of strong lensing $\boldsymbol{d}$. We assume throughout that the PSF is known and constant across the field of observation. As it is not our focus here, we also assume that the lens light $\boldsymbol{\ell}$ has been accurately modeled and subtracted from the data in each case, although our modeling framework enables it to be straightforwardly included. Blending of the lens light with lensed source features can make disentangling the two difficult as well as introduce degeneracies. We therefore leave a full treatment of Eq. (4) including lens light for future work.

Smooth lens mass distributions are often described by analytic profiles in terms of (dimensionless) convergence κ, where

$$\kappa(\boldsymbol{\theta}) = \frac{1}{2}\nabla^2\psi(\boldsymbol{\theta}). \tag{5}$$

For example, the singular isothermal ellipsoid (SIE) is commonly used to model the large-scale features of the lensing mass and is defined as a function $\kappa(\boldsymbol{\theta})$. The profile has analytic formulas for the corresponding potential ψ and deflection angle $\boldsymbol{\alpha}$ (see, for example, Appendix B of Galan et al. 2022). Small-amplitude deviations from smoothness arising from various physical mechanisms (see Sect. 4) can be included as an additional analytic or pixelated component in the potential. In such cases, we can distinguish the smooth analytic part $\psi_{\rm sm}$ from the perturbations $\psi_{\rm pert}$ and write

$$\psi = \psi_{\rm sm} + \psi_{\rm pert}. \tag{6}$$

The neural network model for ψ that we explore in this work is general and can used to represent both perturbations on top of a smooth potential, namely $\psi_{\rm pert}$ alone, as well as the full potential ψ itself.

3.2. Neural networks for implicit representation

Given an input vector $\boldsymbol{x} \in \mathbb{R}^d$, the action of the simplest deep neural network model, the multilayer perceptron (MLP), is the sequential application of affine transformations $T(\boldsymbol{x})$, defined by weight matrices $\mathbf{W}$ and biases $\boldsymbol{b}$, and an element-wise nonlinear activation function ϕ:

$$\Phi(\boldsymbol{x}) = \left(T^{(n_l)} \circ \phi \circ T^{(n_l-1)} \circ \ldots \circ \phi \circ T^{(1)}\right)(\boldsymbol{x}), \tag{7}$$

where

$$T^{(i)}(\boldsymbol{x}) = \mathbf{W}^{(i)}\boldsymbol{x} + \boldsymbol{b}^{(i)} \tag{8}$$

is the ith transformation and n_l is the number of layers. In the case of representing image values based on input coordinates, $d = 2$ and the final transformation $T^{(n_l)}$ maps the output of the penultimate layer to $\mathbb{R}$.

Much of the interest in deep neural networks stems from the fact that these models are known to be universal function approximators. In other words, given a continuous function f on a compact set of d-dimensional space $f : \mathbb{R}^d \to \mathbb{R}^k$, there exists a neural network Φ with (at least) one hidden layer and a nonlinear activation function that approximates it to any desired degree of precision (Cybenko 1989). It is thanks to this property and to the development of powerful nonconvex optimization algorithms (Kingma & Ba 2014) that neural networks have been extensively applied in a variety of fields where datasets in the form of input-output pairs are available.

Recent work has demonstrated that neural networks can be employed to represent complex signals including time series, images, and 3D scenes with a high level of detail (Mildenhall et al. 2020; Sitzmann et al. 2020; Tancik et al. 2020). In particular, standard MLPs can be trained to map a generic input coordinate $\boldsymbol{x}$ to the corresponding signal value at that coordinate. In our present application to strong lensing, this means pixel values of an observed image (see Fig. 1). Given that such a neural network is defined over a continuous domain, the ability of the resulting model to capture fine details of an image is not limited by any predefined grid resolution. Its accuracy depends only on the capacity and expressive power of the network architecture.

In the most general case, the target signal we aim to represent with a network is not directly available, but instead only implicitly defined. For instance, we might only have access to the gradients or higher-order derivatives of the image we wish to recover. As a result, following Sitzmann et al. (2020), we can frame the problem of implicit function learning as the search for $\Phi(\boldsymbol{x})$ subject to

$$C_m\Big(\boldsymbol{a}(\boldsymbol{x}), \Phi(\boldsymbol{x}), \nabla\Phi(\boldsymbol{x}), \ldots\Big) = 0, \tag{9}$$

where $\{C_m\}_{m=1}^{M}$ are a set of M constraints that each relates the function Φ and/or its derivatives to a set of known quantities $\boldsymbol{a}(\boldsymbol{x})$. Each constraint, depending on its desired effect, applies to its own subset of the input domain Ω_m such that $C_m = 0$ holds for all $\boldsymbol{x} \in \Omega_m$. Implicit function learning can therefore be translated into the minimization of a loss function penalizing deviations of the collection of constraints from zero:

$$L = \int_{\Omega} \sum_{m=1}^{M} \mathbf{1}_{\Omega_m}(\boldsymbol{x}) \left\| C_m\left(\boldsymbol{a}(\boldsymbol{x}), \Phi(\boldsymbol{x}), \nabla\Phi(\boldsymbol{x}), \ldots\right) \right\| \mathrm{d}\boldsymbol{x}, \tag{10}$$

for a suitably chosen norm $\|\cdot\|$. The indicator function $\mathbf{1}_{\Omega_m}(\boldsymbol{x})$ restricts each constraint only to its domain, giving 1 when $\boldsymbol{x} \in \Omega_m$ and 0 otherwise.

In the simple case of fitting an image $\mathcal{I}(\boldsymbol{x})$ directly from its pixel values, a single constraint is present, namely $C\big(\mathcal{I}(\boldsymbol{x}), \Phi(\boldsymbol{x})\big) = \mathcal{I}(\boldsymbol{x}) - \Phi(\boldsymbol{x})$, where $\boldsymbol{a}(\boldsymbol{x}) = \mathcal{I}(\boldsymbol{x})$. The loss in this case assumes the form

$$L = \frac{1}{N} \sum_{i=1}^{N} \left[\mathcal{I}(\boldsymbol{x}_i) - \Phi(\boldsymbol{x}_i)\right]^2, \tag{11}$$

where N is the number of pixels, and we have chosen, for example, the mean squared error as the norm, as was used to produce the right panel of Fig. 1. In more complex scenarios, the number and types of constraints depend on the details of the problem and on the amount of information available to train the model. In this work, these constraints arise directly from strong lensing physics, namely, the lens equation and its relation to the lens potential (see Sect. 3.1).

A limitation of standard MLPs has been pointed out by Rahaman et al. (2018), where it was shown that these models are biased toward low-frequency solutions. Where reconstructing the fine-grained details of an image is important, such as in representing realistic and complex natural images, this property can hinder the applicability of MLP networks. A number of solutions have been proposed to overcome this issue (Sitzmann et al. 2020; Tancik et al. 2020). In this work, we implement the approach of Tancik et al. (2020), in which the input coordinate vector is first passed through a Fourier feature mapping γ_σ before being fed into the network:

$$\gamma_\sigma(\boldsymbol{x}) = \left[\, \cos(2\pi \mathbf{A}\boldsymbol{x}),\ \sin(2\pi \mathbf{A}\boldsymbol{x}) \,\right]^\top. \tag{12}$$

Each entry in $\mathbf{A} \in \mathbb{R}^{n\times d}$ is sampled from $\mathcal{N}\left(0, \sigma^2\right)$, and the cos and sin functions act element-wise on their arguments. The input $\boldsymbol{x}$ is therefore mapped to $2n$ components via γ_σ, and n is typically chosen to be larger than d. This straightforward operation has proven to effectively limit the spectral bias of standard neural networks and enable more control over the output frequencies. Applying γ_σ can thus be thought of as a preprocessing step whereby Eq. (7) becomes $\Phi(\boldsymbol{x}) \to \Phi(\gamma_\sigma(\boldsymbol{x}))$.

3.3. Hybrid differentiable lensing model with a neural network potential

We model the lensing potential $\psi(\boldsymbol{\theta})$ with a neural network $\Phi(\boldsymbol{x})$ as defined by Eq. (7) and including the γ_σ mapping. It is parameterized by a collection of weights and biases, which we denote by $\boldsymbol{\eta}_\Phi$, and $\boldsymbol{x} \in \mathbb{R}^2$ is a point in the (continuous) lens plane. The network has been implemented as a module within the HERCULENS code (Galan et al. 2022) using Flax (Heek et al. 2020), a JAX-compatible library for neural network development. Since the model is fully differentiable, training the network (i.e., optimizing $\boldsymbol{\eta}_\Phi$) can be carried out using gradient descent optimization. It can also be done simultaneously with the optimization of other model components, such as analytic smooth lens parameters or source light parameters.

Supervised learning for deep neural networks requires a large number of labeled input-output pairs. In strong lensing applications, this typically means training on tens or hundreds of thousands of example images (outputs) generated from known input lensing parameters (inputs). In our case, the training set is instead comprised of the set of all pixel positions and their corresponding brightness values as provided in the original data image. No other prior training set needs to be created for our hybrid model to learn its implicit representation of ψ. We note, though, that the learned potential is specific to the data being modeled and does not necessarily generalize to other observations.

Given the observed image data $\boldsymbol{d}$ of a strongly lensed system, Eq. (9) can be written as the single constraint

$$C(\boldsymbol{d}, \nabla\Phi) = \boldsymbol{d} - \mathsf{BL}_\psi(\nabla\Phi)\boldsymbol{s}, \tag{13}$$

where, as described in Sect. 3.1, we assume that the lens light $\boldsymbol{\ell}$ has been accurately modeled beforehand and subtracted from the data. Because the lens equation maps image to source plane coordinates via the gradient of the lens potential, we write the lensing operator L_ψ here as an explicit function of $\nabla\Phi$.

Combining Eqs. (10) and (13) while taking into account noise gives rise to the loss function

$$L(\boldsymbol{\eta}) = \frac{1}{2}(\boldsymbol{d} - \boldsymbol{m})^\top \mathbf{C}_{\boldsymbol{d}}^{-1}(\boldsymbol{d} - \boldsymbol{m}), \tag{14}$$

where $\mathbf{C}_{\boldsymbol{d}}$ is the covariance matrix of the data, and we have defined the model

$$\boldsymbol{m}(\boldsymbol{\eta}) \equiv \mathsf{BL}_\psi(\boldsymbol{\eta}_\Phi, \boldsymbol{\eta}_\psi)\boldsymbol{s}(\boldsymbol{\eta}_s). \tag{15}$$

Equation (15) is the general form of our model, where $\boldsymbol{\eta}_s$ are the source light parameters (whether the source is analytic or pixelated), and $\boldsymbol{\eta}$ refers to the collection of all model parameters. We distinguish $\boldsymbol{\eta}_\psi$ from $\boldsymbol{\eta}_\Phi$, where the former refers to parameters of an analytic smooth potential, which can be included when Φ is only modeling perturbations. Without loss of generality, we assume the noise to be uncorrelated, which makes $\mathbf{C}_{\boldsymbol{d}}$ diagonal. Equation (14) can be understood as the negative log-likelihood of normally distributed data, and is similar to Eq. (11) but with the mean squared error replaced by the χ^2.

A schematic illustration of our modeling approach is shown in Fig. 2. The resulting framework consists of a hybrid model, where the neural network lens potential operates seamlessly within a differentiable strong lensing simulator. Our main goal is to leverage a general-purpose neural network to accurately model the potential without any specific inductive biases on its form and with the only constraints stemming from the physics of strong lensing and the observed data.

The use of implicit neural networks to model ψ also presents some immediate advantages with respect to pixelated models of the lensing potential: once trained, the network can be queried at any continuous input point without the need to interpolate between predefined grid positions. We note as well that owing to the differentiable nature of Φ, the spatial gradients of the modeled potential can be efficiently computed via autodifferentiation with respect to the input $\boldsymbol{x}$. We therefore have direct access to the map of deflection angles $\boldsymbol{\alpha}(\boldsymbol{\theta})$, which can be used as an additional test of the method's accuracy.

3.4. Network architecture and parameter optimization

Throughout our study, we keep the complexity of the architecture relatively low, using five hidden layers of 100 nodes each to specify $\Phi(\boldsymbol{x})$. The activation ϕ was chosen to be the Swish function (Ramachandran et al. 2017), which we found to outperform the common rectified linear unit (ReLU) in terms of the quality and smoothness of the reconstructed potential gradients (i.e., deflection angles).

We chose $N = 100$ for the dimension of the Fourier embedding and a standard deviation $\sigma = 0.05$ (see Eq. (12)). We recall that this mapping is not a trained parameter and serves only to

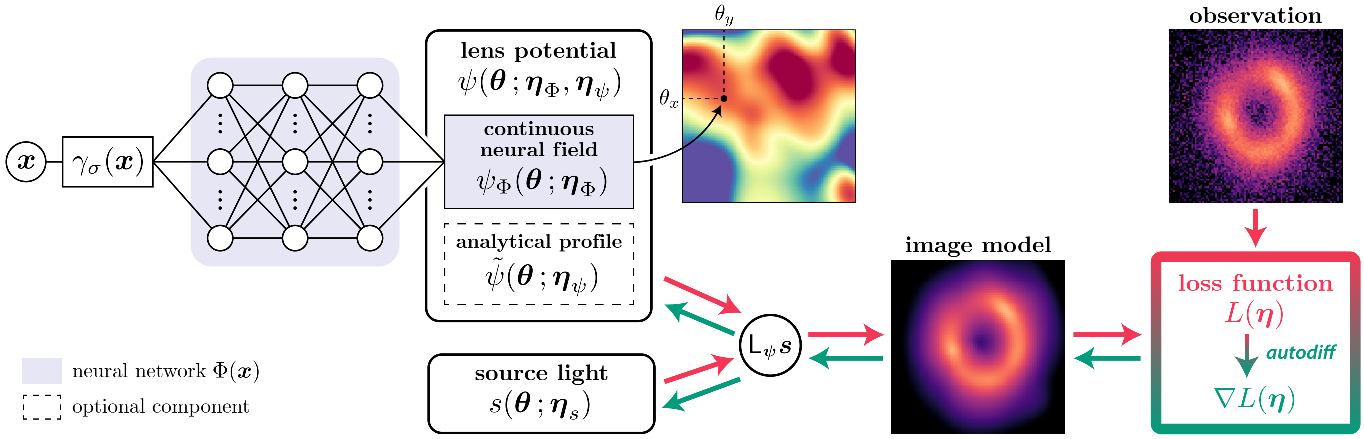
$\boldsymbol{x}$
$\gamma_\sigma(\boldsymbol{x})$
lens potential
$\psi(\boldsymbol{\theta}\,;\boldsymbol{\eta}_\Phi,\boldsymbol{\eta}_\psi)$
continuous neural field
$\psi_\Phi(\boldsymbol{\theta}\,;\boldsymbol{\eta}_\Phi)$
analytical profile
$\tilde{\psi}(\boldsymbol{\theta}\,;\boldsymbol{\eta}_\psi)$
θ_y
θ_x
observation
image model
loss function
$L(\boldsymbol{\eta})$
autodiff
$\nabla L(\boldsymbol{\eta})$
$\mathsf{L}_\psi s$
source light
$s(\boldsymbol{\theta}\,;\boldsymbol{\eta}_s)$
neural network $\Phi(\boldsymbol{x})$
optional component

Fig. 2. Flowchart of the proposed method. The lens potential is modeled with a continuous neural field that takes as input any position in the image plane and outputs the value of the lens potential at that position. Alternatively, an analytical profile (e.g., a SIE) models the smooth component of the lens potential, while the neural network captures deviations from that smooth component. The input coordinates are first passed through a Fourier feature mapping (γ_σ) to increase the dynamic range of the recovered features. The different model components follow Eq. (4) except for the blurring operator, omitted to avoid clutter. Since the model is fully differentiable, automatic differentiation is used to compute the exact gradient of the highly nonlinear loss function.

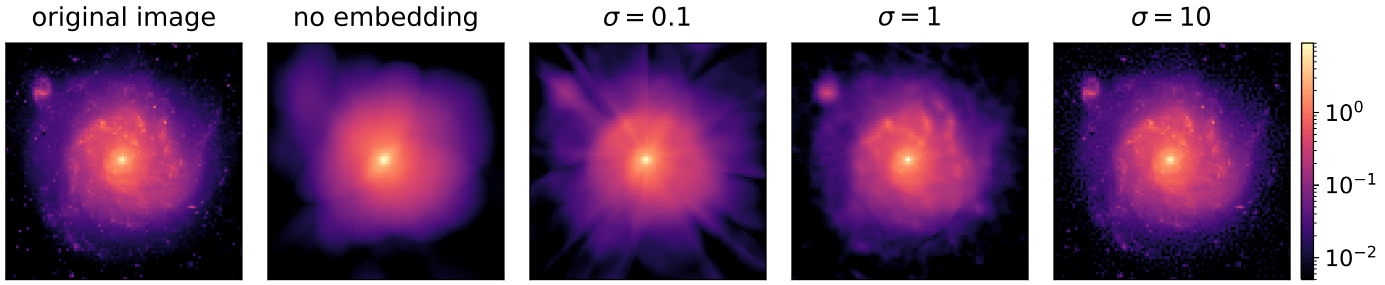

Fig. 3. Illustration of the effect of scale σ on fitting an image using a neural network including the Fourier feature mapping γ_σ. The original image is shown on the far left, and the remaining images show fully connected network representations with varying σ values for the embedding. Lower values (including no embedding) give smoother representations but fail to capture high spatial frequencies in the image. Higher values are able to capture the fine details but will also overfit any high-frequency noise present. The same network architecture and training scheme were used to produce all of the fits.

focus the network toward a certain range of spatial frequencies. The value we chose was to prevent the network from fitting the noise and provided good results across all of the experiments carried out. Apart from the choice of activation function, we did not fine-tune all hyperparameters, a procedure that would most likely improve results, but which is beyond our present scope.

We illustrate the effect of varying σ on fitting an image of a strongly lensed quasar system in Fig. 3. The same data and network architecture as were used in Fig. 1 were used here, namely a simple MLP fitting of a galaxy image. Low values of σ (as well as no embedding at all) lead the network to only capture the low spatial frequencies of the image, meaning the resulting fit is smoother than the original while noise is effectively filtered out. This is in agreement with the recent finding of Rahaman et al. (2018) that neural networks are intrinsically biased to learn low-frequency functions. On the other hand, higher σ values can result in overfitting the data, so that Φ learns to reproduce the noise in addition to the underlying signal. This example shows how σ acts as a tunable knob for the network allowing control over the desired output level of detail. We underline again that our method is adapted ("trained") to each given lens system. Hence, we cannot really generalize it and that applies to the value of sigma as well. In Fig. 3, we simply represent an image of a galaxy with our neural field and good values of σ are between 1 and 10. In the rest of the paper, we represent the lens potential so a completely different value ($\sigma = 0.05$) happens to work best (which also depends on the chosen source).

In Sect. 5, we perform experiments where we model different combinations of lens potential perturbations and the full potential directly. In all the experiments, we train our network using the ADAM optimizer (Kingma & Ba 2014) with learning rate 10^{-3} as implemented in the Optax framework (Babuschkin et al. 2020), a gradient processing and optimization library for JAX.

As an illustration of the typical training procedure of Φ, we consider the concrete example of modeling a single localized subhalo perturbation to a smooth potential. Progress of Φ toward the true $\psi_{\rm pert}$ as a function of epoch is shown in Fig. 4, with the only training signal coming from the loss in Eq. (14). At the start of training, the network is randomly initialized, and thus its output is far from the ground truth (e.g., at epoch 2). As training progresses, the model finds the right location and scale of the perturbation, obtaining a good fit after only around 300 epochs, as shown in Fig. 5.

4. Experimental setup

Any lens modeling effort consists of three main components: the brightness distributions of the source and lens, and the mass distribution of the lens that is linked to the lens potential. None of these components is precisely known, and analyzing any

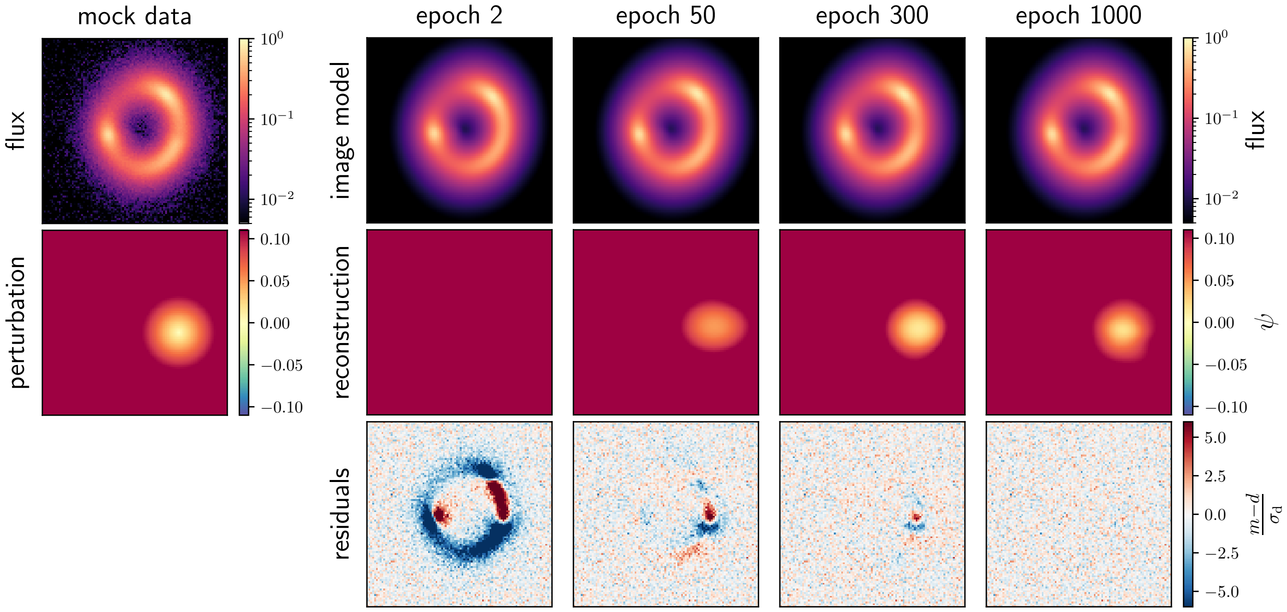

Fig. 4. Evolution of the network output prediction when modeling an example localized perturbation to a smooth lens potential. Mock noisy lensing image data and the true perturbation are shown on the left. The predicted lensing image, the network reconstruction of ψ_{pert}, and residuals between the data and the predicted image at different training epochs are also shown. Training is performed via back-propagation of the gradients through the simulator, as presented in Fig. 2.

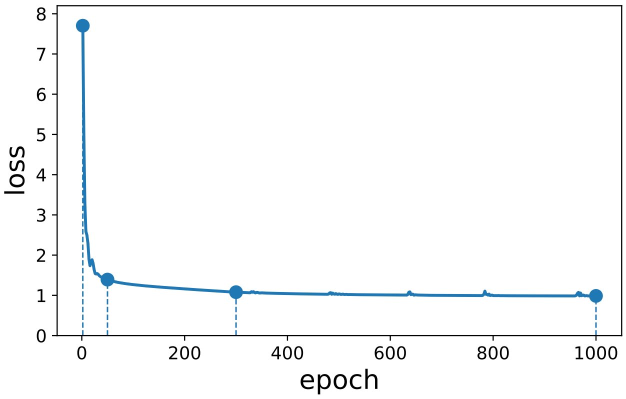

Fig. 5. Evolution of the loss function during the training of ψ_{pert} in the example of Fig. 4. Epochs for which the image model, the reconstructed perturbation, and image plane residuals are shown in Fig. 4 are indicated on the plot. By epoch 300, the network already produces a good fit to the data with a small loss value and residuals largely consistent with noise.

lensing system relies on priors, the extent to which depends on the method used. In addition, all components can have detailed structure on small scales that cannot be fully captured by the typically used macroscopic profiles, like SIE and Sérsic (Sérsic 1963). For example, there is a known but difficult-to-capture degeneracy between source brightness and lens potential that can complicate any attempt to model the system down to the noise (e.g., see Vernardos & Koopmans 2022).

In this work, we explore how model-agnostic neural networks can provide a reconstruction of the lens potential across all scales with only little prior information. Therefore, in the mock data used in our experiments, we assume a smooth, analytic, elliptical Sérsic brightness profile for the source light, an SIE profile as the main lens potential, and no light from the lens (the choices of parameters for the analytic profiles used can be found in Tables 1 and 2 and in Sect. 4.4). We then perturb the main lens potential in different, physically motivated ways, which we explain in detail below. We note that the experimental setup is very similar to Galan et al. (2022), and we therefore focus on emphasizing the differences here.

4.1. Single localized subhalo

We simulate a single, localized, spherical DM subhalo, assumed to have an isothermal profile with Einstein radius $\theta_{\text{E,sub}} = 0.07$ arcsec (for reference, the main deflector has $\theta_{\text{E}} = 1.8$ arcsec). Assuming typical redshifts of Early-type, Early-type lensing systems (Oldham & Auger 2018) for the lens and the source, 0.3 and 0.7 respectively, and that the perturbing subhalo is in the plane of the lens, the mass within $\theta_{\text{E,sub}}$ is $10^9\ M_\odot$. This is comparable to massive subhalos detected via their lensing effect that have been previously reported in the literature (e.g. Vegetti et al. 2010). The mean perturbation level computed within the region containing the lensed arcs corresponds to 6.1% of the main lens potential. The resulting perturbing lens potential is shown in Fig. 6.

4.2. Population of subhalos

The net effect on the lens potential of a population of DM subhalos along the line of sight can be approximated by a Gaussian Random Field (GRF, Chatterjee & Koopmans 2018; Bayer et al. 2018; Vernardos et al. 2020; Vernardos & Koopmans 2022). The statistical properties of such a random field can be defined by its (isotropic) power spectrum, which follows a power law:

$$\text{PS}(k) = A_{\text{GRF}}\, k^{-\beta_{\text{GRF}}}, \tag{16}$$

where k is the wave number, β_{GRF} is the power-law slope, and A_{GRF} is a normalizing factor that depends on β_{GRF} and the size of the field of view and is related to the variance of the GRF, σ^2_{GRF} (for the exact formula, see Chatterjee 2019). The value of β_{GRF} determines the distribution of power at each length scale: a large value leads to extended and smooth variations, whereas a small value creates a large number of localized and grainy structures.

Table 1. Evaluation of the reconstructed perturbing potentials (Sect. 5) by fitting analytical profiles to the SIS, Shear, and Combo models, and calculating the power spectrum of the GRF model.

	Name	Units	Truth	ψ_{pert} only	$\psi_{\mathrm{sm}} + \psi_{\mathrm{pert}}$
SIS	$x_{0,\mathrm{sub}}$	arcsec	1.9	$1.940^{+0.270}_{-0.249}$	$1.950^{+0.284}_{-0.252}$
	$y_{0,\mathrm{sub}}$	arcsec	−0.4	$-0.411^{+0.218}_{-0.221}$	$-0.395^{+0.194}_{-0.229}$
	$\theta_{\mathrm{E,sub}}$	arcsec	0.07	$0.069^{+0.004}_{-0.004}$	$0.069^{+0.003}_{-0.004}$
Shear	$x_{0,\mathrm{ext}}$	arcsec	0.0	$0.006^{+0.134}_{-0.138}$	$-0.001^{+0.130}_{-0.141}$
	$y_{0,\mathrm{ext}}$	arcsec	0.0	$0.013^{+0.136}_{-0.135}$	$0.004^{+0.128}_{-0.145}$
	γ_{ext}	–	0.032	$0.032^{+0.002}_{-0.002}$	$0.032^{+0.002}_{-0.002}$
	ϕ_{ext}	°	144.0	$144.126^{+2.258}_{-2.793}$	$143.913^{+2.494}_{-2.779}$
Combo	$x_{0,\mathrm{ext}}$	arcsec	0.0	$0.085^{+0.507}_{-0.458}$	$0.078^{+0.443}_{-0.391}$
	$y_{0,\mathrm{ext}}$	arcsec	0.0	$-0.060^{+0.304}_{-0.351}$	$-0.033^{+0.270}_{-0.316}$
	γ_{ext}	–	0.032	$0.031^{+0.002}_{-0.002}$	$0.031^{+0.002}_{-0.002}$
	ϕ_{ext}	°	144.0	$143.536^{+2.960}_{-3.045}$	$143.637^{+2.905}_{-2.798}$
	$x_{0,\mathrm{sub}}$	arcsec	1.9	$1.749^{+0.462}_{-0.483}$	$1.812^{+0.523}_{-0.481}$
	$y_{0,\mathrm{sub}}$	arcsec	−0.4	$-0.248^{+0.816}_{-0.770}$	$-0.275^{+0.735}_{-0.699}$
	$\theta_{\mathrm{E,sub}}$	arcsec	0.07	$0.070^{+0.004}_{-0.004}$	$0.070^{+0.004}_{-0.004}$
GRF	$\log_{10} A_{\mathrm{GRF}}$	–	$-8.231^{+0.005}_{-0.005}$	$-8.242^{+0.005}_{-0.005}$	$-8.088^{+0.001}_{-0.001}$
	β_{GRF}	–	$-3.044^{+0.019}_{-0.019}$	$-3.106^{+0.02}_{-0.02}$	$-3.135^{+0.004}_{-0.004}$

Notes. For the Combo model, the first four parameters refer to the shear and the others to the SIS. The power spectrum parameters of the GRF were fitted within the mask, as were the true perturbations.

Table 2. Smooth potential parameters obtained by fitting an SIE profile to the $\psi_{\mathrm{sm}} + \psi_{\mathrm{pert}}$ models.

Name	Units	Truth	SIS	Shear	Combo	GRF
x_0	arcsec	0.0	$0.124^{+0.009}_{-0.009}$	$0.036^{+0.010}_{-0.009}$	$0.126^{+0.009}_{-0.010}$	$0.043^{+0.010}_{-0.009}$
y_0	arcsec	0.0	$0.031^{+0.008}_{-0.008}$	$0.038^{+0.008}_{-0.008}$	$0.026^{+0.007}_{-0.008}$	$0.045^{+0.008}_{-0.008}$
q_{m}	–	0.73	$0.729^{+0.014}_{-0.014}$	$0.626^{+0.012}_{-0.011}$	$0.627^{+0.011}_{-0.010}$	$0.737^{+0.014}_{-0.015}$
ϕ_{m}	°	82.0	$83.583^{+1.263}_{-1.252}$	$72.790^{+1.043}_{-1.136}$	$71.460^{+1.119}_{-1.055}$	$84.253^{+1.424}_{-1.441}$
θ_{E}	arcsec	1.6	$1.678^{+0.003}_{-0.003}$	$1.612^{+0.004}_{-0.003}$	$1.691^{+0.003}_{-0.003}$	$1.600^{+0.003}_{-0.003}$

Notes. Full reconstructed potential (including the perturbations) is fitted.

Typical ranges for σ_{GRF} and β_{GRF} that have been explored in the literature are $\sigma_{\mathrm{GRF}} \in [10^{-5}, 10^{-2}]$ and $\beta_{\mathrm{GRF}} \in [3, 8]$ (Chatterjee 2019; Vernardos et al. 2020), while Bayer et al. (2018) exclude a GRF variance larger than $\sigma^2_{\mathrm{GRF}} = 10^{-2.5}$, based on HST observations of the strong lens system SDSS J0252+0039. In this work, we set $\log_{10} A_{\mathrm{GRF}} = -8$ (i.e., $\sigma^2_{\mathrm{GRF}} \approx 10^{-3.8}$) and $\beta_{\mathrm{GRF}} = -3$, such that it leads to a GRF that is not unphysically large and contains both small and large-scale features. The resulting GRF realization is shown in Fig. 6. The mean perturbation level computed within the region containing the lensed arcs is about 0.4 percent of the main lens potential.

4.3. External shear

Perturbations on the same scale as the typically used smooth profiles (that is, scales comparable to the Einstein radius θ_{E}) can occur in a lensing galaxy due to higher order moments in its mass distribution. For example, there can be ellipticity gradients (Van de Vyvere et al. 2022), the presence of a bar (Hsueh et al. 2018), or deviations from ellipticity that can be attributed to multipolar azimuthal perturbations (Van de Vyvere et al. 2022). In the latter case, Galan et al. (2022) consider the effects of an octupole; instead, here we consider the more common case of a quadrupole through an external shear perturbation, which simulates first-order effects of massive structures in the vicinity of the lens. This moment is widely used to describe lensing deflections at distances much larger than the Einstein radius, usually induced at the location of a given lens by mass concentrations (e.g., galaxies) that lie close to it and perpendicular to the line of sight. Although this external shear lens potential component has an analytical expression, we choose to treat it agnostically as a perturbation. Motivated by typical values found in the literature, we choose a shear magnitude $\gamma_{\mathrm{ext}} = 0.032$ and direction $\phi_{\mathrm{ext}} = 144°$. The corresponding perturbing lens potential is shown in Fig. 6 and it amounts to 1.1% of the main lens potential within the region containing the lensed arcs.

4.4. Generating the mocks

We simulate data with the independent code MOLET (Vernardos 2022) in order to prevent the occurrence of potentially advantageous minima during optimization. This choice

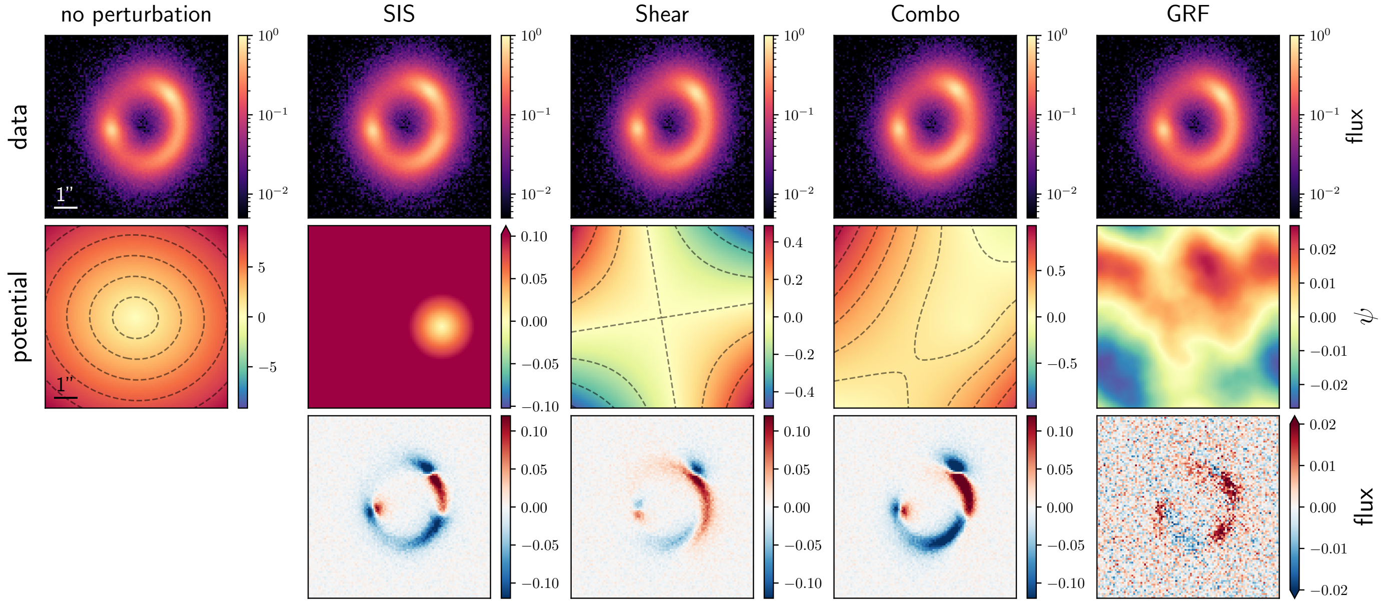

Fig. 6. Mock data modeled in this work. The leftmost column shows the image and lensing potential of a mock system without any perturbations, while subsequent columns show the mock image (first row) of the same system under the influence of a different perturbing potential (second row). The difference with the unperturbed case is shown in the bottom row. Details on generating the mocks and the parameters of the system and perturbations are given in Sect. 4 and in Tables 1 and 2. The Combo case is simply a combination of the SIS and Shear perturbations. Contours are shown for the Shear and Combo cases to better trace the shape of the lensing potential. We also note that the GRF perturbing potential has 2–5 times lower amplitude than the other ones.

also better mimics real-world situations, as the data never exactly correspond to any model generated by the modeling code itself.

We consider typical observations of strongly lensed galaxies as observed with HST and the Wide Field Camera 3 (WFC3) instrument, in the near-infrared ($F160W$ filter). The pixel size is 0.08 arcsec, and the field of view is 8×8 arcsec^2. We do not explore effects due to incorrect PSF modeling, hence for simplicity, we use a Gaussian PSF with 0.3 arcsec full-width-at-half-maximum. We consider an exposure time of 9600 s, which corresponds approximately to 4 HST orbits. The magnitude zero-point is 25.9463 mag, the sky brightness is 22 mag arcsec^{-2}, and the detector readout noise is four electrons, in line with the assumed instrument and filter (Gennaro 2018). The simulated noise contains both a Gaussian component from readout noise and a Poisson component from shot noise.

The source light is taken as a Sérsic profile, with half-light radius $\theta_{\rm eff} = 0.8$ arcsec, Sérsic index $n_{\rm Sersic} = 2$, axis ratio $q = 0.82$, and orientation angle $\phi = 170°$. The center of the distribution is located in the source plane at $(0.40, 0.15)$ arcsec, resulting in the lensed annulus seen in the image plane. We use the same parameter definitions as in Galan et al. (2022) and therefore refer the reader to their Appendix B.

These assumptions for generating the mock data result in a sufficiently realistic experimental setup in which we can evaluate our method. The first column of Fig. 6 shows the image of an unperturbed simulated system, while the remaining columns show mock data corresponding to the different perturbation cases to which we apply our method.

5. Results

We analyze the mock data described above with our neural network-based method to model the lens potential. Our objective is to recover the perturbations to the main lens potential as accurately as possible without any prior assumption on their form. We also assess how well we can reconstruct the full lensing potential without assuming any analytical profile. In all our experiments, we assume perfect knowledge of the source.

5.1. Recovering perturbations to the smooth potential

In this case, the main lens potential is fixed and our continuous neural field model is used to discover a hidden perturbation for which little or no prior information is available. This situation can commonly occur when a smooth model is able to fit the largest-scale features of a system, but as it does not explicitly account for perturbations, localized features are present in the residuals.

Figure 7 shows the model, reconstructed perturbations, and residuals for the four different mock observations that we examine. Visually, the residuals are very close to the noise level (the reduced chi-squared χ^2_ν is unity), and the reconstructed perturbation fields very closely match the truth. In the case of the localized SIS substructure, the model is able to recover the correct shape, magnitude, and location of the perturbation. Similarly, the shape and amplitude of the external Shear and Combo perturbations are very close to the truth. For the GRF perturbation, the network is able to retrieve its general properties and detailed features, but small artifacts appear near the edges of the field of view. In fact, we should expect the fidelity of the reconstructions to decrease the further we are from the gravitationally lensed flux of the source, roughly indicated by the mask in the second row of Fig. 7, due to the lack of signal. Despite the apparent lack of signal, our method also reconstructs the perturbations outside the mask remarkably well, even in the case of the complex GRF, albeit with less accuracy.

To further evaluate the quality of the reconstructions, we fit analytic SIS and external shear profiles to the reconstructions of the SIS, Shear, and Combo models. The results are listed in Table 1, where we see that the true value of the parameters is

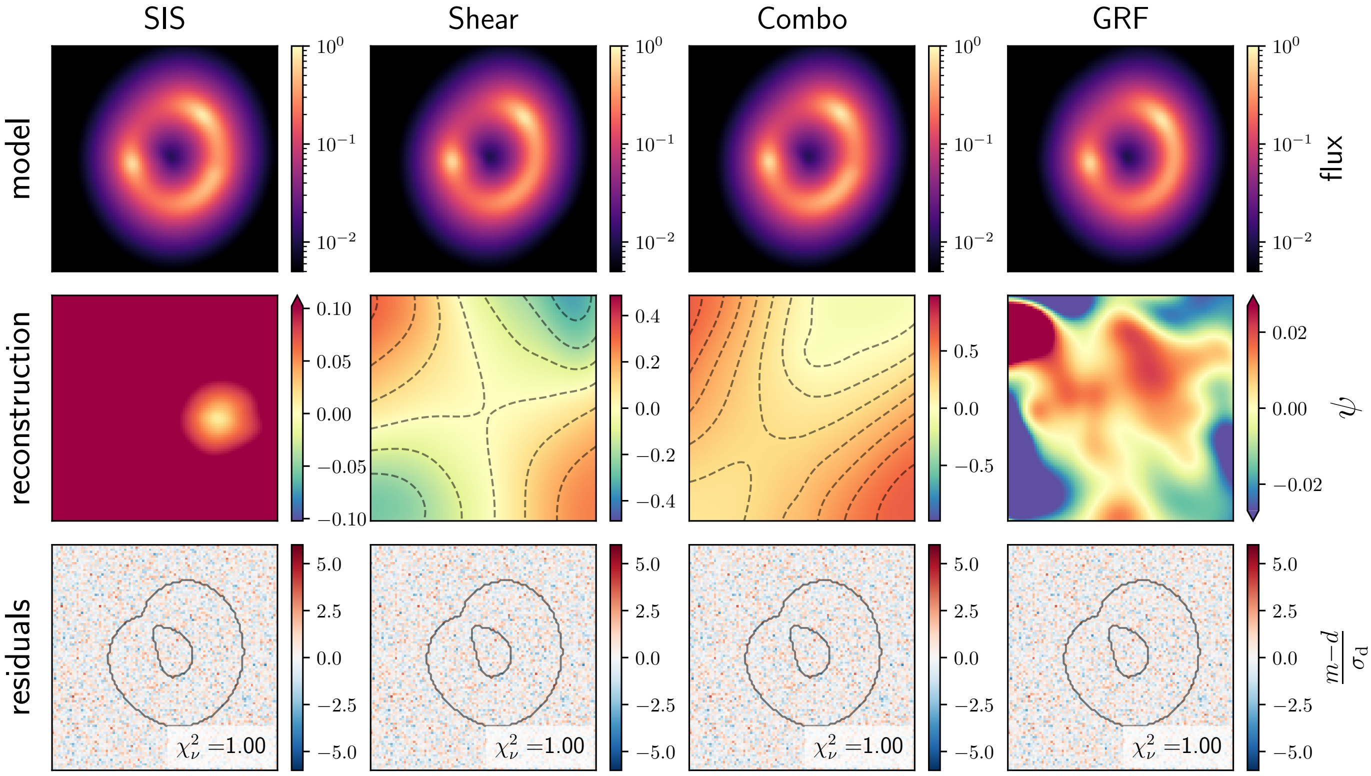

Fig. 7. Top row: models of the mock data shown in Fig. 6. Middle row: reconstructed perturbations. Bottom: model residuals in normalized units, i.e., divided by the diagonal, σ_d, of the data covariance matrix from Eq. (14). We note in particular that within the area of the most significant lensed flux, as indicated by the mask (solid contour), the residuals are identical to noise. The reduced chi-squared χ^2_ν of the models is also given in the panels of the bottom row.

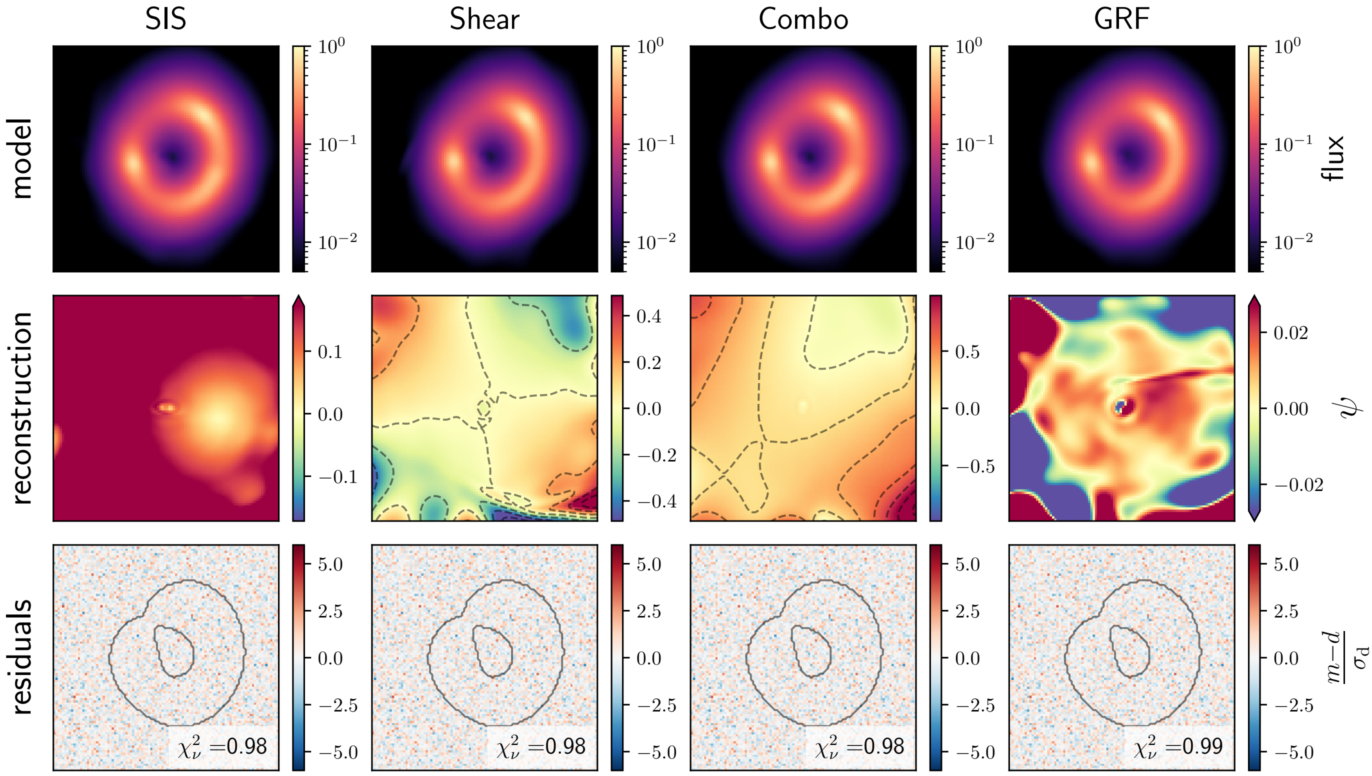

Fig. 8. Same as Fig. 7, but in this case, the full potential was modeled and then the true main SIE was subtracted in order to obtain the reconstructions shown here. We changed the amplitude scale of the reconstructed SIS (left column) to better outline some detailed structure (artifacts) that appear for this model.

always within the 1σ confidence interval of our fits. In a similar fashion, we calculate the amplitude and slope of the power spectrum of the GRF reconstruction. This is done within the mask (as well for the true underlying GRF), where we find systematic errors of only a few percent.

5.2. Modeling the full potential

We also model the same mock data without any knowledge of the main, smooth lens potential component of the lens, namely without assuming an SIE profile. We expect this less constrained version of the problem to be more degenerate and harder to optimize, as our neural network method must now account for both an extended smooth potential and local perturbations. However, based on the Combo case examined above, which does combine extended and local perturbations albeit in a limited way, we do not expect this to fail entirely.

Figure 8 shows the model, reconstructed perturbations after subtracting the true SIE main potential, and residuals for the four different data cases. As when fitting only the perturbations, the residuals here are visually very close to the noise level, with an indication of slight overfitting from the χ^2_ν being below unity. The reconstructed fields for the SIS, Shear, and Combo cases closely match the true perturbations, although artifacts outside the mask are amplified, in particular around the edges and in the center of the field of view. This is also supported by the analytic fits, for which parameters are listed in the last column of Table 1, and which are almost always within 1σ confidence interval from the truth. The reconstructed GRF perturbation visually bears little resemblance to the true field, although its power spectrum is recovered remarkably well, and its fitted power-law parameters have a systematic bias of just a few percent.

Having fitted the full potential, we can also investigate how well the main smooth potential is reconstructed. We do so by fitting an analytic SIE profile directly to the full reconstruction (meaning the perturbations are included). The fitted parameters are listed in Table 2 and show small systematic biases of a few percent, except for the ellipticity and position angle in the Shear and Combo models. This reflects the well-known degeneracy between these two parameters and the magnitude and orientation of the shear (e.g., see Fig. 3 of Vernardos & Koopmans 2022). The presence of a massive subhalo in the SIS and Combo models also biases the value of $\theta_{\rm E}$ toward higher values, because more mass is directly added to the total lens potential. This is not the case with the GRF perturbations, which should average to zero within the mask and whose amplitude is too low to bias $\theta_{\rm E}$ anyway.

6. Discussion and conclusion

In this work, we present a new way of using neural networks to address the problem of mass reconstruction in galaxy-galaxy gravitational lensing. By being embedded in a model that explicitly satisfies the lens equation, our neural network is agnostic to any specific mass distribution and converges to the solution using automatically calculated derivatives. The method is able to capture both large-scale (i.e., spanning the field of view) and small-scale (i.e., on the order of a few pixels) features in lensing potentials, corresponding to roughly two orders of magnitude in terms of angular scale.

We were able to successfully model large-scale potentials, such as an external shear component or the combination of a primary smooth potential (e.g., SIE) and a perturbing field representing a single massive subhalo, an external shear, or a population of lower-mass subhalos. In the latter case, the parameters of the smooth potential were recovered in a subsequent fit with biases of only a few percent (including the effect of the perturbations), without using any analytical profile or other prior knowledge during the reconstruction (see Table 2).

The parameters of the perturbing potentials were well recovered in the case of perfect knowledge of the main potential (see Fig. 7 and Table 1). This is also true of the reconstruction in areas outside of the mask and away from where most of the lensed flux is observed. Given the much lower signal-to-noise ratio in these pixels, such accuracy is not necessarily anticipated. When reconstructing the full potential, it is primarily in these regions that artifacts appear. Nevertheless, the reconstructions are still good enough without model residuals such that we could recover the parameters of the perturbations with very small biases.

Regarding the case of a subhalo population, which is the most complicated but also the weakest perturbation we examined, its detailed structure is remarkably well reconstructed if perfect knowledge of the main potential is assumed (see the bottom right panel in Fig. 4). On the other hand, assuming no knowledge of the underlying potential, the statistical properties of the population as described by its power spectrum are still very well recovered (see Table 1), despite the actual reconstructed potential being qualitatively different from the truth. A similar result was obtained by Vernardos & Koopmans (2022), who used the semi-linear inversion technique to reconstruct the perturbations.

Our method relies on the standard deviation σ of the Fourier feature mapping (see Sect. 3 and Eq. (12)) as a hyperparameter. As illustrated in Fig. 3, the higher its value, the finer the features (including possible noise) in the data that can be captured by the neural network. This is in direct analogy to the regularization strength in pixelated models, such as the semi-linear inversion technique (e.g., Vernardos & Koopmans 2022) and when using wavelets (Galan et al. 2021, 2022). In this work, we chose a value obtained by trial and error, which works well for analyzing the mock data and the given signal-to-noise explored here. However, an interesting avenue for future research is to integrate this hyperparameter into our differentiable inference scheme and obtain its most probable value in a Bayesian way. We leave the development of such a framework for future work.

In this initial presentation of our new method, we examine lens configurations that have a smooth, analytical source brightness profile. This is a deliberate choice due to the strong degeneracy known to exist between the small-scale structure in the source light and the lens mass (see Vernardos & Koopmans 2022, for a study of this degeneracy). Systems with a complicated source structure are expected to affect the signal-to-noise ratio per pixel and consequently the choice of σ. In addition, we kept the source fixed to the truth and did not attempt to reconstruct it simultaneously with the lens potential. Although this is clearly not a realistic scenario, it has allowed us to validate our method in a more restricted parameter space and controlled set of experiments. We intend to explore both effects of a structured and unknown source in future work.

From the point of view of machine learning, improvements to the current model can be approached from several angles. First, the network design can be modified to allow for the quantification of uncertainty. In this way, its output would also include an estimate of the error associated with its prediction. This could be done, for example, by using the MC-Dropout technique as in Levasseur et al. (2017) to model the so-called epistemic uncertainty. One can also add an additional output node to the network in order to model the covariance of predictions and capture the so-called aleatoric uncertainty, as in Fluri et al. (2019). Another interesting direction would be to study ways to use better initialization schemes of the network weights, or use pretraining routines, followed by a fine-tuning phase for each new lensing potential. Finally, to the best of the authors' knowledge, this is one of the earliest applications of implicit representation learning to cosmology. In light of the extraordinary success achieved by modern methods designed within this framework, particularly for the synthesis of 3D images (Mildenhall et al. 2020), we believe that the exploration of their application to cosmology represents an exciting avenue for future research.

In conclusion, we have presented a promising machine learning method applied for the first time to lens potential reconstruction problems. Our method can accurately reconstruct, in a continuous way, single massive subhalos, complex subhalo populations, and large-scale potentials that span the entire field of view. For the latter, the simple external shear that we have explored here could be straightforwardly extended to include ellipticity gradients, twists, and higher-order moments (e.g., Van de Vyvere et al. 2022). Our experiments covered two extreme cases, both with successful results: either perfect knowledge or no knowledge of the main lens potential. This motivates the development of a more traditional approach that reconstructs the lens potential as a combination of an analytical profile for the main potential (e.g., SIE or power law) and a free-form field that captures any remaining perturbations. Given that our method is implemented within HERCULENS, we will explore this in future work, as well as degeneracies with the source surface brightness.

Acknowledgements. GV has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodovska-Curie grant agreement No 897124. This research was made possible by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program. This programme is supported by the Swiss National Science Foundation (SNSF) and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (COSMICLENS: grant agreement No 787886).

References

Adam, A., Perreault-Levasseur, L., & Hezaveh, Y. 2022, ArXiv e-prints [arXiv:2207.01073]
Alexander, S., Gleyzer, S., McDonough, E., Toomey, M. W., & Usai, E. 2020a, ApJ, 893, 15
Alexander, S., Gleyzer, S., Parul, H., et al. 2020b, ArXiv e-prints [arXiv:2008.12731]
Babuschkin, I., Baumli, K., Bell, A., et al. 2020, http://github.com/deepmind
Barnes, J. E., & Hernquist, L. 1996, ApJ, 471, 115
Bayer, D., Chatterjee, S., Koopmans, L. V. E., et al. 2018, ArXiv e-prints [arXiv:1803.05952]
Biggio, L., Galan, A., Peel, A., Vernardos, G., & Courbin, F. 2021, in Machine Learning and the Physical Sciences, NeurIPS 2021 Workshop
Boylan-Kolchin, M., Bullock, J. S., & Kaplinghat, M. 2011, MNRAS, 415, L40
Bradbury, J., Frostig, R., Hawkins, P., et al. 2018, JAX: composable transformations of Python+NumPy programs
Brehmer, J., Mishra-Sharma, S., Hermans, J., Louppe, G., & Cranmer, K. 2019, ApJ, 886, 49
Chatterjee, S. 2019, PhD thesis, University of Groningen, The Netherlands
Chatterjee, S., & Koopmans, L. V. E. 2018, MNRAS, 474, 1762
Chianese, M., Coogan, A., Hofma, P., Otten, S., & Weniger, C. 2020, MNRAS, 496, 381
Coogan, A., Karchev, K., & Weniger, C. 2020, ArXiv e-prints [arXiv:2010.07032]
Coros, S., Macklin, M., Thomaszewski, B., & Thürey, N. 2021, in SIGGRAPH Asia 2021 Courses, SA '21 (New York, NY, USA: Association for Computing Machinery)
Cuomo, S., di Cola, V. S., Giampaolo, F., et al. 2022, Scientific Machine Learning through Physics-Informed Neural Networks: Where we are and What's Next (Berlin: Springer)
Cybenko, G. 1989, Math. Control Signals Syst., 2, 303
de Blok, W. J. G. 2010, Adv. Astron., 2010, 789293
Diaz Rivero, A., & Dvorkin, C. 2020, Phys. Rev. D, 101, 023515
Doerr, S., Majewski, M., Pérez, A., et al. 2021, J. Chem. Theor. Comput., 17, 2355
Dubinski, J. 1994, ApJ, 431, 617
Fluri, J., Kacprzak, T., Lucchi, A., et al. 2019, Phys. Rev. D, 100, 063514
Galan, A., Peel, A., Joseph, R., Courbin, F., & Starck, J. L. 2021, A&A, 647, A176
Galan, A., Vernardos, G., Peel, A., Courbin, F., & Starck, J.-L. 2022, A&A, 668, A155
Gennaro, M. 2018, in WFC3 Data Handbook (Baltimore: STScI), 4, 4
Gu, A., Huang, X., Sheu, W., et al. 2022, ApJ, 935, 49
Heek, J., Levskaya, A., Oliver, A., et al. 2020, Flax: A neural network library and ecosystem for JAX
Hermann, J., Schätzle, Z., & Noé, F. 2020, Nat. Chem., 12, 891
Hezaveh, Y. D., Dalal, N., Marrone, D. P., et al. 2016, ApJ, 823, 37
Hezaveh, Y. D., Perreault Levasseur, L., & Marshall, P. J. 2017, Nature, 548, 555
Hsueh, J.-W., Despali, G., Vegetti, S., et al. 2018, MNRAS, 475, 2438
Karchev, K., Coogan, A., & Weniger, C. 2022, MNRAS, 512, 661
Kingma, D. P., & Ba, J. 2014, arXiv e-print [arXiv:1412.6980]
Klypin, A., Kravtsov, A. V., Valenzuela, O., & Prada, F. 1999, ApJ, 522, 82
Koopmans, L. V. E. 2005, MNRAS, 363, 1136
Levasseur, L. P., Hezaveh, Y. D., & Wechsler, R. H. 2017, ApJ, 850, L7
Mildenhall, B., Srinivasan, P. P., Tancik, M., et al. 2020, NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis
Mishra-Sharma, S., & Yang, G. 2022, arXiv e-prints [arXiv:2206.14820]
Moore, B. 1994, Nature, 370, 629
Moore, B., Ghigna, S., Governato, F., et al. 1999, ApJ, 524, L19
Oldham, L. J., & Auger, M. W. 2018, MNRAS, 476, 133
Ostdiek, B., Diaz Rivero, A., & Dvorkin, C. 2022, A&A, 657, L14
Papastergis, E., Giovanelli, R., Haynes, M. P., & Shankar, F. 2015, A&A, 574, A113
Pearson, J., Li, N., & Dye, S. 2019, MNRAS, 488, 991
Perreault Levasseur, L., Hezaveh, Y. D., & Wechsler, R. H. 2017, ApJ, 850, L7
Rahaman, N., Baratin, A., Arpit, D., et al. 2018, arXiv e-print [arXiv:1806.08734]
Raissi, M., Perdikaris, P., & Karniadakis, G. E. 2019, J. Comput. Phys., 378, 686
Ramachandran, P., Zoph, B., & Le, Q. V. 2017, arXiv preprint [arXiv:1710.05941]
Scannapieco, C., Tissera, P. B., White, S. D. M., & Springel, V. 2008, MNRAS, 389, 1137
Schuldt, S., Suyu, S. H., Meinhardt, T., et al. 2021, A&A, 646, A126
Sérsic, J. L. 1963, Boletin de la Asociacion Argentina de Astronomia La Plata Argentina, 6, 41
Shajib, A. J., Treu, T., Birrer, S., & Sonnenfeld, A. 2021, MNRAS, 503, 2380
Sitzmann, V., Martel, J. N., Bergman, A. W., Lindell, D. B., & Wetzstein, G. 2020, arXiv e-print [arXiv:2006.09661]
Springel, V., Frenk, C. S., & White, S. D. M. 2006, Nature, 440, 1137
Tancik, M., Srinivasan, P. P., Mildenhall, B., et al. 2020, arXiv preprint [arXiv:2006.10739]
Thuerey, N., Holl, P., Mueller, M., et al. 2021, arXiv preprint [arXiv:2109.05237]
Toomre, A., & Toomre, J. 1972, ApJ, 178, 623
Van de Vyvere, L., Gomer, M. R., Sluse, D., et al. 2022, A&A, 659, A127
Varma, S., Fairbairn, M., & Figueroa, J. 2020, arXiv e-prints, [arXiv:2005.05353]
Vegetti, S., Koopmans, L. V. E., Bolton, A., Treu, T., & Gavazzi, R. 2010, MNRAS, 408, 1969
Vernardos, G. 2022, MNRAS, 511, 4417
Vernardos, G., & Koopmans, L. V. E. 2022, MNRAS, 516, 1347
Vernardos, G., Tsagkatakis, G., & Pantazis, Y. 2020, MNRAS, 499, 5641
Wagner-Carena, S., Park, J. W., Birrer, S., et al. 2021, ApJ, 909, 187
Wagner-Carena, S., Aalbers, J., Birrer, S., et al. 2023, ApJ, 942, 75
Yao-Yu Lin, J., Yu, H., Morningstar, W., Peng, J., & Holder, G. 2020, arXiv e-prints [arXiv:2010.12960]
Zubovas, K., & King, A. 2012, ApJ, 745, L34