Featured in Physics

DeepLSS: Breaking Parameter Degeneracies in Large-Scale Structure with Deep-Learning Analysis of Combined Probes

Tomasz Kacprzak[*]

Institute for Particle Physics and Astrophysics, ETH Zurich, 8093 Zurich, Switzerland
and Swiss Data Science Center, Paul Scherrer Institute, 5232 Villigen, Switzerland

Janis Fluri

Institute for Particle Physics and Astrophysics, ETH Zurich, 8093 Zurich, Switzerland

(Received 17 March 2022; revised 14 June 2022; accepted 12 July 2022; published 19 August 2022)

In classical cosmological analysis of large-scale structure surveys with two-point functions, the parameter measurement precision is limited by several key degeneracies within the cosmology and astrophysics sectors. For cosmic shear, clustering amplitude σ_8 and matter density Ω_m roughly follow the $S_8 = \sigma_8(\Omega_m/0.3)^{0.5}$ relation. In turn, S_8 is highly correlated with the intrinsic galaxy alignment amplitude A_{IA}. For galaxy clustering, the bias b_g is degenerate with both σ_8 and Ω_m, as well as the stochasticity r_g. Moreover, the redshift evolution of intrinsic alignment (IA) and bias can cause further parameter confusion. A tomographic two-point probe combination can partially lift these degeneracies. In this work we demonstrate that a deep-learning analysis of combined probes of weak gravitational lensing and galaxy clustering, which we call DeepLSS, can effectively break these degeneracies and yield significantly more precise constraints on σ_8, Ω_m, A_{IA}, b_g, r_g, and IA redshift evolution parameter η_{IA}. In a simulated forecast for a stage-III survey, we find that the most significant gains are in the IA sector: the precision of A_{IA} is increased by approximately 8 times and is almost perfectly decorrelated from S_8. Galaxy bias b_g is improved by 1.5 times, stochasticity r_g by 3 times, and the redshift evolution η_{IA} and η_b by 1.6 times. Breaking these degeneracies leads to a significant gain in constraining power for σ_8 and Ω_m, with the figure of merit improved by 15 times. We give an intuitive explanation for the origin of this information gain using sensitivity maps. These results indicate that the fully numerical, map-based forward-modeling approach to cosmological inference with machine learning may play an important role in upcoming large-scale structure surveys. We discuss perspectives and challenges in its practical deployment for a full survey analysis.

DOI: 10.1103/PhysRevX.12.031029 Subject Areas: Cosmology

I. INTRODUCTION

Combined probes of large-scale structure (LSS) contain information about the late-time evolution of the Universe and thus are a unique laboratory for testing cosmological models. The structure of the matter density field, observed through weak gravitational lensing and galaxy clustering, is used to constrain the present-day matter and dark energy densities, Ω_m and Ω_Λ, as well as matter clustering strength σ_8 and the dark energy equation of state w, and other parameters (see Refs. [1–3] for reviews).

In recent years, dedicated LSS observing programs, such as the Dark Energy Survey (DES) [4], the Kilo Degree Survey [5], and the Hyper-Suprime Cam Survey [6], measured cosmological parameters with less than 5% precision [7–9]. Upcoming surveys, such as Euclid [10] and Vera Rubin Observatory (LSST) [11] will provide orders-of-magnitude richer datasets and enable subpercent measurements of these parameters.

In a LSS analysis, cosmology is typically constrained jointly with a large number of other parameters, corresponding to both astrophysical uncertainties and measurement systematics. These uncertainties can constitute a significant part of the cosmology error budget. In particular, the key degeneracies that limit our ability to constrain σ_8 and Ω_m are galaxy intrinsic alignments and galaxy bias [12,13].

Intrinsic alignments (IA) of galaxy shape and its large-scale environment arise due to tidal fields acting on them during their formation and evolution [12,14]. IA effects can perfectly mimic weak gravitational lensing and thus cause a

[*] tomaszk@phys.ethz.ch

degeneracy between their amplitude A_{IA} and S_8 [15]. Galaxy biasing describes how galaxies trace the underlying dark matter density field. Changes in biasing can have a similar effect on clustering maps as changes in cosmology, causing a sharp degeneracy between linear bias parameter b_g and σ_8 and Ω_m [16]. Varying galaxy stochasticity parameter r_g can also mimic the effects of varying cosmology. These degeneracies are even stronger for models that include redshift evolution of IA and biasing, or their higher-order properties. Finally, for weak lensing (WL), the σ_8 and Ω_m are itself degenerate, roughly following the $S_8 = \sigma_8(\Omega_m/0.3)^{0.5}$ relation.

The tomographic probe combination can alleviate these degeneracies due to the use of joint information about galaxy positions and shapes. For the scales considered, two-point (2-pt) functions-based analyses (3×2) managed to reduce these degeneracies, but did not remove them completely [7,8,17,18]. Moreover, this type of analysis has a very limited ability to constrain higher-order properties and redshift evolution of IA and biasing. This can cause further challenges for accurate cosmology measurement: it is now known that the use of wrong IA or biasing models can cause significant errors in inferred cosmology [15,19], due to limited information on these effects that can be extracted from the data. In that case, even the full Bayesian marginalization of these unconstrained parameters can lead to biases due to prior volume effects [20]. Addressing these challenges is particularly important in the light of recent hints of tensions between the S_8 measurements between LSS and the early Universe, as extrapolated from cosmic microwave background measurements [7,8,21–24].

At intermediate and small scales, the late-time LSS density field contains non-Gaussian information, which can also be extracted from lensing and clustering data. Recently, peak statistics [25] and deep learning [26,27] have been demonstrated to achieve significant measurement precision gain of σ_8 and Ω_m from weak lensing maps. The convolutional neural network (CNN) results [26] and the DES Y3 peaks analysis [25] hinted at the possibility of delivering improved IA constraints compared to the power spectrum. However, a map-level probe combination has not been extensively explored to date.

In this paper, we propose a deep-learning analysis of combined probes of weak lensing and galaxy clustering and investigate its potential to break degeneracies between σ_8, Ω_m, r_g, A_{IA}, b_g, as well as their redshift evolution, parametrized by a simple power law with $\eta_{A_{\mathrm{IA}}}$ and η_{b_g}. We use a fully forward-modeling approach, where we train CNNs on a stacked tomographic weak lensing convergence κ_g and galaxy counts δ_g. This is similar to the 2D "counts-in-cells" technique [28,29], where the galaxy catalogs are binned into pixel or voxel "cells," which are then analyzed directly without using 2-pt functions.

We compare the results of the CNN analysis with the equivalent tomographic power spectrum (PSD) analysis, which uses data vectors similar to the 3×2 method. We use a stage-III simulated survey configuration with 900 deg^2, 10 galaxies/arcmin^2, and four broad redshift bins. The analysis is limited to intermediate scales, corresponding to smoothing FWHM of 8 Mpc/h for galaxy clustering and 4 Mpc/h for weak lensing. We also explore a small-scales configuration with nonlinear bias, with smoothing at 4 Mpc/h for both probes. We report the forecasted cosmology constraints expected from the CNN and PSD, and describe the information gain for all parameters considered. Finally, we investigate the sensitivity maps for CNNs to gain more intuition about the origin of the information used by CNN.

This paper is structured as follows. We describe the simulated lensing and clustering data in Sec. II. The measurement methods for deep learning and power spectrum are presented in Sec. III. We describe our results in Sec. V. We investigate the origin of the information used by our methods in Sec. VI and conclude in Sec. VII.

II. THEORY MODELING

We create a set of consistent tomographic maps of weak lensing convergence κ_g and galaxy clustering δ_g. We use a flat ΛCDM model for cosmology with fixed dark energy equation of state $w = -1$. We leave out the Ω_b, n_s, and H_0 parameters, as they are practically impossible to constrain with this data practically impossible to constrain with this data; we do not expect significant differences to the results if these parameters were marginalized out, as in Ref. [25]. For intrinsic alignments, we employ the nonlinear-linear alignment model [30–32]. The galaxy number counts follow a biasing model, with optional nonlinear terms. The redshift evolution of intrinsic alignments and galaxy biasing is also included in this model. We do not include the redshift space distortions or magnification contributions to clustering in this work, as it is subdominant to other effects included. The full model

TABLE I. Summary of parameters used in the data models. Parameters $b_{g,2}$ and $\eta_{b_{g,2}}$ are used only in the nonlinear galaxy bias model. Parameters Ω_m and σ_8 are additionally restricted by convex hull prior defined by the simulation grid.

	Description	Prior	Fiducial
Ω_m	Matter density today	[0.15, 0.45]	0.29
σ_8	Clustering amplitude	[0.5, 1.2]	0.71
A_{IA}	Intrinsic alignment amplitude	$[-6, 6]$	0.5
b_g	Linear galaxy bias	[0.5, 2.5]	1.5
r_g	Galaxy stochasticity	[0.4, 1]	0.7
$\eta_{A_{\mathrm{IA}}}$	IA redshift evolution	$[-4, 6]$	1.6
η_{b_g}	Linear bias evolution	$[-3, 3]$	0.5
$b_{g,2}$	Nonlinear galaxy bias	$[-3, 1]$	0.5
$\eta_{b_{g,2}}$	Nonlinear bias evolution	$[-2, 2]$	0.0

has the parameters described below and summarized in Table I.

(i) We vary cosmology parameters Ω_m, σ_8, with other parameters fixed at $\Omega_b = 0.0493$, $H_0 = 67.36$, and $n_s = 0.9649$, which corresponds to the baseline results (ΛCDM, TT, TE, EE + lowE + lensing) of Planck 2018 [33].

(ii) Intrinsic alignment amplitude is controlled by A_{IA}, while its redshift dependence is controlled by the power law $\eta_{A_{\mathrm{IA}}}$ [see Eq. (8) below]. We do not include the luminosity or color dependence of intrinsic alignments.

(iii) Galaxy density field is controlled by linear bias b_g and its redshift evolution η_{b_g}, which is also a power law parameter [Eq. (9)]. The galaxy stochasticity r_g captures the degree of correlation between the galaxy and matter density field. We reduce the correlation between these fields by adding uniform noise to phases of the galaxy density field. In our extended model, we also add nonlinear galaxy bias $b_{g,2}$ and its redshift evolution $\eta_{b_{g,2}}$.

We consider a stage-III-like survey configuration with 900 deg^2 with 10 galaxies/arcmin^2 distributed evenly in 4 broad redshift bins. The redshift bins are shown in Appendix A, and have the following mean redshifts: $\langle z\rangle = 0.31,\ 0.48,\ 0.75,\ 0.94$. We use the same galaxy selection for both lensing and clustering; we do not create a separate "lens" sample, as is often done [16,34,35]. We also do not include uncertainties in measurement systematic biases, such as redshift errors, shear calibration, or selection function uncertainty for clustering. These may play an important role amplifying degeneracies in the parameter measurements, especially in the redshift error and IA sector. We leave the optimization of the lens sample and the investigation of the systematics effects to future work.

A. Multiprobe maps

We closely follow the method we introduced in Ref. [26] for calculating convergence κ_g maps and create the corresponding clustering δ_g maps using the same pencil-beam simulations. The simulations in [26] consist of 57 unique cosmologies spanning the Ω_m–σ_8 plane. In that work we used the PkdGrav3 code [36] to run a total of 12 simulations at each cosmology. Each simulation used 256^3 particles in a volume of 500^3 Mpc^3, and the initial conditions were generated at redshift $z_{\mathrm{init}} = 50$, using the MUSIC code [37]. All simulations were run with 500 time steps, writing snapshots at the interval of $\Delta z = 0.1$ from $z = 3.45$ to $z = 1.55$ and $\Delta z = 0.05$ down to redshift $z = 0$. See [26] for more details of these simulations.

The 2D map of these fields m_{2D} is projected from the 3D simulated overdensity δ_{3D} using the Born approximation. Maps are calculated using the UFALCON code [38], in a very similar way as Refs. [25–27,39]. The 2D maps are calculated using the following equation:

$$m^{\mathrm{pix}}_{\mathrm{2D}} \approx \sum_b W^m \int_{\Delta z_b} \frac{dz}{E(z)} \delta_{\mathrm{3D}}\left[\frac{c}{H_0}\mathcal{D}(z)\hat{n}^{\mathrm{pix}}, z\right], \qquad (1)$$

where W_m is the weight kernel corresponding to the considered field, $\mathcal{D}(z)$ is the dimensionless comoving distance, $\hat{n}^{\mathrm{pix}}$ is a unit vector pointing to the pixel's center, and $E(z)$ is given by $d\mathcal{D} = dz/E(z)$. The sum runs over all redshift shells and Δz_b is the thickness of shell b.

The kernels corresponding to weak lensing W^{WL}, intrinsic alignments W^{IA}, and galaxy clustering W^G are

$$W^{\mathrm{WL}} = \frac{3}{2}\Omega_m \frac{\int_{\Delta z_b} \frac{\mathrm{d}z}{E(z)} \int_z^{z_s} dz' n(z') \frac{\mathcal{D}(z)\mathcal{D}(z,z')}{\mathcal{D}(z')} \frac{1}{a(z)}}{\int_{\Delta z_b} \frac{dz}{E(z)} \int_{z_0}^{z_s} dz' n(z')}, \qquad (2)$$

$$W^{\mathrm{IA}} = \frac{\int_{\Delta z_b} dz F(z) n(z)}{\int_{\Delta z_b} \frac{dz}{E(z)} \int_{z_0}^{z_s} dz' n(z')}, \qquad (3)$$

$$W^G = \frac{\int_{\Delta z_b} dz n(z)}{\int_{\Delta z_b} \frac{dz}{E(z)} \int_{z_0}^{z_s} dz' n(z')}, \qquad (4)$$

where $n(z)$ is the redshift distribution of galaxies in a given bin, z_s and z_0 are the source and observer redshifts, respectively, and $F(z)$ is a cosmology and redshift-dependent term:

$$F(z) = -C_1 \rho_{\mathrm{crit}} \frac{\Omega_m}{D_+(z)}, \qquad (5)$$

where $C_1 = 5 \times 10^{-14}\ \mathrm{h}^{-2}\ M_\odot\,\mathrm{Mpc}^3$ is a normalization constant, ρ_{crit} is the critical density at $z = 0$, and $D_+(z)$ is the normalized linear growth factor, so that $D_+(0) = 1$.

We use kernels W^{WL}, W^{IA}, and W^G with Eq. (1) to create convergence maps of lensing κ_m, intrinsic alignment κ_{IA}, and galaxy density contrast δ_m. Then, we subtract the mean of the convergence fields and normalize the galaxy density contrast:

$$\kappa \leftarrow \kappa - \langle\kappa\rangle, \qquad (6)$$

$$\delta \leftarrow (\delta - \langle\delta\rangle)/\langle\delta\rangle. \qquad (7)$$

The redshift dependence of IA and biasing is calculated for a given redshift bin $n^i(z)$ by using a power law. A similar method was previously used in Ref. [25]. We use a simplified formulation of this model, where we calculate a single effective scaling value per redshift bin i using the $n^i(z)$ only:

$$A^i_{\mathrm{IA}} = A_{\mathrm{IA}} \int_z dz n^i(z) \left(\frac{1+z}{1+z_0}\right)^{\eta_{A_{\mathrm{IA}}}}, \qquad (8)$$

$$b_g^i = b_g \int_z dz n^i(z) \left(\frac{1+z}{1+z_0} \right)^{\eta_{b_g}}, \qquad (9)$$

where $z_0 = 0.7$ is the fixed pivot redshift. This allows us to remove the redshift dependence in the process of summing the shell maps to create projected maps. By doing this, we can precompute a set of field maps and apply the IA and biasing variation on the fly during training and predictions steps. We verified that this approximation gives similar results to the full formulation. The exact implementation of this dependence is not crucial in this work, as its main focus is a constraining power forecast. To make the interpretation of the values of η_{b_g} parameter easier, we calculated the values corresponding to uncertainty of current measurements from the results in DES Y3 combined probes [7]. We used the linear bias measurements (their Table V) and fitted them with the η-evolution model in Eq. (9). The uncertainty in these measurements translates to the uncertainty on bias evolution $\sigma[\eta_{b_g}] = 0.21$.

To create the forward-modeled probe maps, we add the noise to the pixels directly. The observed lensing map κ_g and galaxy counts maps δ_g are

$$\delta_g = \text{Poisson}[\bar{n}_{\text{gal}}(1 + b_g \delta_m^r)], \qquad (10)$$

$$\kappa_g = \text{Normal}\left[\kappa_{\text{WL}} + A_{\text{IA}} \kappa_{\text{IA}}, \frac{\sigma_e}{\sqrt{\delta_g}} \right], \qquad (11)$$

where $\bar{n}_{\text{gal}}$ is the mean number of galaxies per map pixel, and $\sigma_e = 0.4$ is the galaxy shape noise, which does not change across redshift bins as we assume the same number of galaxies from each bin. We create an auxiliary map δ_m^r to be partially decorrelated from the density contrast δ_m, such that the correlation coefficient is r_g. This is calculated from Fourier transform $\tilde{\delta}_m$, such that their phase angles $\angle$ are related by

$$\angle \tilde{\delta}_m^r = \angle \tilde{\delta}_m + (1 - r_g)^{2/3} \times \text{Uniform}[-\pi, \pi], \qquad (12)$$

which reduces to $\delta_m^r = \delta_m$ for $r_g = 1$. We create this relation empirically for the δ_m maps we considered. We find that this relation gives Pearson's correlation coefficient of pixel values roughly $\text{corr}[\delta_m^r, \delta_m] = r_g$. This method is described in more detail in Appendix B. While this implementation may differ from others used in previous work [13,28,40], it gives a full degree of variation and should be sufficient for the purpose of this study.

For a model that includes nonlinear galaxy bias, the galaxy counts map is calculated as

$$\delta_g = \text{Poisson}[\bar{n}_{\text{gal}}[1 + b_g \delta_m^r + b_{g,2}(\delta_m^r)^2]], \qquad (13)$$

where $b_{g,2}$ controls the strength of nonlinear biasing and follows the same redshift evolution of the other fields [Eq. (9)], controlled by the parameter $\eta_{b_{g,2}}$.

We use these maps consistently for both individual and combined probe analyses; for the lensing maps, the noise level is calculated using the actual galaxy density map in each bin. While lensing-only inference does not aim to constrain b_g, r_g, η_{b_g} parameters, they are still used to make the maps. Their values are always taken from the same

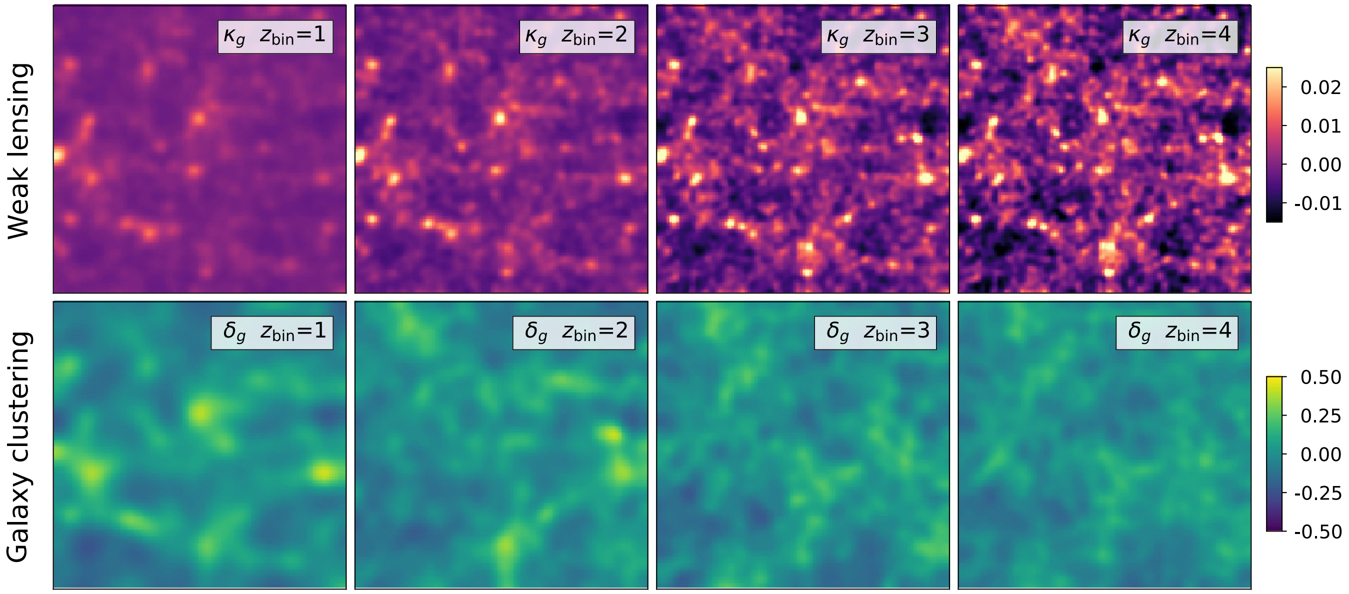

FIG. 1. Example maps for weak lensing convergence κ_m and galaxy clustering δ_m, before the addition of noise. The size of the maps is 5×5 deg. The maps were smoothed with the redshift bin-dependent Gaussian kernel, with FWHM corresponding to ~4 Mpc/h at the mean redshift of the bin for weak lensing and ~8 Mpc/h for galaxy clustering (see Sec. II B).

prior, shown in Table I. This way the lensing-only networks effectively marginalize over the uncertainty on these parameters without aiming to constrain them.

As the Poisson random generator does not have graphics processing unit (GPU) kernels available in TensorFlow, we use the inverse Anscombe transform [41] to approximate it, as described in Appendix F. We perform a number of additional scalings on the map to ensure numerical stability of the training for all parameters in the prior range; see Appendix E.

B. Scale cuts and smoothing

In this work we do not attempt to model baryonic effects, which can strongly modify the matter distribution at small scales [42,43]. Similarly, the galaxy biasing on small scales can significantly deviate from the linear bias model [16]. To avoid using small scales, we smooth the maps with a redshift-dependent kernel. Following choices in Ref. [16], we use the smoothing scales of $R = 4$ Mpc/h for lensing, $R = 8$ Mpc/h for clustering maps for the linear bias model, and $R = 4$ Mpc/h for the nonlinear bias model. Including the pixel kernel, which is of size $d = 4.68$ arcmin, we calculate that the additional Gaussian smoothing to apply is

$$R = 4\ \mathrm{Mpc/h} \Rightarrow \sigma = [4.8, 3.5, 2.8, 2.5]\ \mathrm{arcmin},$$

$$R = 8\ \mathrm{Mpc/h} \Rightarrow \sigma = [9.8, 7.4, 6.0, 5.6]\ \mathrm{arcmin}$$

for four redshift bins used in our analysis. An example of the maps is shown in Fig. 1.

In the original simulations from [26], the field size is 5×5 deg and 128×128 pixels. As we use larger scales in this work, we down-sample the original maps to the size of 64×64, which results in pixel size of $d = 4.68$ arcmin. The standard deviation of this top hat kernel is $\sigma = 1.35$. In [26], the survey consisted of 20 fields, which were passed to the networks as channels. We construct random simulated surveys of 900 deg^2 by using 36 fields. We employ, however, a different method of including these fields: we create a mosaic of size 6×6 fields. We add noise on the fly at each realization during training and prediction. We avoid repeating the same mosaic by placing each field randomly within it, as well as performing a random flip. The smoothing is done before creating the mosaic to avoid blurring over sharp edges.

III. DEEP LEARNING AND POWER SPECTRA

We analyze the maps using two approaches: a convolutional neural network and a power spectrum. The CNN maps from pixel maps to summary statistics corresponding to the model parameters. In this work, we use a residual network architecture [44]. CNNs pass the input maps through convolutional layers, which create a large set of feature maps by convolving the input with a set of learnable filters. This process is repeated from layer to layer, with feature maps being produced at lower resolution each time. They are followed by residual layers, which operate on feature maps created from a residual between input and output of the previous layer, keeping the resolution constant. Finally, the feature maps are flattened and passed to the output layer, which gives informative summary statistics. For a general introduction to deep learning, see Ref. [45].

For the PSD, we also use the summary statistics compression. We calculate all auto- and cross-spectra from the maps, and then employ a simple neural network (NN) to compress these PSD vectors into summary statistics. We use this approach for operational reasons, as it removes the necessity of creating a dedicated likelihood analysis for PSDs only. A similar approach was used in Ref. [27]. The PSD NNs can be trained simultaneously with the CNN, with minimal time overhead. Both CNN and PSD NNs give then the same form of summary output that can be used in a likelihood analysis the same way.

We create a separate network for combined probes (*CP* CNN), which takes a stack of 4 κ_g and 4 δ_g maps, as well as separate networks for lensing (KG CNN) and clustering (DG CNN). It starts with 4 convolutional layers, which create 128 feature maps, reducing the dimensionality of the mosaic images from 384×384 to 24×24. It is followed by 10 residual layers, and kernel size of 5 with stride of 2, and the Relu activation function. The final residual layer is flattened and fully connected to the output, which contains the summaries and their covariances. That gives a network with ∼10–12 million trainable parameters, depending on the model.

As the equivalent NN for the PSD case, we also create separate networks for combined probes (*CP* PSD NN), lensing (KG PSD NN), and clustering (DG PSD NN). The power spectra were calculated using FFT and averaged in 20ℓ bins in the range of $\ell \in [36, 4536]$. The bins are spaced in logarithmic way, with the minimum interval of $\delta\ell = 36$, which is the resolution of the FFT given the pixel sizes used. We first concatenate the auto- and cross-spectra from the maps, which gives 36 and 10 spectra for combined and individual probes, respectively. We do not cut the scales, as the smoothing already removes most small-scale power. Then we pass the PSDs through 2 fully connected layers with 1024 hidden units, also with the Relu activation. Finally, the output layer predicts the summaries and their covariances, as for the CNN. This gives ∼12–1.8 million trainable parameters, depending on the output size. For the PSD NNs, we test a number of architectures and model sizes. We find that they all give very similar results and that choice does not change in comparison with the CNNs. This test is described in Appendix D. Further details of the implementation of these networks can be found in the public DeepLSS code repository.

The training spanned the Ω_m and σ_8 parameters fixed at the 57 simulation points, and the values for the remaining parameters were drawn from a flat prior using the first 16m samples from the Sobol sequence, with parameter ranges shown in Table I. We employ a similar training strategy as [26], both for CNNs and PSD NNs. We use the negative log-likelihood loss function:

$$L = \ln(|\Sigma_p|) + (\theta_p - \theta_t)^\top \Sigma_p^{-1} (\theta_p - \theta_t), \qquad (14)$$

where θ_t is the true parameter vector, θ_p is the summary vector, and Σ_p is the predicted covariance matrix. Note that Σ_p is not used later for inference; it is simply just a part of the likelihood loss formulation. We found that the likelihood is not well described by a single Gaussian, and decided to create the likelihood empirically, using a more complicated model, described in Sec. IV. We train the networks using stochastic gradient descent with the ADAM optimizer [46] with batch size of 32 mosaic maps and learning rates of 0.000 05 and 0.0025 for CNNs and PSD NNs, respectively. We additionally applied gradient clipping using the method of Ref. [47], using 50% percentile. The training took 885000 batches and the loss did not improve over the last 200000 batches, which indicates reasonable convergence. The networks were created in TensorFlow [48] and trained on NVIDIA A100-SXM4–40 GB GPUs.

To avoid overfitting the training set, we calculate the loss for the test set, which is not used during training. The test set consisted of ≈8% of the full simulation set. We did not notice significant differences between the training and testing loss at any point in the training process. This indicates that the networks have a good generalization ability and do not overfit the training set. The reason for this is the on-the-fly addition of noise, which is a very efficient regularizer.

IV. LIKELIHOOD ANALYSIS

The CNN and PSD NN networks output a number of summary statistics, which are then interpreted in a Bayesian framework. The final posterior distribution follows the Bayes rule $p(\theta_t|\theta_p) \propto p(\theta_p|\theta_t)p(\theta_t)$, where θ_p is the network output summary statistic and θ_t is the true parameter value, $p(\theta_p|\theta_t)$ is the likelihood, and $p(\theta_t)$ is the prior.

We estimate the conditional probability distribution $p(\theta_p|\theta_t)$ in the following way. We run a prediction for a set of 7 615 200 samples from $p(\theta_p|\theta_t)$, with 228 000 unique parameter combinations. The prediction set was created using the same parameter sampling scheme as the training set. We then create a model of $p(\theta_p|\theta_t)$ using a mixture density network (MDN), which uses a mixture of Gaussians at each θ_t; it predicts the relative weights $w_j(\theta_t)$ of components, their means $\mu_j(\theta_t)$ and covariances $\Psi_j(\theta_t)$, where $j = 1, \ldots, K$, where K is the number of Gaussians in the mixture model. We train this model using stochastic gradient descent and monitor its validation loss to prevent overfitting; see Appendix C for details. We confirmed that this network predicts the right means and variances for our choice of $K = 4$, for all θ_t, in Appendix C.

The final constraints are calculated using the Markov chain Monte Carlo method using the EMCEE algorithm [49]. We use flat priors on parameters, with ranges shown in Table I, and obtain a chain with 128k samples (1.28m for plotted chains). As the Ω_m and σ_8 parameters were sampled on a grid, we use the convex hull of this grid as the prior for this parameter combination.

V. RESULTS

We calculate constraints for CNNs and PSDs for the fiducial true parameter set θ_t^{fid} given in Table I. We create a mock observation θ_p^{obs} for CNNs and PSDs by taking the most likely prediction for θ_t^{fid}. Figure 2 shows the comparison of the constraints for the combined probes analysis, with CNN result in pink and PSD in blue. The regions correspond to 68% and 95% confidence intervals.

The CNN yields more precise measurement than the PSD for all parameters considered. The strongest constraining power gain is for the intrinsic alignment amplitude A_{IA}, which is of the order of 8 times. CNN measurement effectively breaks the degeneracy between the IA and other parameters. We explore this further in the following paragraphs. Another significant gain comes from breaking the degeneracy between stochasticity r_g and σ_8 and b_g, with r_g improved by 4 times. The measurement of galaxy bias b_g and its evolution η_{b_g} is also decorrelated, with respective improvements of 1.5 times and 1.2 times. Moreover, for CNN, the cosmology parameters Ω_m and σ_8 are also significantly less dependent on the bias evolution, with constraints improved by 1.6 times and 1.3 times, respectively. The redshift evolution of intrinsic alignments is weakly constrained, compared to no constraint by PSD. Overall, breaking of these degeneracies contributes to a very substantial improvement of the Ω_m–σ_8 figure of merit (FOM): the FOM is 15 times higher for the deep learning analysis.

Figure 3 shows the equivalent comparison for individual probes: weak lensing and galaxy clustering. For weak lensing, the CNNs are able to break the σ_8–Ω_m "banana" degeneracy, as previously shown by [26]. There also is an

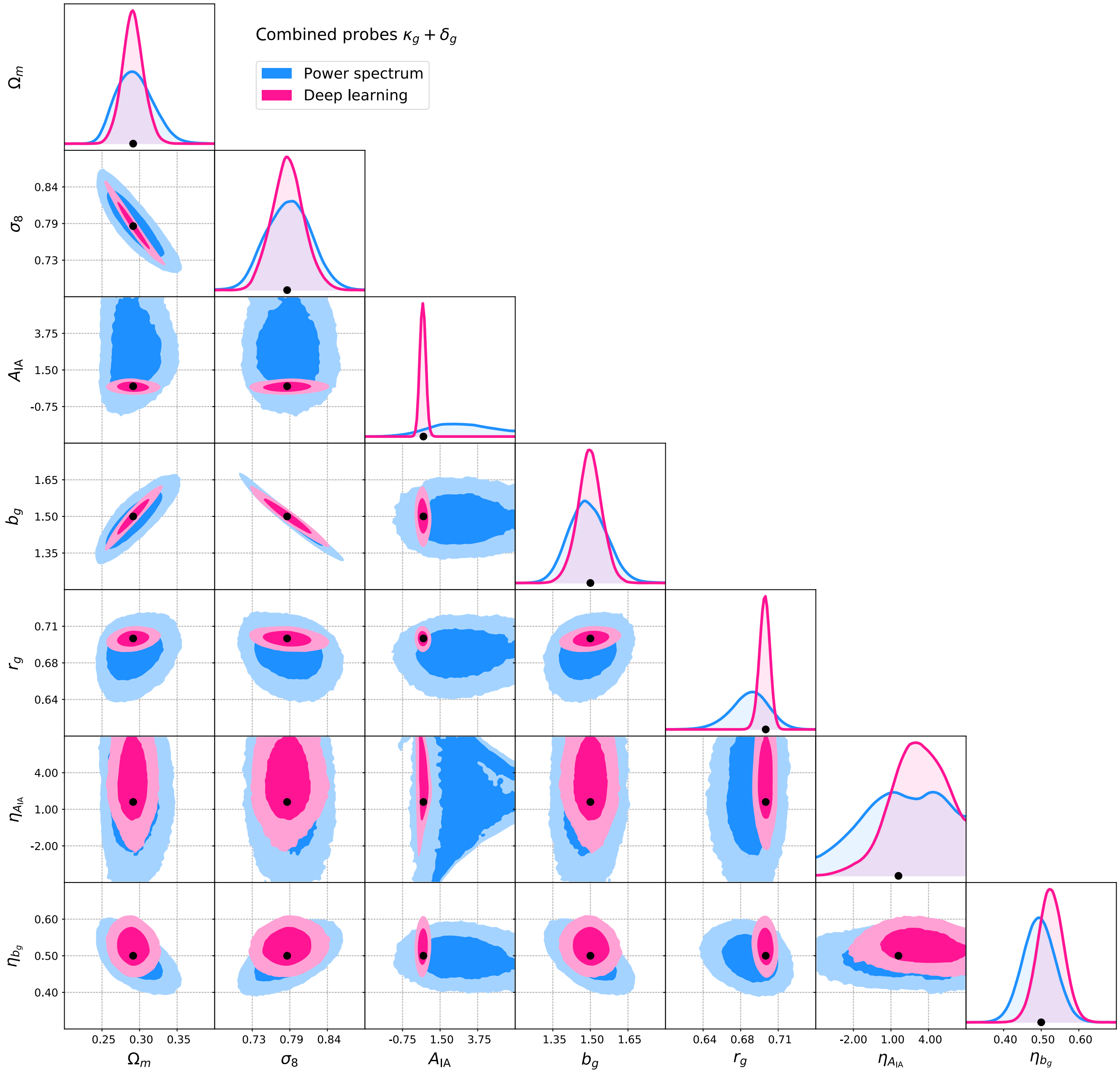

FIG. 2. Constraints for combined probes with deep learning (pink), compared to power spectra (blue), for the main model with linear galaxy bias. The black dots mark the true value of parameters used here, from the fiducial model summarized in Table I. The mock observation used here was taken as the most likely model prediction for the fiducial model.

improvement for the A_{IA} parameter, as also found in [26]. It is, however, significantly smaller than for the combined probes analysis. Moreover, the IA redshift evolution $\eta_{A_{IA}}$ is unconstrained for both CNNs and PSDs. This suggests that the $\eta_{A_{IA}}$ information from combined probes come purely from a better understanding of the κ_g and δ_g maps jointly, since the clustering maps are independent of this IA.

The constraints from galaxy clustering are also significantly improved. The measurement of stochasticity benefits the most from using CNNs, with 4 times better precision. Again, the CNNs break the degeneracy between b_g and its evolution η_{b_g}, which in turn helps with constraining σ_8 and Ω_m. Generally, these parameters are improved by a factor of around 1.3 times to 1.8 times, which is also a significant information gain. The degeneracy between σ_8 and b_g is slightly reduced, with a precision gain on the level of 1.6 times.

The comparison for all parameters and probes is shown in Table II. The gain value was calculated using 200 mock observations selected at random from all the predictions for

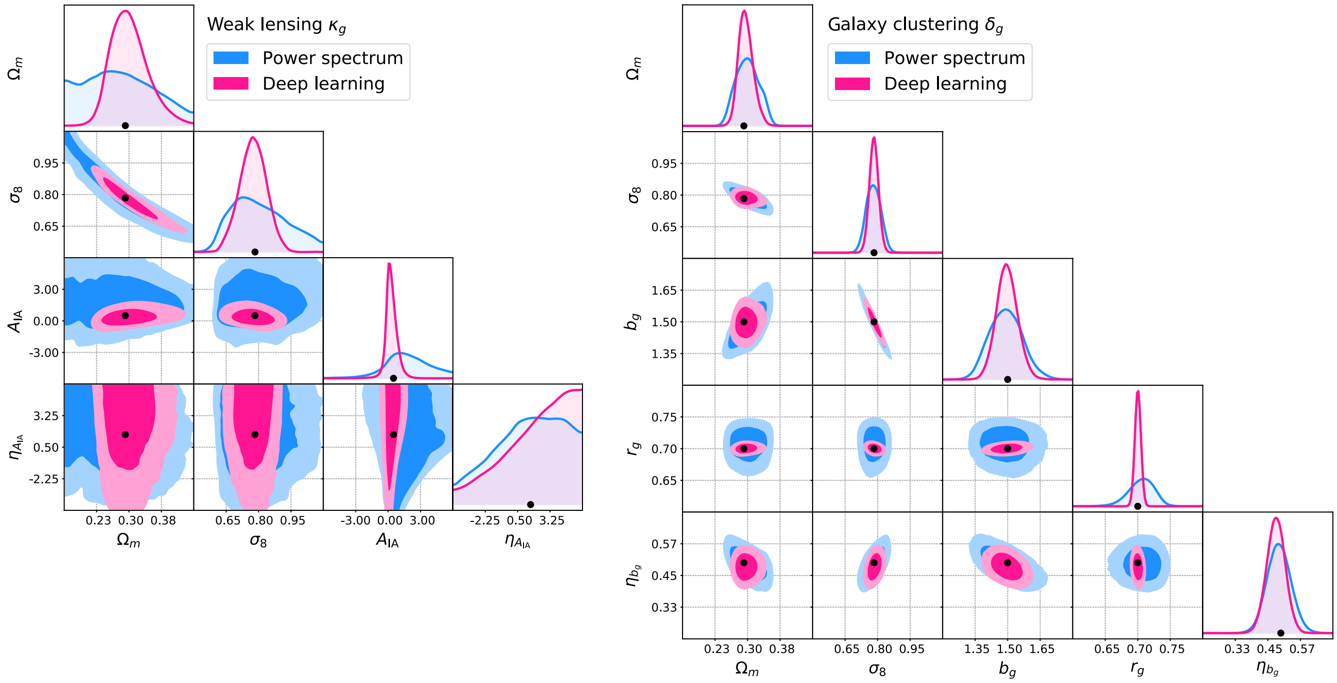

FIG. 3. Constraints for weak lensing (left) and galaxy clustering (right), with deep learning (pink) and power spectra (blue), for the main model with linear galaxy bias. As in Fig. 2, the black dots show the true parameters.

the fiducial parameter set θ_t^{fid}. The value cited is the median of the ratios of parameter standard deviations.

We further investigate the powerful IA constraints obtained by the CNN. Figure 4 shows the uncertainty on S_8 versus A_{IA}. For the PSD, there is a clear degeneracy between these parameters of both lensing and combined probes. For lensing, the degeneracy is still present also for the CNN. It is, however, efficiently broken by CNNs for the probe combination, yielding a 2.3 times improvement in S_8, on average.

We present the results for the nonlinear galaxy biasing model analysis in Fig. 5. For the sake of clarity, we limit the panels to Ω_m, b_g, the nonlinear bias strength $b_{g,2}$, as well as its evolution $\eta_{b_{g,2}}$. For the PSD, a clear degeneracy appears between the linear and nonlinear galaxy bias parameters, b_g and $b_{g,2}$, as well as between the $b_{g,2}$ and its redshift evolution $\eta_{b_{g,2}}$. We also notice that the PSD constraints on linear galaxy bias evolution η_{b_g} are much worse than for the linear bias case, despite using smaller smoothing scales. Deep learning is able to break these degeneracies

TABLE II. Measurement precision gain using with DeepLSS over the power spectrum analysis. Results for galaxy clustering and combined probes are given for two redshift-dependent smoothing scales of galaxy position maps, with FWHM of 8 and 4 Mpc/h for “large scales” and “small scales,” respectively. Weak lensing maps were always smoothed with 4 Mpc/h kernel. The difference is expressed as $\mathrm{SD}(\theta_{\mathrm{PSD}})/\mathrm{SD}(\theta_{\mathrm{CNN}})$, where value of 1× means no precision gain. It was calculated as a median of 200 noisy mock observations taken at the fiducial cosmology.

		Galaxy clustering		Combined probes	
	Weak lensing	Large scales	Small scales	Large scales	Small scales
Ω_m	1.7×	1.5×	1.3×	1.6×	1.5×
σ_8	2.1×	1.6×	1.2×	1.3×	1.2×
S_8	3.0×	0.9×	1.3×	2.3×	1.5×
FOM Ω_m–σ_8	11.0×	1.7×	6.2×	14.9×	13.2×
A_{IA}	3.9×	⋯	⋯	8.3×	8.1×
b_g	⋯	1.6×	1.5×	1.5×	1.4×
r_g	⋯	4.4×	4.3×	3.6×	4.8×
$\eta_{A_{\mathrm{IA}}}$	1.0×	⋯	⋯	1.4×	1.6×
η_{b_g}	⋯	1.3×	1.7×	1.2×	1.7×
$b_{g,2}$	⋯	⋯	2.6×	⋯	2.7×
$\eta_{b_{g,2}}$	⋯	⋯	1.4×	⋯	1.7×

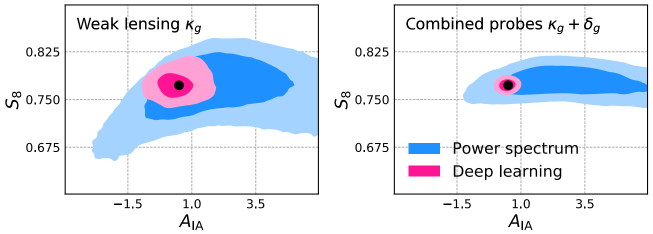

FIG. 4. Constraints on S_8, intrinsic alignments $A_{\rm IA}$, and IA evolution $\eta_{A_{\rm IA}}$, for the combined probes (left) and lensing only (right), using the linear galaxy bias model.

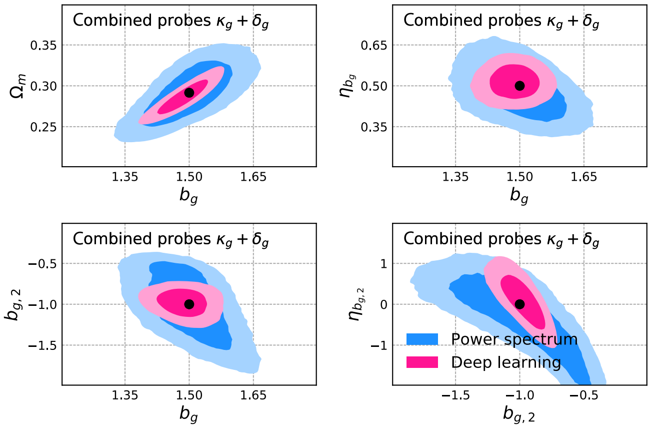

FIG. 5. Galaxy bias constraints for the nonlinear bias model for combined probes. Deep-learning results are shown in pink, while power spectra are shown in blue. To give some intuition about the meaning of the η_{b_g} parameter, we calculated its uncertainty using bias measurements from DES Y3 [7] and obtained $\sigma[\eta_{b_g}] = 0.21$.

and constrain the b_g and $b_{g,2}$ with 1.4 times and 2.7 times precision increase, respectively. The clustering-only analysis of the nonlinear model gives similar gains, as shown in Table II.

VI. WHERE IS THE ADDITIONAL INFORMATION COMING FROM?

Given the significant improvement in constraining power for the CNNs, it is important to gain an intuitive understanding about where the additional information is coming from. Firstly, the nonlinear part of the matter density field contains significant information, even at intermediate scales [25,27]. Multiple approaches have been designed to extract it from the lensing convergence field [39,50,51], most recently tomographic shear peak statistics. Secondly, the full forward analysis on map level enables us to include the phase information of the field; the 2-pt functions are effectively discarding it.

But this may not be the only mechanism employed by the CNNs to increase constraining power. To investigate this further, we look at *sensitivity* maps. For a trained network, one can create such maps by calculating gradients of the network output with respect to the image pixels. In our configuration, where the network outputs are summaries that correspond directly to a given parameter, such a map is easy to create and interpret. Given a set of input maps m and output summaries θ_p, the sensitivity map s is defined as $s = \partial\theta_p/\partial m$. Changes in the regions of the map with highest gradient absolute value will have the most impact on the final prediction, while regions with sensitivity close to zero have almost no impact. However, as all pixels are considered, it does not directly indicate what the value of the predicted parameter will be. Intuitively, one may also interpret these maps as a weight function, which shows which parts of the map are used by the network, and which are ignored. This is a first-order analysis; i.e., it ignores the correlations between pixels and only focuses on the leading term. An approach similar to this has been introduced for lensing maps in Ref. [52], and more broadly in the field of machine learning called *interpretability* [53,54].

We focus on the $A_{\rm IA}$ and σ_8 parameter constraints, as these display the strongest gain and the most benefit from the combined probes analysis. Figure 6 shows input maps and the corresponding sensitivity maps. For the purpose of this figure, additional Gaussian smoothing was applied to the κ_g maps to suppress the noise and make them easier to read. For $A_{\rm IA}$, we show the maps for redshift bin 1, which contain the largest sensitivity signal. We notice that there is a large overdensity in the middle of the δ_g^1 map, as well as in the upper left-hand and right-hand corners. The sensitivity to the κ_g^1 is focused on precisely these regions, while the rest of the map is practically ignored. This makes intuitive sense: the IA signal is expected to be present in the convergence maps at the positions of overdensities. For σ_8, the sensitivity is the strongest for κ_g^2 and δ_g^4. The mean redshifts of these bins are $\langle z^2\rangle \approx 0.5$ and $\langle z^4\rangle \approx 1$, which is exactly the configuration that leads to the strongest lensing signal. The κ_g sensitivity map seems to be focused at positions of peaks, additionally matching their sizes. The δ_g sensitivity is also focused on position of peaks; it seems that the CNNs detect overdensities in the κ_g map and look for corresponding signal in the δ_g map.

The fact that the network only takes information from specific regions of the map seems quite simple, but has profound consequences, especially for the $A_{\rm IA}$ constraint. While the power spectrum analysis also takes advantage of the $\kappa_g \times \delta_g$ cross-correlation, it does not employ such weighting. There, the total cross-correlation signal is an average of regions that are important with regions that are not informative at all. This leads to the dilution of the signal, as the rest of the field contributes only random noise. This in turn lowers the signal-to-noise ratio on the measured parameter and consequently makes the constraints less precise. This can explain the 8 times information gain for $A_{\rm IA}$, as it draws information mostly from the lowest redshift bin. For that bin, the maps are usually sparsely populated

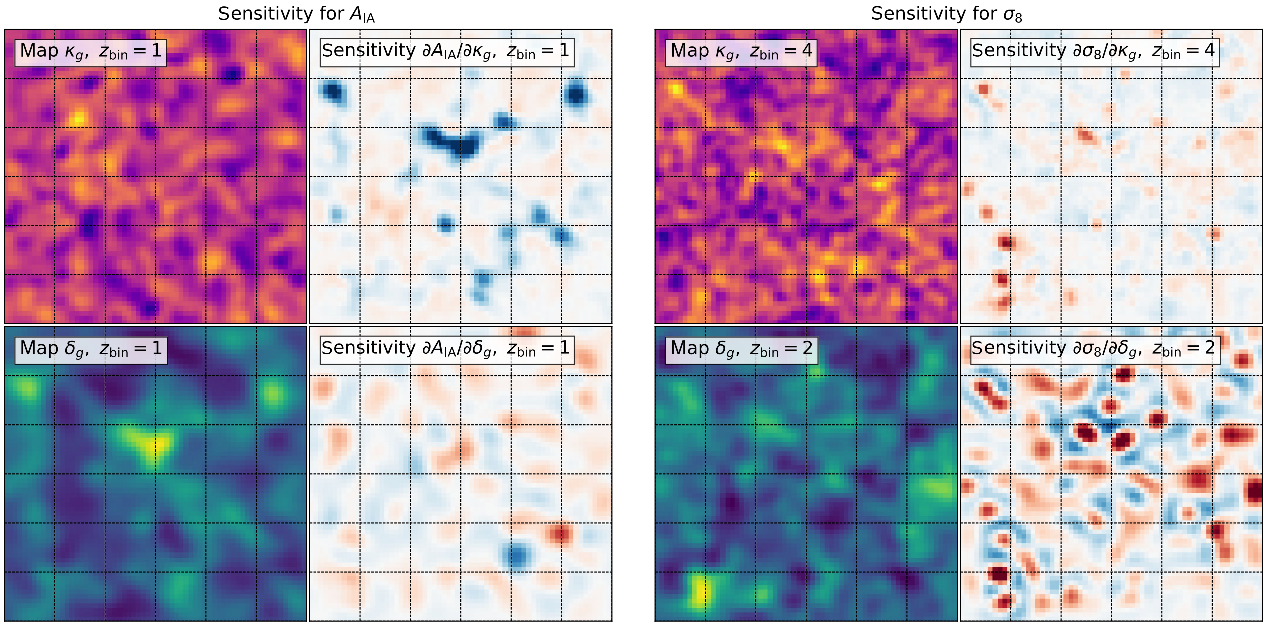

FIG. 6. Sensitivity analysis for the A_{IA} (left-hand panels) and σ_8 (right-hand panels). We show the κ_m and δ_g maps on the left, and the corresponding sensitivity maps on the right. To suppress the noise, the lensing maps are additionally smoothed using a Gaussian kernel with $\sigma = 1$ pixel. Regions with higher absolute sensitivity contribute more to the decision of the network about the predicted value of A_{IA} and σ_8. Note the correspondence between the galaxy overdensity map and the lensing sensitivity map for bin 1; the areas of large overdensities are weighted higher in the κ_m map.

by overdensities; as seen in the bottom left-hand panel of Fig. 6, the dilution of the signal is the most significant. The CNNs automatically select the important regions and ignore the rest, which avoids signal dilution by unimportant regions.

While this interpretation is qualitative and subjective, it can already shed some initial light on the origin of the information used by CNNs. The impact of non-Gaussian information, access to phases of the map, and optimal map weighting, could all be quantitatively investigated further. However, we leave it to the dedicated future work.

VII. CONCLUSIONS

We present a novel approach to analysis of large-scale structure data by using full forward modeling on map level and its interpretation with deep learning. In a method dubbed DeepLSS, we create a combined probes analysis of weak lensing convergence κ_g and galaxy clustering maps δ_g. We focus on improving the constraints on the Ω_m and σ_8 parameters through breaking their degeneracies with intrinsic galaxy alignments and galaxy biasing, specifically, between S_8 and intrinsic alignment amplitude A_{IA}, linear bias b_g and matter density Ω_m, and linear b_g and nonlinear bias $b_{g,2}$. Internal degeneracies between these parameters and their redshift evolution cause further increase in marginalized parameter uncertainty.

We create consistent sets of simulations of galaxy clustering and weak lensing within the space of two models: (i) large smoothing and linear bias and (ii) small smoothing and nonlinear bias. These models have 7 and 9 free parameters, respectively.

As the most common way to analyze combined probes of LSS is to use 2-pt functions, we focus on a fair comparison between the deep-learning and the power spectrum methods. We create a set of residual convolutional neural networks on the individual probes κ_g and δ_g, and the probe combination $\kappa_g + \delta_g$. We pass the sets of tomographic maps to the networks as channels: 4 for individual probes and 8 for probe combination. The networks are trained using likelihood loss between outputs θ_p and the true input parameters θ_t. This way, the network creates informative output summaries corresponding to the model parameters. We interpret these summaries by performing conditional density estimation on the likelihood $p(\theta_p|\theta_t)$. The constraints are obtained using Bayesian analysis with a Markov chain Monte Carlo sampler.

We report a remarkable ability of deep learning to break degeneracies between the LSS parameters and cosmology, as compared to the power spectrum method. The most significant gain is for intrinsic galaxy alignments, where the S_8–A_{IA} correlation is effectively broken and the A_{IA} constraint improves by a factor of 8 times. Galaxy stochasticity constraints improve by a factor of 3 times. The improvement in the galaxy biasing sector is around 1.3 times. For the nonlinear model, the CNN effectively breaks the degeneracy between the linear and nonlinear bias parameters, b_g and $b_{g,2}$. We also observe a significant gain in the constraining power of the redshift evolution of intrinsic alignments $\eta_{A_{\mathrm{IA}}}$, where the gain is 1.4 times. Similarly, for z evolution of galaxy biasing, the CNN achieves 1.2 times improvement for η_{b_g} together with 1.7 times improvement in $\eta_{b_{g,2}}$. Overall, breaking these degeneracies leads to a very significant gain in constraints along the σ_8 and Ω_m degeneracy, with the figure of merit improved by 15 times.

We investigate the source of the information gained in the CNN analysis by looking at sensitivity maps. These maps show which regions of the input κ_g and δ_g maps have the most impact on the network's prediction. We focus on the sensitivity to A_{IA} and σ_8, which gain the most from the DeepLSS analysis. We notice that, for the A_{IA} parameter, the network draws most information from κ_g maps in the regions corresponding to high overdensities in the δ_g maps, while other pixels are heavily down-weighted. This picture suggests the following intuitive explanation. The highly-weighted regions are exactly where the IA signal comes from: galaxy shape alignment around overdense regions at the same redshift. Ignoring the rest of the map decreases the dilution of the signal by the noise coming from uninformative regions. The PSD method, on the other hand, does not perform such weighting and considers all pixels in the field equally, which leads to increased impact of the noise from regions that have no signal. As high density regions in low redshift maps are rare, the gain of the networks due to this effect becomes very significant. This interpretation, however, is probably not the full story: the PSD also ignores the phases of the maps, as well as all non-Gaussian information. Further work would be needed to gain more insight about the relative importance of these effects.

Previous works that apply deep learning to weak lensing mass maps predicted around 1.3–1.5 times precision gain for stage-III surveys. In this work, we find somewhat larger gains for weak lensing alone, and significantly larger gains for the probe combination. We note that the results of this and previous work are not directly comparable: here, we used a more complicated intrinsic alignment model, which included redshift evolution. We found that evolving IAs have a large impact on PSD results from lensing only. However, compared to previous work, the main difference is the combination of lensing and clustering, which enables CNNs to break the degeneracies between S_8 and intrinsic alignment.

This work serves as a demonstration of the constraining power of the DeepLSS method, and the exact gains are specific to the analysis configuration used here. Our theory and data models, while simpler than for a typical survey analysis, are highly realistic and contain most of the needed degrees of freedom. This suggests that the gains presented here should also be recovered by a LSS analysis with more precise theory and data modeling.

More development on the forward-modeling side is needed before practical deployment of this method. Firstly, the galaxy biasing prescriptions could be compared to the more advanced modeling using halo occupation distributions, subhalo abundance matching models, or using rapid halo and subhalo simulations with PINOCCHIO [55,56], and others [57–60]. Secondly, the baryonic effects [61] should be included in the forward model, similarly to Ref. [27]. Finally, given the large constraining power gain for intrinsic alignments, it may be feasible to constrain much more complex models with AI, such as tidal alignment tidal torque [14], or color-dependent IAs [62]. Conversely, one may imagine using the clustering maps with even larger smoothing scales for simple AI models. This would reduce the dependence of the forward model on details of galaxy clustering modeling, nonlinear effects, and systematics, while sill providing a large improvement on A_{IA}. Such large-scale-only model would be much easier to implement in practice.

A careful study of the impact of survey systematic effects, such as galaxy selection function uncertainty, redshift measurement errors, and shear calibration, should be performed before the practical application of DeepLSS. Their requirements may prove to be different for the AI analysis compared to the 2-pt analysis. For shear peak statistics, the DES Y3 analysis [25] revisited the shear calibration requirements; similar strategies can be used here. Conversely, the fidelity of the simulated maps to the observed maps should be assessed. For deep learning, intrinsic alignment model biases can have potentially much larger impact than for the 2-pt functions, which is already significant [15,19]. These biases will need to be carefully studied to make sure that the analysis is robust to the modeling choices.

The machine-learning methods used in this work were very simple and we did not optimize them for better information gain. We did, however, optimize the architectures of the networks interpreting the PSD vectors. It is therefore conceivable that the gain from CNNs can be further increased by using larger models and better optimization techniques available now in the field of machine learning.

In the near future, large light cone simulation grids will become more available. Along with the growth in processing capacities of hardware, the full forward model will become more practical and reproducible. Given that the advantage of this analysis is substantial and able to break degeneracies in the model, it is therefore possible that a map-level probe combination using deep learning may play an important role in future large-scale structure surveys.

The code is made publicly available at Ref. [63]. The data will be made public as a part of larger data release of CosmoGrid at Ref. [64]. In the meantime, the data can be made available upon request.

ACKNOWLEDGMENTS

T. K. thanks Javier Francisco Castander, Pablo Fosalba, Martin Crocce, and Enrique Gaztañaga for very helpful discussions. T. K. also thanks Marc Caubet Serrabou for excellent HPC support for the Merlin cluster at PSI, as well as Andreas Adelmann for access to the Gwendolen machine. We would like to thank Aurelien Lucchi for ongoing collaboration on deep learning for cosmology

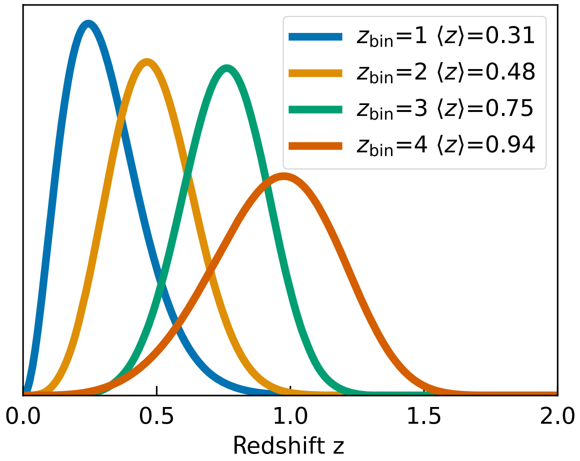

FIG. 7. Redshift bins used in this work.

and providing machine-learning expertise and advice. T. K. thanks David Bacon for support and useful discussions.

APPENDIX A: REDSHIFT BINS

In this work we use a generic stage-III simulated survey with four redshift bins. These bins have the mean redshifts of $\langle z\rangle = 0.31$, 0.48, 0.75, 0.94. The shape of the bins is shown in Fig. 7. The last bin is particularly wide, as is the case for photometric surveys; the uncertainty on the redshift for distant galaxies is high, which causes the bin to be broader.

APPENDIX B: MAPS OF GALAXY CLUSTERING WITH STOCHASTICITY

Galaxy stochasticity describes the degree of correlation between the galaxy density contrast δ_g and the underlying dark matter density field δ_m. The parameter r is the correlation coefficient between these fields, so that $r = \langle\delta_g\delta_m\rangle/\langle\delta_m\delta_m\rangle$. To create maps with varying degrees of correlation, we add random uniform noise to phases of the δ_m field, as described in Eq. (12). This equation uses a proxy parameter r_s, which is part of the model. We find the factor $(1 - r_s)^{2/3}$ empirically; we calculate the cross-correlations $\langle\delta_g\delta_m\rangle/\langle\delta_m\delta_m\rangle$ and find that this expression brings the relation $r_s/r \approx 1$, with deviations of 10%–20%. Figure 8 shows this correlation as a function of the parameter r_s, for four redshift bins. Each line is an average cross-correlation from 36 different 5×5 deg maps. The function in Eq. (12) could probably be improved, but it is sufficient for the purpose of this paper.

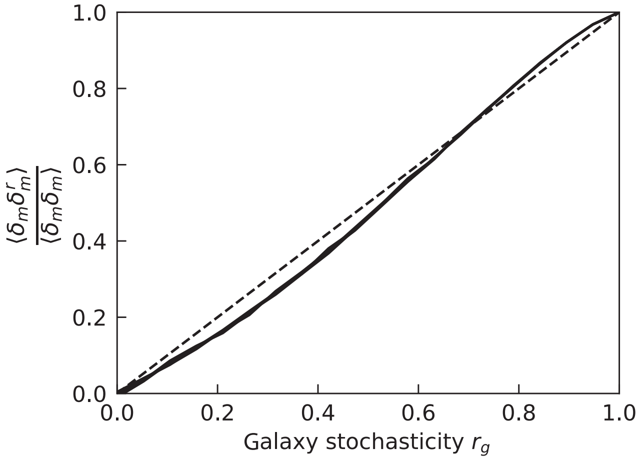

FIG. 8. Stochasticity parameter r_g used in the model and the corresponding cross-correlation between the galaxy density contrast δ_g and the underlying dark matter density field δ_m.

APPENDIX C: LIKELIHOOD MODELING

As introduced in Sec. IV, we model the likelihood of the predicted summary θ_p given true parameter value θ_t using a mixture density network. This network takes in values of θ_t and outputs the parameter of a Gaussian mixture model: means of components μ_j, their covariance matrices Ψ_j, and their relative weights w_j. This is a simple network with 3 layers and 256 hidden units each, and a Relu activation. The loss is the negative log-likelihood of the samples

$$L = -\log\mathcal{N}[\theta_p|\mu_{1,\ldots,K}(\theta_t), \Psi_{1,\ldots,K}(\theta_t), w_{1,\ldots,K}(\theta_t)],$$

where $\mathcal{N}$ is the normal distribution probability density function. We use $K = 4$ G in our model. As described in Sec. IV, we use a training set of 7 615 200 samples from $p(\theta_p|\theta_t)$, for each CNN and PSD network. We train these networks using the ADAM optimizer [46] with the initial learning rate of 0.001. We leave 25% of the training set as the validation set. Before passing them the neural networks, we rescale θ_t using the robust scaler and θ_p with minmax scaler [65] in the range $[10^{-5}, 1 - 10^{-5}]$, followed by a normal percentile function. These invertible transformations make it easier for the Gaussian mixture model to model this data. We use early stopping and pick the model with the best validation loss.

To validate the precision of the likelihood model, we perform the following test. First, we create a validation set of $\theta_p^{\rm val}$ at fixed $\theta_t^{\rm val}$, which contains 5700 parameter combinations. For each $\theta_t^{\rm val}$, we sample new predictions $\theta_s^{\rm val}$ from the MDN model. We compare these predictions with the corresponding summaries in the training set $\theta_p^{\rm val}$. For each θ_t, we calculate the significance of the difference of means of these samples Δ_μ, as shown in Eq. (C1). We also compare the scatter in the summaries in the samples predicted by the CNNs or PSDs and the MDN density estimator. To do this, we calculate the standard deviations for each parameter in the summary vector, and compare their fractional differences Δ_σ, shown in Eq. (C2):

$$\Delta_\mu = (\mathrm{mean}\theta_s^{\rm val} - \theta_p^{\rm val})/(\mathrm{SD}\theta_p^{\rm val}), \tag{C1}$$

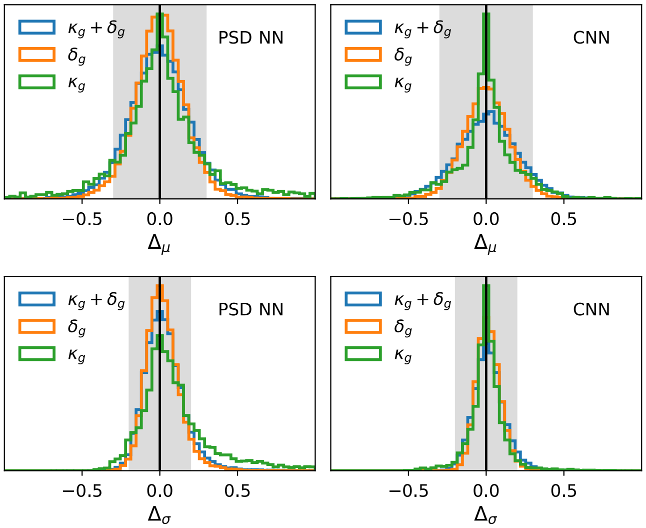

FIG. 9. Verification of the accuracy of the conditional density estimation of the likelihood $p(\theta_p|\theta_t)$ of predicted summary θ_p given true parameter value θ_t, modeled using a mixture density network (MDN). The upper panels show histograms of Δ_μ: fractional differences between the summaries predicted by the CNN and PSD NN models θ_p and samples θ_s from the estimated density $p(\theta_p|\theta_t)$, compared to the standard deviation of θ_p [see Eq. (C1)]. The shaded regions correspond to $<0.3\sigma$, which is the level of error that will not have significant impact on the results. The lower panels show the fractional difference Δ_σ between standard deviations of θ_p and θ_s, as defined in Eq. (C2), with shaded regions corresponding to $<0.2\sigma$. The histograms consist of 5700 parameter combinations from the prior space defined in Table I, concatenating all model parameters.

$$\Delta_\sigma = (\mathrm{SD}\theta_s^{\mathrm{val}} - \mathrm{SD}\theta_p^{\mathrm{val}})/(\mathrm{SD}\theta_p^{\mathrm{val}}). \qquad \text{(C2)}$$

We then plot the distribution of Δ_μ from all 5700 parameter sets and all 6 models, as shown in the upper panel of Fig. 9. To make this histogram, we use fractional differences for all parameters inside the θ vector. The MDN is unbiased and much smaller than the uncertainty of the summary, on the level of $<0.3\sigma$. The fact that the MDN is unbiased and with low scatter indicates that it estimates the conditional density of $p(\theta_p|\theta_t)$ sufficiently well. We plot the histogram of Δ_σ in the bottom panels of Fig. 9. The fractional difference distribution is, again, centered around zero, with differences on the level of 0.1σ The MDN modeling for κ_g is slightly worse than for other probes, most likely due to its comparatively worse overall constraining power, with posterior probability mass often hitting the prior boundaries. Given that both the central values of the summaries from the MDN model and their uncertainty are unbiased and with relatively low scatter, we consider this method to be sufficiency precise for the purpose of this paper.

Finally, we test if a more complicated mixture model would be needed. We run the fiducial combined probes analysis with number of Gaussians increased to $K = 8$ and the number of neural network hidden units to 512. We find no fundamental differences in the results, which suggests that there is no need to use a larger model. More advanced density estimation methods can be used for this step, such as Refs. [66–68]; we leave these improvements to future work.

APPENDIX D: NEURAL NETWORKS FOR PSD

In this work we calculate constraints from power spectra by compressing the PSD vectors into summaries using a simple neural network. This requires choosing an architecture and training of the PSD NN networks. These choices can affect the size of the final constraint. For example, a NN with a small number of neurons would not be able to capture all variation in the PSD vectors in the training set. A sufficiently large network would, however, be able to extract almost all the information from PSDs, as long as the number of network outputs is greater than or equal to the number of model parameters [69].

In order to have a fair comparison between the PSD and CNN results, we make sure that the PSD NN extract close to full information from the PSD vectors. For a single case of *CP* PSD NN, we perform a convergence test: we increase the number of trainable parameters in the networks to see if this makes a difference. We also test different network architectures to make sure that the results are not significantly affected by this choice. To test this, we create 5 alternatives to the main model described in Sec. III. We use three different architectures, and for each of them we try a standard and an extra-large (XL) version. The architectures are as follows.

(1) Classic NN (as in main model) with flattened PSD vector connected to 2 dense hidden layers with 1024 units, which then map to the outputs. Relu activation was used. The total number of trainable parameters was 1 823 779. The XL version had 3 hidden layers and 2 873 379 parameters.

(2) Locally connected convolutional neural net, which treated the PSD vectors as 1D channels. The channels were passed through locally connected, 1D convolutional layers [70]. This architecture has 128 filters with size of 3 and separated by stride of 2, for each part of the PSD vector. Here the filters are shared across channels. We use two such layers, followed by 4 residual layers, which are then fully connected to the output. The total number of trainable parameters was 734 883. The XL version had 256 filters, 8 residual layers, and 4 224 291 parameters.

(3) Separable convolutional neural net [71], which also used PSD vectors as 1D channels. This architecture uses a depthwise 1D convolution that acts separately on channels, followed by a pointwise convolution that mixes channels. We used 2 such layers, each with 128 filters, with kernel size of 3 and a stride of 2. They were followed by 4 residual layers, which are later flattened and fully connected to the output. This network has 438 415 trainable parameters. The

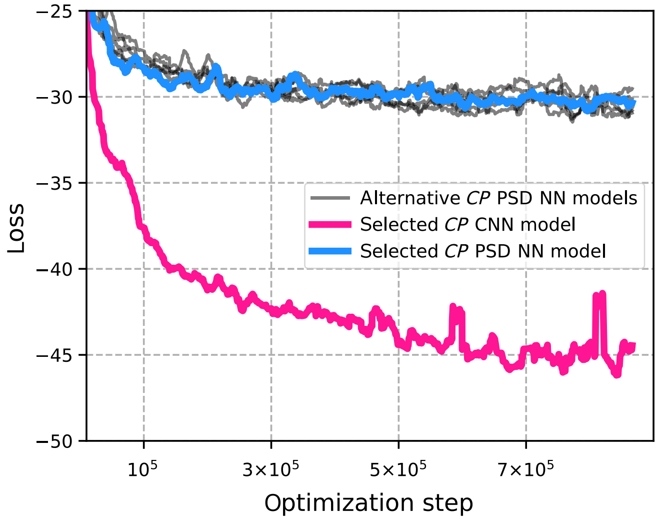

FIG. 10. Training loss during the optimization process for the main *CP* CNN (pink), the main *CP* PSD NN (blue), and all the alternative *CP* PSD NN networks. In this work we use likelihood loss, as defined in Eq. (14), which can have negative values.

XL version had 256 filters and 8 residual layers, which totaled to 3 270 799 parameters.

We train these networks for 885k steps with batch size of 32 and learning rate of 0.0025. The loss function progress during training is shown in Fig. 10. The alternative networks are shown in gray, while the main *CP* PSD NN are shown in blue. For comparison, the main deep-learning result, *CP* CNN, is shown in pink. We notice that the loss has not decreased much for the last 100k steps, which indicates that all networks are almost perfectly converged. There is some small variation in the values of the loss function at the end of training among the PSD networks. Importantly, the XL networks do not achieve significantly lower loss at the end of training than the main *CP* PSD NN. All the PSD networks, however, have much higher final loss than the deep-learning method *CP* CNN. This suggests that training the PSD networks for longer, or using larger models, would not improve the loss function significantly. We conclude that the main PSD NN used in this work effectively captures almost all the information contained in the PSD vectors, and can be used in fair comparison with the deep-learning method. This test was performed only for the combined probes, which is the most complex case; the conclusions of this test should also hold if repeated for the individual probes.

APPENDIX E: DATA TRANSFORMATIONS

To ensure that the results are numerically stable, we apply a number of transformations to the maps before passing them to the neural networks. All these operations are lossless and identical for both CNN and PSD. We bring the dynamic range of the maps close to the range between −1 and 1, which generally helps to improve the training: We scale the κ_g map by a factor of $(0.005^2 + \sigma^2_{\mathrm{pix}})^{-0.5}$, where σ^{pix} is the standard deviation of shape noise in a pixel. For δ_g, we use the transformation $\delta_g \leftarrow \delta_g / N_g^{\mathrm{pix}} - 1$, where N_g^{pix} is the average number of galaxies in a pixel calculated. We also make sure that the number of galaxies in the δ_g map is always ≥0 to prevent negative Poisson λ parameters, that can very rarely occur. We do this by rescaling the density contrast δ_m such that the lowest pixel value is −1, but the total number of galaxies is preserved. To do this, we first clip δ_m to the value of −1 and create δ_m^{clip}, and the final map is $\delta_m \leftarrow \delta_m^{\mathrm{clip}} \sum \delta_m / (\sum \delta_m^{\mathrm{clip}})$. We also transform the output parameters θ by using S_8 instead of σ_8, and multiplying the A_{IA} by a factor of 0.1. This operation is reversed at prediction step.

APPENDIX F: APPROXIMATE POISSON NOISE

As currently there is no Poisson noise generator available on GPU in TensorFlow, we use the inverse Anscombe transform [41] to approximate it. We sample random Poisson number z with mean λ using this equation:

$$z \sim \mathrm{Normal}\left[2\sqrt{\lambda + \frac{3}{8}} - \frac{1}{4}\lambda^{-1/2}, 1\right], \tag{F1}$$

$$z \leftarrow \begin{cases} 0 & z < 1 \\ \frac{1}{4}z^2 - \frac{1}{8} + \frac{1}{4}\sqrt{\frac{3}{2}}z^{-1} - \frac{11}{8}z^{-2} + \frac{5}{8}\sqrt{\frac{3}{2}}z^{-3} & z \geq 1, \end{cases} \tag{F2}$$

$$z \leftarrow \mathrm{round}[\max[z, 0]]. \tag{F3}$$

We verify that this approximation is very close to Poisson and should be sufficiently precise for the purpose of this work.

[1] M. Kilbinger, *Cosmology with Cosmic Shear Observations: A Review*, Rep. Prog. Phys. **78**, 086901 (2015).
[2] H. Zhan and J. A. Tyson, *Cosmology with the Large Synoptic Survey Telescope: An Overview*, Rep. Prog. Phys. **81**, 066901 (2018).
[3] A. Albrecht, G. Bernstein, R. Cahn, W. L. Freedman, J. Hewitt, W. Hu, J. Huth, M. Kamionkowski, E. W. Kolb, L. Knox, J. C. Mather, S. Staggs, and N. B. Suntzeff, *Report of the Dark Energy Task Force*, arXiv:astro-ph/0609591.
[4] https://www.darkenergysurvey.org.
[5] https://kids.strw.leidenuniv.nl.
[6] https://hsc.mtk.nao.ac.jp/ssp/survey.
[7] T. M. C. Abbott *et al.* (DES Collaboration), *Dark Energy Survey Year 3 Results: Cosmological Constraints from Galaxy Clustering and Weak Lensing*, Phys. Rev. D **105**, 023520 (2022).
[8] C. Heymans, T. Tröster, M. Asgari, C. Blake, H. Hildebrandt, B. Joachimi, K. Kuijken, C.-A. Lin, A. G.

Sánchez, J. L. van den Busch *et al.*, *KiDS-1000 Cosmology: Multi-Probe Weak Gravitational Lensing and Spectroscopic Galaxy Clustering Constraints*, Astron. Astrophys. **646**, A140 (2021).

[9] C. Hikage, M. Oguri, T. Hamana, S. More, R. Mandelbaum, M. Takada, F. Köhlinger, H. Miyatake, A. J. Nishizawa, H. Aihara *et al.*, *Cosmology from Cosmic Shear Power Spectra with Subaru Hyper Suprime-Cam First-Year Data*, Publ. Astron. Soc. Jpn. **71**, 43 (2019).

[10] https://www.euclid-ec.org.

[11] https://lsstdesc.org.

[12] D. Kirk, M. L. Brown, H. Hoekstra, B. Joachimi, T. D. Kitching, R. Mandelbaum, C. Sifón, M. Cacciato, A. Choi, A. Kiessling, A. Leonard, A. Rassat, and B. M. Schäfer, *Galaxy Alignments: Observations and Impact on Cosmology*, Space Sci. Rev. **193**, 139 (2015).

[13] M. Eriksen and E. Gaztañaga, *Combining Spectroscopic and Photometric Surveys Using Angular Cross-Correlations—III. Galaxy Bias and Stochasticity*, Mon. Not. R. Astron. Soc. **480**, 5226 (2018).

[14] J. A. Blazek, N. MacCrann, M. A. Troxel, and X. Fang, *Beyond Linear Galaxy Alignments*, Phys. Rev. D **100**, 103506 (2019).

[15] L. F. Secco *et al.* (DES Collaboration), *Dark Energy Survey Year 3 Results: Cosmology from Cosmic Shear and Robustness to Modeling Uncertainty*, Phys. Rev. D **105**, 023515 (2022).

[16] A. Porredon, M. Crocce, J. Elvin-Poole, R. Cawthon, G. Giannini, J. De Vicente, A. Carnero Rosell, I. Ferrero, E. Krause, X. Fang *et al.*, *Dark Energy Survey Year 3 Results: Cosmological Constraints from Galaxy Clustering and Galaxy-Galaxy Lensing Using the MagLim Lens Sample*, arXiv:2105.13546.

[17] T. Eifler, E. Krause, P. Schneider, and K. Honscheid, *Combining Probes of Large-Scale Structure with COSMOLIKE*, Mon. Not. R. Astron. Soc. **440**, 1379 (2014).

[18] E. Krause, X. Fang, S. Pandey, L. F. Secco, O. Alves, H. Huang, J. Blazek, J. Prat, J. Zuntz, T. F. Eifler *et al.*, *Dark Energy Survey Year 3 Results: Multi-Probe Modeling Strategy and Validation*, arXiv:2105.13548.

[19] D. Kirk, A. Rassat, O. Host, and S. Bridle, *The Cosmological Impact of Intrinsic Alignment Model Choice for Cosmic Shear*, Mon. Not. R. Astron. Soc. **424**, 1647 (2012).

[20] P. R. V. Chintalapati, G. Gutierrez, and M. H. L. S. Wang, *Systematic Study of Projection Biases in Weak Lensing Analysis*, Phys. Rev. D **105**, 043515 (2022).

[21] A. Amon, N. C. Robertson, H. Miyatake, C. Heymans, M. White, J. DeRose, S. Yuan, R. H. Wechsler, T. N. Varga, S. Bocquet *et al.*, *Consistent Lensing and Clustering in a Low-S_8 Universe with BOSS, DES Year 3, HSC Year 1 and KiDS-1000*, arXiv:2202.07440.

[22] A. Leauthaud, S. Saito, S. Hilbert, A. Barreira, S. More, M. White, S. Alam, P. Behroozi, K. Bundy, J. Coupon *et al.*, *Lensing is Low: Cosmology, Galaxy Formation or New Physics?*, Mon. Not. R. Astron. Soc. **467**, 3024 (2017).

[23] A. Leauthaud, A. Amon, S. Singh, D. Gruen, J. U. Lange, S. Huang, N. C. Robertson, T. N. Varga, Y. Luo, C. Heymans *et al.*, *Lensing without Borders—I. A Blind Comparison of the Amplitude of Galaxy-Galaxy Lensing between Independent Imaging surveys*, Mon. Not. R. Astron. Soc. **510**, 6150 (2022).

[24] E. Abdalla, G. F. Abellán, A. Aboubrahim, A. Agnello, O. Akarsu, Y. Akrami, G. Alestas, D. Aloni, L. Amendola, and L. A. Anchordoqui *et al.*, *Cosmology Intertwined: A Review of the Particle Physics, Astrophysics, and Cosmology Associated with the Cosmological Tensions and Anomalies*, J. High Energy Astrophys. **34**, 49 (2022).

[25] D. Zürcher *et al.* (DES Collaboration), *Dark Energy Survey Year 3 Results: Cosmology with Peaks Using an Emulator Qpproach*, Mon. Not. R. Astron. Soc. **511**, 2075 (2022).

[26] J. Fluri, T. Kacprzak, A. Lucchi, A. Refregier, A. Amara, T. Hofmann, and A. Schneider, *Cosmological Constraints with Deep Learning from KiDS-450 Weak Lensing Maps*, Phys. Rev. D **100**, 063514 (2019).

[27] J. Fluri, T. Kacprzak, A. Lucchi, A. Schneider, A. Refregier, and T. Hofmann, *A Full w*CDM *Analysis of KiDS-1000 Weak Lensing Maps Using Deep Learning*, Phys. Rev. D **105**, 083518 (2022).

[28] D. Gruen *et al.* (DES Collaboration), *Density Split Statistics: Cosmological Constraints from Counts and Lensing in Cells in DES Y1 and SDSS Cata*, Phys. Rev. D **98**, 023507 (2018).

[29] A. I. Salvador *et al.* (DES Collaboration), *Measuring Linear and Non-Linear Galaxy Bias Using Counts-in-Cells in the Dark Energy Survey Science Verification Data*, Mon. Not. R. Astron. Soc. **482**, 1435 (2019).

[30] C. M. Hirata and U. Seljak, *Intrinsic Alignment-Lensing Interference as a Contaminant of Cosmic Shear*, Phys. Rev. D **70**, 063526 (2004).

[31] S. Bridle and L. King, *Dark Energy Constraints from Cosmic Shear Power Spectra: Impact of Intrinsic Alignments on Photometric Redshift Requirements*, New J. Phys. **9**, 444 (2007).

[32] B. Joachimi, R. Mandelbaum, F. B. Abdalla, and S. L. Bridle, *Constraints on Intrinsic Alignment Contamination of Weak Lensing Surveys Using the MegaZ-LRG Sample*, Astron. Astrophys. **527**, A26 (2011).

[33] N. Aghanim *et al.* (Planck Collaboration), *Planck 2018 Results VI. Cosmological Parameters*, Astron. Astrophys. **641**, A6 (2020).

[34] M. Rodríguez-Monroy *et al.* (DES Collaboration), *Dark Energy Survey Year 3 Results: Galaxy Clustering and Systematics Treatment for Lens Galaxy Samples*, Mon. Not. R. Astron. Soc. **511**, 2665 (2022).

[35] A. Porredon *et al.* (DES Collaboration), *Dark Energy Survey Year 3 Results: Optimizing the Lens Sample in a Combined Galaxy Clustering and Galaxy-Galaxy Lensing Analysis*, Phys. Rev. D **103**, 043503 (2021).

[36] D. Potter, J. Stadel, and R. Teyssier, *PkdGrav3: Beyond Trillion Particle Cosmological Simulations for the Next Era of Galaxy Surveys*, Comput. Astrophys. Cosmol. **4**, 2 (2017).

[37] O. Hahn and T. Abel, *Multi-Scale Initial Conditions for Cosmological Simulations*, Mon. Not. R. Astron. Soc. **415**, 2101 (2011).

[38] R. Sgier, A. Réfrégier, A. Amara, and A. Nicola, *Fast Generation of Covariance Matrices for Weak Lensing*, J. Cosmol. Astropart. Phys. 01 (2019) 044.

[39] D. Zürcher, J. Fluri, R. Sgier, T. Kacprzak, and A. Refregier, *Cosmological Forecast for Non-Gaussian Statistics in Large-Scale Weak Lensing Surveys*, J. Cosmol. Astropart. Phys. 01 (2021) 028.

[40] O. Friedrich *et al.* (DES Collaboration), *Density Split Statistics: Joint Model of Counts and Lensing in Cells*, Phys. Rev. D **98**, 023508 (2018).

[41] M. Makitalo and A. Foi, *A Closed-Form Approximation of the Exact Unbiased Inverse of the Anscombe Variance-Stabilizing Transformation*, IEEE Trans. Image Process. **20**, 2697 (2011).

[42] A. Schneider, N. Stoira, A. Refregier, A. J. Weiss, M. Knabenhans, J. Stadel, and R. Teyssier, *Baryonic Effects for Weak Lensing. Part I. Power Spectrum and Covariance Matrix*, J. Cosmol. Astropart. Phys. 04 (2020) 019.

[43] A. J. Weiss, A. Schneider, R. Sgier, T. Kacprzak, A. Amara, and A. Refregier, *Effects of baryons on weak lensing peak statistics*, J. Cosmol. Astropart. Phys. 10 (2019) 011.

[44] K. He, X. Zhang, S. Ren, and J. Sun, *Deep Residual Learning for Image Recognition*, in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, New York, 2016), p. 770.

[45] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016).

[46] D. P. Kingma and J. L. Ba, *Adam: A Method for Stochastic Optimization*, arXiv:1412.6980.

[47] P. Seetharaman, G. Wichern, B. Pardo, and J. Le Roux, *Autoclip: Adaptive Gradient Clipping for Source Separation Networks*, in *Proceedings of the 2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)* (IEEE, New York, 2020), pp. 1–6.

[48] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*, https://tensorflow.org.

[49] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, *EMCEE: The MCMC Hammer*, Publ. Astron. Soc. Pac. **125**, 306 (2013).

[50] J. M. Kratochvil, E. A. Lim, S. Wang, Z. Haiman, M. May, and K. Huffenberger, *Probing Cosmology with Weak Lensing Minkowski Functionals*, Phys. Rev. D **85**, 103513 (2012).

[51] J. Harnois-Déraps, N. Martinet, T. Castro, K. Dolag, B. Giblin, C. Heymans, H. Hildebrandt, and Q. Xia, *Cosmic Shear Cosmology Beyond Two-Point Statistics: A Combined Peak Count and Correlation Function Analysis of DES-Y1*, Mon. Not. R. Astron. Soc. **506**, 1623 (2021).

[52] Jose Manuel Zorrilla Matilla, M. Sharma, D. Hsu, and Z. Haiman, *Interpreting Deep Learning Models for Weak Lensing*, Phys. Rev. D **102**, 123506 (2020).

[53] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev, *The Building Blocks of Interpretability, Distill*, https://distill.pub/2018/building-blocks.

[54] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Mller, *Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications*, Proc. IEEE **109**, 247 (2021).

[55] P. Berner, A. Refregier, R. Sgier, T. Kacprzak, L. Tortorelli, and P. Monaco, *Rapid Simulations of Halo and Subhalo Clustering*, arXiv:2112.08389.

[56] G. Taffoni, P. Monaco, and T. Theuns, *PINOCCHIO and the Hierarchical Build-Up of Dark Matter Haloes*, Mon. Not. R. Astron. Soc. **333**, 623 (2002).

[57] J. DeRose, R. H. Wechsler, M. R. Becker, M. T. Busha, E. S. Rykoff, N. MacCrann, B. Erickson, A. E. Evrard, A. Kravtsov, D. Gruen *et al.*, *The Buzzard Flock: Dark Energy Survey Synthetic Sky Catalogs*, arXiv:1901.02401.

[58] P. Fosalba, E. Gaztañaga, F. J. Castander, and M. Crocce, *The MICE Grand Challenge Light-Cone Simulation—III. Galaxy Lensing Mocks from All-Sky Lensing Maps*, Mon. Not. R. Astron. Soc. **447**, 1319 (2015).

[59] O. Friedrich, A. Halder, A. Boyle, C. Uhlemann, D. Britt, S. Codis, D. Gruen, and C. Hahn, *The PDF Perspective on the Tracer-Matter Connection: Lagrangian Bias and Non-Poissonian Shot Noise*, Mon. Not. R. Astron. Soc. **510**, 5069 (2022).

[60] J. Harnois-Déraps, A. Amon, A. Choi, V. Demchenko, C. Heymans, A. Kannawadi, R. Nakajima, E. Sirks, L. van Waerbeke, Y.-C. Cai, B. Giblin, H. Hildebrandt, H. Hoekstra, L. Miller, and T. Tröster, *Cosmological Simulations for Combined-Probe Analyses: Covariance and Neighbour-Exclusion Bias*, Mon. Not. R. Astron. Soc. **481**, 1337 (2018).

[61] A. Schneider, A. Refregier, S. Grandis, D. Eckert, N. Stoira, T. Kacprzak, M. Knabenhans, J. Stadel, and R. Teyssier, *Baryonic Effects for Weak Lensing. Part II. Combination with X-Ray Data and Extended Cosmologies*, J. Cosmol. Astropart. Phys. 04 (2020) 020.

[62] S. Samuroff *et al.* (DES Collaboration), *Dark Energy Survey Year 1 Results: Constraints on Intrinsic Alignments and Their Colour Dependence from Galaxy Clustering and Weak Lensing*, Mon. Not. R. Astron. Soc. **489**, 5453 (2019).

[63] https://github.com/tomaszkacprzak/DeepLSS.

[64] https://www.cosmogrid.ai.

[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-Learn: Machine Learning in PYTHON*, J. Mach. Learn. Res. **12**, 2825 (2011), https://arxiv.org/abs/1201.0490.

[66] Q. Liu, J. Xu, R. Jiang, and W. H. Wong, *Density Estimation Using Deep Generative Neural Networks*, Proc. Natl. Acad. Sci. U.S.A. **118**, e2101344118 (2021).

[67] V. Dutordoir, H. Salimbeni, M. P. Deisenroth, and J. Hensman, *Gaussian Process Conditional Density Estimation*, in *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)* (Curran Associates Inc., Red Hook, NY, 2018), pp. 2391–2401.

[68] G. Papamakarios, T. Pavlakou, and I. Murray, *Masked Autoregressive Flow for Density Estimation*, Adv. Neural Inf. Process. Syst. **30** (2017), https://arxiv.org/abs/1705.07057.

[69] A. Heavens, E. Sellentin, D. de Mijolla, and A. Vianello, *Massive Data Compression for Parameter-Dependent Covariance Matrices*, Mon. Not. R. Astron. Soc. **472**, 4244 (2017).

[70] Yu-hsin Chen, I. L. Moreno, T. Sainath, M. Visontai, R. Alvarez, and C. Parada, *Locally-Connected and Convolutional Neural Networks for Small Footprint Speaker Recognition*, in Interspeech (2015), https://research.google/pubs/pub43970.

[71] L. Kaiser, A. N. Gomez, and F. Chollet, *Depthwise Separable Convolutions for Neural Machine Translation*, in *Proceedings of the International Conference on Learning Representations*, 2017, https://arxiv.org/abs/1706.03059.