The Astrophysical Journal, 941:132 (12pp), 2022 December 20 https://doi.org/10.3847/1538-4357/ac9f14

Inpainting Hydrodynamical Maps with Deep Learning

Faizan G. Mohammad[1,2], Francisco Villaescusa-Navarro[3,4], Shy Genel[4,5], Daniel Anglés-Alcázar[4,6], and Mark Vogelsberger[7]
[1] Waterloo Center for Astrophysics, University of Waterloo, Waterloo, ON N2L 3G1, Canada; sgenel@flatironinstitute.org
[2] Department of Physics and Astronomy, University of Waterloo, Waterloo, ON N2L 3G1, Canada
[3] Department of Astrophysical Sciences, Princeton University, Peyton Hall, Princeton, NJ 08544-0010, USA
[4] Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, 10010, New York, NY, USA
[5] Columbia Astrophysics Laboratory, Columbia University, 550 West 120th Street, New York, NY 10027, USA
[6] Department of Physics, University of Connecticut, 196 Auditorium Road, U-3046, Storrs, CT 06269, USA
[7] Kavli Institute for Astrophysics and Space Research, Department of Physics, MIT, Cambridge, MA 02139, USA
Received 2021 September 14; revised 2022 October 27; accepted 2022 October 30; published 2022 December 16

Abstract

From 1000 hydrodynamic simulations of the CAMELS project, each with a different value of the cosmological and astrophysical parameters, we generate 15,000 gas temperature maps. We use a state-of-the-art deep convolutional neural network to recover missing data from those maps. We mimic the missing data by applying regular and irregular binary masks that cover either 15% or 30% of the area. We quantify the reliability of our results using two summary statistics: (1) the distance between the probability density functions, estimated using the Kolmogorov–Smirnov (K-S) test, and (2) the 2D power spectrum. We find an excellent agreement between the model prediction and the unmasked maps when using the power spectrum: better than 1% for $k < 20\,h\,\mathrm{Mpc}^{-1}$ for any irregular mask. For regular masks, we observe a systematic offset of ∼5% when covering 15% of the maps, while the results become unreliable when 30% of the data is missing. The observed K-S test p-values favor the null hypothesis that the reconstructed and the ground-truth maps are drawn from the same underlying distribution when irregular masks are used. For regular-shaped masks, on the other hand, we find a strong evidence that the two distributions do not match each other. Finally, we use the model, trained on gas temperature maps, to inpaint maps from fields not used during model training. We find that, visually, our model is able to reconstruct the missing pixels from the maps of those fields with great accuracy, although its performance using summary statistics depends strongly on the considered field.

Unified Astronomy Thesaurus concepts: Large-scale structure of the universe (902); Computational methods (1965); Astrostatistics (1882)

1. Introduction

Cosmology is in a transformative stage. Nowadays, we know the value of the main cosmological parameters with a relatively high precision. This has allowed us to claim, with high confidence, the existence of a substance that is responsible for the accelerated expansion of the universe: dark energy. The nature and properties of dark energy remain the biggest mysteries in modern physics. In order to shed light on these and other open questions, such as the sum of the neutrino masses, the community has spent billions of dollars on surveys like the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration et al. 2016), Euclid (Laureijs et al. 2011), Prime Focus Spectrograph (PFS; Tamura et al. 2016), extended ROentgen Survey with an Imaging Telescope Array (eROSITA; Merloni et al. 2012), Roman Observatory (Spergel et al. 2015), Rubin Observatory (The LSST Dark Energy Science Collaboration et al 2018), Square Kilometer Array (SKA; Square Kilometre Array Cosmology Science Working Group et al. 2020), and Simons Observatory (Ade et al. 2019), whose data may contain the answers to all these fundamental questions.

The traditional method used to transform the data from cosmological surveys into constraints is this: (1) the data is compressed into a lower-dimensional summary statistic, (2) theoretical predictions for that summary statistic are provided as a function of the value of the cosmological parameters, and (3) a likelihood function is evaluated to find the parameter constraints. Currently, there is a large debate on what summary statistics should be employed to extract the maximum information from these surveys (e.g., Allys et al. 2020; Banerjee et al. 2020; Dai et al. 2020; de la Bella et al. 2021; Friedrich et al. 2020; Giri & Smith 2022; Hahn et al. 2020; Uhlemann et al. 2020; Villaescusa-Navarro et al. 2020; Banerjee & Abel 2021a, 2021b; Bayer et al. 2021; Gualdi et al. 2021a, 2021b; Hahn & Villaescusa-Navarro 2021; Kuruvilla & Aghanim 2021; Massara et al. 2021; Samushia et al. 2021; Valogiannis & Dvorkin 2022). Another possibility is to extract information from the field itself without relying on summary statistics, using machine-learning methods (Ravanbakhsh et al. 2017; Schmelzle et al. 2017; Gupta et al. 2018; Fluri et al. 2019; Ntampaka et al. 2020; Ribli et al. 2019; Hassan et al. 2020; Jeffrey et al. 2021; Zorrilla Matilla et al. 2020).

Unfortunately, the data from the cosmic surveys are affected by numerous issues, such as instrument noise. Among these problems, there are some effects that can induce spatial discontinuities in the data. For instance, in the case of galaxy redshift surveys, the presence of stars, fiber collisions, and bad observations will create masks in the survey geometry (Ross et al. 2012; de la Torre et al. 2013; Bianchi & Verde 2020; Mohammad et al. 2020). Another example is when such masks are created to avoid contamination by systematic effects; e.g., Cosmic Microwave Background (CMB) and 21 cm observations

may be masked near the galactic plane to avoid the bright foregrounds.

In general, the complicated geometry induced by these masked regions represents a challenge for both the theoretical predictions and the computation of the (optimal) summary statistic. This problem may also get worse when working at the field level with machine-learning methods, as one needs to make sure that no information from the mask itself is used by the network.

One potential solution to this problem will be to reconstruct the missing data within the masked region. In most of the cases, however, this is a very difficult task, as the clustering properties of the considered field (e.g., galaxy redshift surveys or 21 cm surveys) are not well-understood theoretically (see discussion about summary statistics above). On the other hand, the statistical properties of the considered field can be learned by neural networks and used to reconstruct the masked region. This idea has been developed in the machine-learning community to *inpaint* the missing pixels of images (Pathak et al. 2016; Yang et al. 2016; Demir & Unal 2018; Liu et al. 2018; Yan et al. 2018; Yu et al. 2018, 2019; Nazeri et al. 2019; Zhu et al. 2021).

The use of image inpainting techniques based on deep learning has recently gained increasing interest in the cosmological community. Several works have successfully used deep convolutional neural networks to reconstruct missing data in 2D maps of the cosmic microwave background (Raghunathan et al. 2019; Yi et al. 2020; Montefalcone et al. 2021; Vafaei Sadr & Farsian 2021) and in the galactic foreground intensity and polarization maps (Puglisi & Bai 2020).

In this work, we use these techniques to investigate whether we can reconstruct masked regions from 2D images generated from state-of-the-art magnetohydrodynamic simulations. For this, we make use of data from the Cosmology and Astrophysics with MachinE Learning Simulations (CAMELS; Villaescusa-Navarro et al. 2021) Multifield Data set (CMD; Villaescusa-Navarro et al. 2022b), a collection of hundreds of thousands of 2D maps and 3D grids containing 13 different fields from thousands of different cosmological and astrophysical models. To our knowledge, this is the first time that such a study has been carried out with data from state-of-the-art hydrodynamic simulations over a vast range of cosmological and astrophysical models.

This paper is organized as follows. In Section 2, we describe the data we use in this work. We outline the architecture and training procedure in Section 3. The main results of this work are shown in Section 4. Finally, we summarize and discuss the findings of this paper in Section 5.

2. Data

In this work, we make use of 2D maps from the CAMELS Multifield Data set,[8] CMD, a collection of hundreds of thousands of 2D maps showing different properties of the gas, dark matter, and stars at $z = 0$ from 2000 state-of-the-art (magneto-)hydrodynamic simulations of the CAMELS project (Villaescusa-Navarro et al. 2021, 2022a). All simulations follow the evolution of 256^3 dark matter particles and 256^3 fluid elements from $z = 127$ down to $z = 0$ in a periodic comoving volume of $(25\ h^{-1}\ \mathrm{Mpc})^3$. Half of the hydrodynamic simulations have been run with the AREPO code (Weinberger et al. 2019) and employ the same subgrid model as the IllustrisTNG simulations (Weinberger et al. 2017; Pillepich et al. 2018), while the other half have been run with the GIZMO code (Hopkins 2015) and utilize the subgrid model of the SIMBA simulation (Davé et al. 2019).

All simulations share the same values of these cosmological parameters: baryon density $\Omega_\mathrm{b} = 0.049$, Hubble parameter $h = 0.67$, spectral index $n_s = 0.96$, sum of the neutrino mass $\sum m_\nu = 0\ \mathrm{eV}$, and the equation-of-state parameter for dark energy $w = -1$. On the other hand, each simulation has a different value for the total matter density parameter Ω_m and σ_8, the amplitude of the linear power spectrum on scales of $8\ h^{-1}\ \mathrm{Mpc}$, and they also differ in the values of four astrophysical parameters that characterize the efficiency of supernova and active galactic nuclei feedback. CMD contains maps for 13 different fields: (1) gas density, (2) gas velocity, (3) gas temperature, (4) gas pressure, (5) gas metallicity, (6) neutral hydrogen density, (7) electron number density, (8) magnetic fields, (9) magnesium-to-iron ratio, (10) dark matter density, (11) dark matter velocity, (12) stellar mass density, and (13) total matter density. Each 2D map covers an area of $25 \times 25\ (h^{-1}\ \mathrm{Mpc})^2$, contains 256×256 pixels, and has a specific value of the cosmological and astrophysical parameters. For each field, CMD provides 15,000 maps. We refer the reader to Villaescusa-Navarro et al. (2022b) for further details on CMD.

In this work, we focus our attention on the gas temperature maps, which represent the mass-weighted temperature field of the gas particles in the different simulations.

3. Technique

In this section, we describe the method used to evaluate the performance of the inpainting model. We start in Section 3.1 by describing the construction of the binary masks that we later apply to the CMD maps in order to mimic the missing data. In Section 3.2, we present the architecture of the deep convolutional neural network used to inpaint the masked regions in the data. In Section 3.3, we discuss the loss function used to train the neural network, while the training process is described in Section 3.4.

3.1. Binary Masks

We generate two types of masks: (1) *regular masks* that have either a rectangular or circular shape and cover a continuous portion of the field of view, and (2) *irregular masks* that consist of a set of segments of different width and length randomly placed over the field. For each of these two types, we build masks that cover different fractions of the total area. In particular, we use masks, both regular and irregular, that cover 15% and 30% of the total area. These are realistic numbers that one may encounter in galaxy redshift surveys. In particular, in the Dark Energy Survey (DES) photometric sample for cosmology (Sevilla-Noarbe et al. 2021), the masked regions amount to roughly 10% of the total survey area. In spectroscopic surveys, such as the Baryon Oscillation Spectroscopic Survey (BOSS) LOWZ and CMASS (Dawson et al. 2013) samples, ∼7% of the total area is lost due to the veto masks. In the extended Baryon Oscillation Spectroscopic Survey (eBOSS), ∼17% of the area covered by the Luminous Red Galaxy (LRG) and quasar (QSO) (Ross et al. 2020) catalogs

[8] https://camels-multifield-dataset.readthedocs.io

was obscured by different types of veto masks. The choice for the sizes of regular masks is straightforward, given the desired fraction of the area to be masked. Each irregular mask, on the other hand, is built by successively adding segments of randomly chosen width and length until the number of pixels they cover is the target fraction of the total area. In this paper, we will refer to the pixels covered by the mask as the *"hole pixels"* and to the unmasked pixels as *"valid pixels"*.

3.2. Architecture

We use the network architecture presented in Zhu et al. (2021) based on the so-called "Mask-Aware Dynamic Filtering" (MADF) module. This is a deep convolutional neural network consisting of three main stages: the encoder, the recovery decoder, and the refinement decoder. The architecture is similar in nature to a U-shaped encoder-decoder network that encodes the semantic information from the valid pixels of the masked image into multiple-level feature maps, which are later decoded into the low-level pixel values.

The encoder provides the high-level feature maps using the information from the input damaged image and the corresponding binary mask. In particular, rather than using fixed kernels, it uses the MADF module to dynamically generate kernels for each convolutional window based on the features of the corresponding position on the mask. The decoder step is further divided into two stages. The recovery decoder performs a rough filling of the holes in the feature maps and produces the first output. A set of refinement decoders are run in parallel to the recovery decoder, to refine the decoded feature maps. Another distinct feature of this novel network architecture is the use of the so-called "Point-wise Normalization" (PN) in place of the typical "Batch Normalization" (BN) in the refinement decoding steps, to avoid the "covariant shift" problem arising from the difference between the statistical properties of the features of the hole and valid pixels. We refer the reader to Zhu et al. (2021) for a detailed discussion of the advantages of this approach.

Although the architecture proposed in Zhu et al. (2021) is flexible in terms of the model complexity, tuning its hyperparameters would require many tests that are computationally expensive and time-demanding. We thus use the same setup proposed in Zhu et al. (2021) that resulted in excellent results on the benchmark data sets typically used to assess the performance of the image inpainting models. In particular, each of the encoder, recovery decoder, and refinement decoders consists of seven levels, with the kernel size and strides of each convolutional operation set empirically. Also the number of refinement decoders is set to be two, as a compromise between model performance and efficiency.

3.3. Loss Function

We use the "inpainting loss" adopted in Liu et al. (2018) and Zhu et al. (2021) as the optimization objective. The total loss function consists of multiple terms that depend on the output of each decoder and are incrementally added. Different loss terms compare different properties of the predicted and the true maps (ground truth).

The first-order comparison is performed using the so-called "per-pixel reconstruction loss" that is split into two terms, one evaluated over the valid pixels ($L_{\rm valid}$) and one over the hole pixels ($L_{\rm hole}$):

$$L_{\rm valid} = \frac{1}{N_{\boldsymbol{I}_{\rm gt}}} \left\| M \odot (\boldsymbol{I}_{\rm out} - \boldsymbol{I}_{\rm gt}) \right\|_1, \tag{1}$$

$$L_{\rm hole} = \frac{1}{N_{\boldsymbol{I}_{\rm gt}}} \left\| (1 - M) \odot (\boldsymbol{I}_{\rm out} - \boldsymbol{I}_{\rm gt}) \right\|_1. \tag{2}$$

In Equations (1) and (2), $N_{\boldsymbol{I}_{\rm gt}}$ indicates the number of elements in the ground-truth map, M is the binary mask, $\boldsymbol{I}_{\rm out}$ is the model output, $\boldsymbol{I}_{\rm gt}$ is the ground-truth image, and $\odot$ denotes the element-wise product.

While the terms introduced in the previous paragraph encourage accurate per-pixel predictions, they do not account explicitly for larger-scale structures in the contents of the images. The perceptual loss $L_{\rm perc}$, introduced by Gatys et al. (2015), forces the network to output semantically meaningful predictions as encoded by the feature maps Ψ_p extracted using the *pool1*, *pool2*, and *pool3* layers of the pretrained VGG16 ImageNet (Simonyan & Zisserman 2014). By encouraging high accuracy in several pooling layers, the perceptual loss assists in the prediction of structures of general nature on various scales, which is expected to be beneficial for a complex structure like the cosmic web. This perceptual loss is given by

$$L_{\rm perc} = \sum_{p=1}^{3} \frac{\left\| \Psi_p^{\boldsymbol{I}_{\rm out}} - \Psi_p^{\boldsymbol{I}_{\rm gt}} \right\|_1}{N_{\Psi_p^{\boldsymbol{I}_{\rm gt}}}} + \sum_{p=1}^{3} \frac{\left\| \Psi_p^{\boldsymbol{I}_{\rm comp}} - \Psi_p^{\boldsymbol{I}_{\rm gt}} \right\|_1}{N_{\Psi_p^{\boldsymbol{I}_{\rm gt}}}}, \tag{3}$$

where $N_{\Psi_p^{\boldsymbol{I}_{\rm gt}}}$ denotes the number of elements in the feature map extracted from the VGG16 layer p, and $\boldsymbol{I}_{\rm comp}$ results from the model output with the valid pixels set to their ground-truth values.

Further, the style loss $L_{\rm style}$ uses the same feature maps extracted from the VGG16 network as those used for $L_{\rm perc}$, but it computes the L1 loss over their autocorrelation given by the Gram matrix,

$$\begin{aligned} L_{\rm style} = \sum_{p-1}^{3} & \frac{\left\| K_p \left((\Psi_p^{\boldsymbol{I}_{\rm out}})^T (\Psi_p^{\boldsymbol{I}_{\rm out}}) - (\Psi_p^{\boldsymbol{I}_{\rm gt}})^T (\Psi_p^{\boldsymbol{I}_{\rm gt}}) \right) \right\|_1}{C_p C_p} \\ & + \frac{\left\| K_p \left((\Psi_p^{\boldsymbol{I}_{\rm comp}})^T (\Psi_p^{\boldsymbol{I}_{\rm comp}}) - (\Psi_p^{\boldsymbol{I}_{\rm gt}})^T (\Psi_p^{\boldsymbol{I}_{\rm gt}}) \right) \right\|_1}{C_p C_p}, \end{aligned} \tag{4}$$

where $K_p = 1/(C_p H_p W_p)$ is the normalization factor, with $(C_p H_p W_p)$ being the size of the feature vector extracted from layer p. In Equation (4), C_p is the number of channels while W_p and H_p refer to the number of pixels along the width and height of the image, respectively. The style loss $L_{\rm style}$ helps constrain the texture of the predicted maps to match that of the ground truth.

Finally, the total-variation loss $L_{\rm tv}$ is used to allow for the spatial smoothness in the output map,

$$\begin{aligned} L_{\rm tv} = & \sum_{(i,j)\in R,(i,j+1)\in R} \frac{\left\| \boldsymbol{I}_{\rm comp}^{i,j+1} - \boldsymbol{I}_{\rm comp}^{i,j} \right\|_1}{N_{\boldsymbol{I}_{\rm comp}}} \\ & + \sum_{(i+1,j)\in R,(i,j)\in R} \frac{\left\| \boldsymbol{I}_{\rm comp}^{i+1,j} - \boldsymbol{I}_{\rm comp}^{i,j} \right\|_1}{N_{\boldsymbol{I}_{\rm comp}}}, \end{aligned} \tag{5}$$

where R represents the 1 pixel dilation of the hole region.

Different loss terms described above are weighted by the corresponding weights and combined to provide the total loss

function, $L_{\rm tot}$:

$$L_{\rm tot} = L_{\rm valid} + 6L_{\rm hole} + 0.05L_{\rm perc} + 120L_{\rm style} + 0.1L_{\rm tv}. \quad (6)$$

The weights associated with each term in Equation (6) are identical to those set by Liu et al. (2018), found by empirical calibration.

It is worth noting that we deliberately choose to use a loss function that has been trained to work directly on the images rather than to optimize some commonly used summary statistic such as the power spectrum or density pdf. This is because the optimal summary statistics for cosmological information are unknown. We postulate that this loss function, which was developed in previous work and shown to be successful for complex images, will be effective at reproducing salient features of our fields. This postulation is explicitly tested in the next section for common cosmological summary statistics for which the loss function has not been explicitly tuned.

3.4. Training

In order to train the model, we first split the 15,000 IllustrisTNG-based CMD gas temperature maps into the train, validation and test sets. We assign 10,000 maps to the train set, 2000 to the validation set, and 3000 to the test set.

We train the network using four NVIDIA P100 GPUs for 130 epochs. Each epoch consists of multiple iterations, with a single iteration using a batch of 16 maps. At a given iteration, each map is coupled with a randomly selected binary mask from a pool of 12,000 masks for data augmentation purposes. We apply the $\log_{10}$ transformation to the input maps, to reduce the dynamic range of the temperature values, and then normalize the training set to zero mean and unit variance using the mean $\mu_{\rm train}$ and standard deviation $\sigma_{\rm train}$ of the train set. The same parameters ($\mu_{\rm train}$, $\sigma_{\rm train}$) are then used to normalize the $\log_{10}$-transformed validation and test sets. We use the `Adam` optimizer and set the initial learning rate to 0.0002. We use the PyTorch `ReduceLROnPlateau` function to implement the update policy that decays the learning rate by a factor of 10 if no decrease in the training loss $L_{\rm tot}$ is observed for five consecutive epochs. The training process is completed in approximately 24 hr. After each epoch, the model is evaluated on the validation set in order to monitor any overfitting to the training set.

4. Results

We evaluate the model predictions using the holdout test set of 3000 gas temperature maps and binary masks. None of these maps and masks is exposed to the model during training, in order to check how well the results generalize to new data. We first show a visual comparison of the ground truth and predicted maps in Section 4.1. We then quantitatively assess the reliability of the inpainted maps using the probability density function in Section 4.2, and the 2D power spectrum in Section 4.3. In Section 4.4, we also evaluate the performance of the model in recovering missing data in physical fields different from the one exposed during training.

4.1. Visual Comparison

In Figure 1, we show four temperature maps from the CMD test set. Rather than showing the raw maps, we plot the $\log_{10}$ of the temperature values to facilitate a visual inspection. From top to bottom, the first row shows the ground-truth maps, the second row displays the output of the reconstruction, and the last row shows a pixel-by-pixel comparison between the ground truth and the predicted maps. Different columns show the results for different types (regular or irregular) and extent (fraction of the total area covered) of the binary masks. The left two columns contain results using irregular-shaped masks, and a visual comparison between the ground truth and network prediction can barely spot any difference. In the case of regular masks in the right two columns of Figure 1, there are some clear differences between the target and the predicted map, even for the masks with the lower coverage (15%). This naturally arises from the fact that whole structures are wiped out in the masking process and the inpainting model aims at recovering the correct *style* (or statistical properties) in the reconstructed map rather than matching pixel-by-pixel the output and the ground-truth maps. This effect is much more pronounced for the regular masks that cover 30% of the total pixels. Indeed, large structures in the reference map are replaced by an ensemble of smaller structures. This result is not surprising, because the lost information cannot be retrieved from the valid pixels given the size of the mask relative to that of the cosmological structures it covers and the size of the whole map.

We also highlight the near-perfect match between the model output and the ground-truth maps for the unmasked pixels. This can be attributed to the "skip connection" between the input map and the final stage of the recovery decoder (see Figure 4 in Zhu et al. 2021).

Finally, the use of $L_{\rm tv}$ in the total loss $L_{\rm tot}$ allows a continuity and smooth transition between the hole and valid pixels. Indeed, in none of the cases tested in this work do we find any artifacts at the edges of the binary masks.

4.2. Probability Density Function (PDF)

In order to quantify how closely the predicted maps match the ground truth, we compare their probability density functions of the temperature values. We use the p-values of the Kolmogorov–Smirnov test (K-S test hereafter) that quantifies the likelihood that the pixels temperature values in the reconstructed and the ground-truth maps are drawn from the same underlying distribution. In particular, for each map, we estimate the p-value of the K-S test by comparing the reconstructed and ground-truth maps in the masked region and repeat the exercise for all 3000 maps in the test set. Figure 2 shows the histograms of the corresponding 3000 p-values for different choices of binary masks.

Under the null hypothesis, i.e., the temperature values in the reconstructed and the ground-truth maps are drawn from the same underlying distribution, we expect a uniform distribution of the K-S test p-values. However, we notice that, for irregular masks, the distribution peaks at p-value $= 1$, with a near-exponential drop at lower p-values indicating an even stronger agreement than that expected between two samples randomly drawn from the same distribution. In order to observe a distribution, such as those seen in the top panel of Figure 2, the temperature values in a non-negligible fraction of the reconstructed pixels must match very closely their counterparts in the ground-truth map. On the other hand, the observed p-values in the case of regular masks in the lower panel of Figure 2 provide strong evidence against the null hypothesis, especially when 30% of the pixels are masked. Although, for regular masks covering 15% of the data, some of the maps

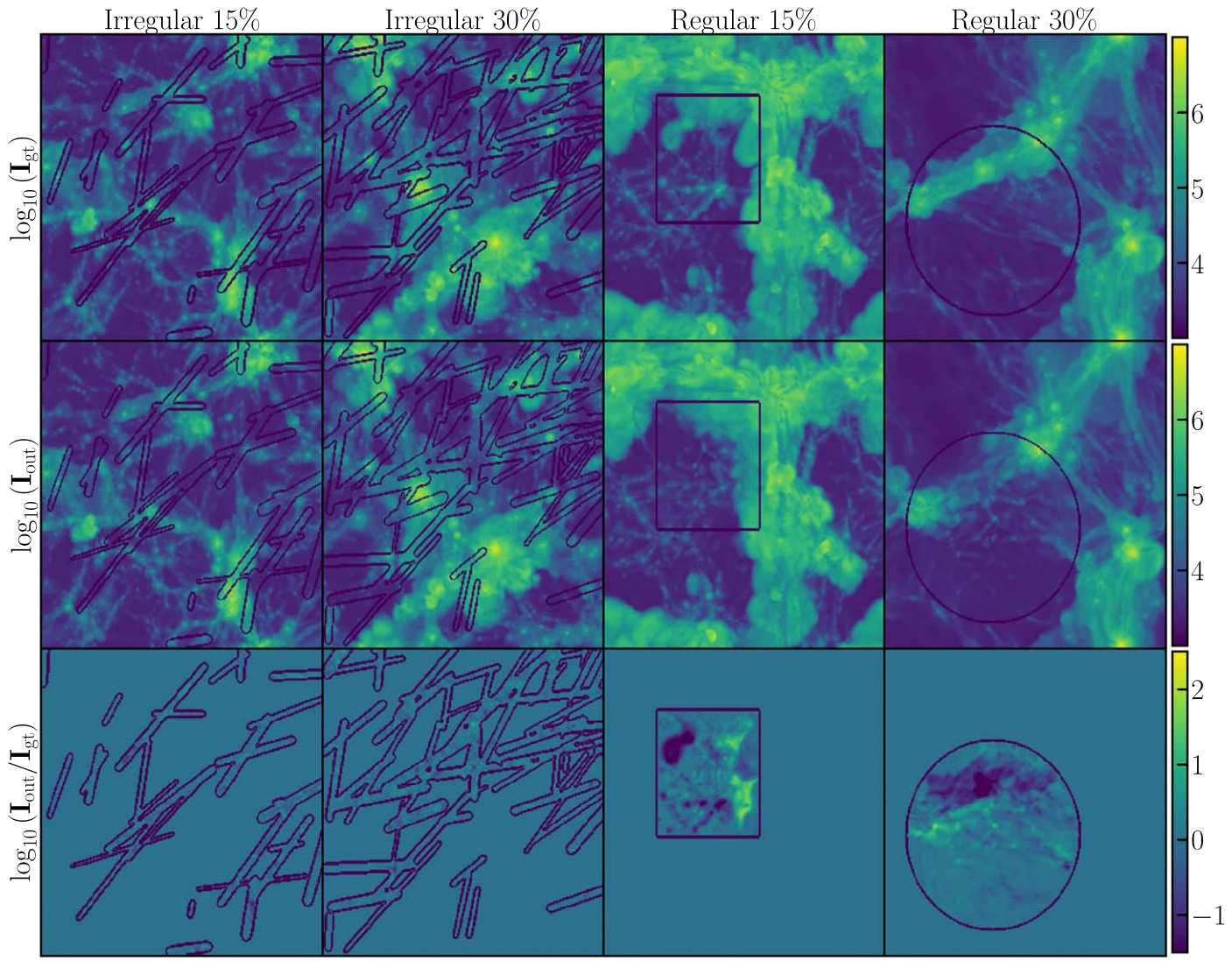

Figure 1. Temperature maps extracted from the CAMELS test set. Different columns display results applying masks of different types and extents. We show the $\log_{10}$ of the temperature maps for better visualization. Rows in order from top to bottom show the ground-truth maps $\log_{10}\boldsymbol{I}_{\rm gt}$, prediction of the network $\log_{10}\boldsymbol{I}_{\rm out}$, and the difference between the ground truth and the output maps $\log_{10}(\boldsymbol{I}_{\rm out}/\boldsymbol{I}_{\rm gt})$. The contours show the area delimited by the masks. The different panels in the top two rows share the same color scale, while the color range in the bottom row is adapted to highlight the structures in the plot.

exhibit p-values larger than the threshold of 0.05 that is typically used to reject (lower p-values) or accept (larger p-values) the null hypothesis; these form only a relatively small fraction of the 3000 test maps.

In order to understand the trend observed in Figure 2, we notice from Figure 1 (last row) that there is a near-perfect match in the temperature values between the reconstructed and the ground-truth map near the edges of the mask. However, this is not surprising, given the use of the "per-pixel" loss and the "total-variation" loss to train the network that together ensure continuity in the reconstructed map between the hole and valid pixels. On the other hand, the neural network struggles to provide accurate reconstruction in the innermost part of larger masked patches. For irregular masks, a smaller fraction of the area being masked results in a lower probability of different segments that form the mask being joined together to create a single large patch. This increases the fraction of the hole pixels that are close to the mask boundaries where the reconstructed temperature field closely matches its ground-truth values. This explains the blue histogram (irregular masks covering 15% of the data) in the top panel of Figure 2 being more skewed toward 1 than the red one (irregular masks erasing 30% of the data). For regular masks, along with the aforementioned cause, another effect that contributes to the bad performance seen in Figure 2 (and later on in Figures 4 and 5) is the unique nature of the structures being removed. In particular, the structure that are erased by the regular masks are unique, and the network is unable to retrieve the semantic features of the missing data from the valid pixels of the map. The latter effect is field-dependent, and we expect a much better performance in terms of recovering both accurate probability density function and the power spectrum for a field that is more homogeneous on the scales of the $(25\ {\rm Mpc}\ h^{-1})^2$ maps, such as the temperature fluctuations seen in the CMB.

To support the aforementioned arguments, we show two extreme cases in Figure 3 when irregular masks covering 15% of the pixels are applied to maps in the test set. The two columns show the cases that result in the lowest (left column) and highest (right column) p-values observed in the blue histogram in the top panel of Figure 2. We notice that the

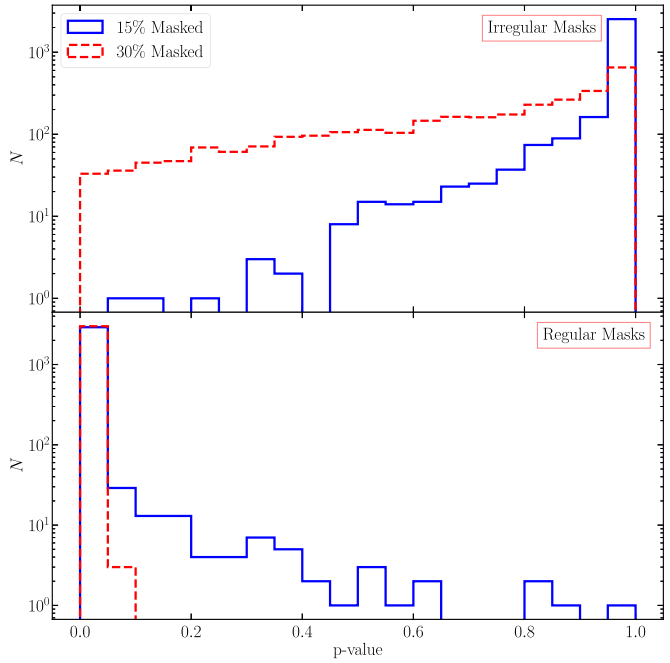

Figure 2. Distribution of the *p*-values of the Kolmogorov–Smirnov test performed on the 3000 maps that form the test set. Blue and red histograms show results for masks covering 15% and 30% of the total area, respectively. Top panel: results applying irregular masks. Bottom panel: results when regular-shaped masks are used. For irregular masks, the observed distribution of the K-S test *p*-values in the top panel supports the null hypothesis that the pixel values of the reconstructed and ground-truth maps are drawn from the same distribution. For regular masks, the distributions seen in the bottom panel indicate that the model fails to match the probability density function of the pixel values in the ground-truth temperature maps.

segments that make the irregular mask in the left column (lowest *p*-value of the Kolmogorov–Smirnov test) cluster together in correspondence with the white arrow form a large blob where the reconstructed image differs significantly from the ground-truth one. Similar large continuous masked regions are absent in the right column, where the model yields a near-perfect reconstruction.

We also investigate whether the *p*-values of the K-S test correlate with any of the six cosmological or astrophysical parameters used to run the simulations. The Pearson correlation coefficients reported in Table 1 show that there is no significant correlation between the K-S test *p*-values and any of the simulation parameters. This indicates that the model performance mainly depends on the properties and extent of the mask —and not that much on the particular cosmological and astrophysical model employed.

While the K-S test quantifies the statistical differences between the probability density functions of the ground truth and the predicted map, it does not indicate where these differences originate from. To investigate if the model's bad performance occurs in specific regimes of the temperature values, we compare the corresponding probability density functions estimated from the ground truth and the predicted maps. In particular, for each map in the test set, we apply the $\log_{10}$ transformation and the `min-max` scaling to both the ground truth and the model output before estimating the probability density functions. We show the results in Figure 4, where the *y*-axis shows the difference between the two distributions averaged over 3000 maps from the test set in units of the standard error on the mean as a function of the `min-max` scaled logarithmic temperature. We note that the disagreement between the model prediction and the ground truth is (i) stronger in the regime of low pixel values and improves in pixels with higher field intensity; (ii) as expected, worse for regular masks compared to the irregular ones; and (iii) higher for regular masks covering a larger extent of the total area.

4.3. Power Spectrum

Besides the probability density function, another widely used statistic in cosmology is the power spectrum, defined in this case as

$$P(k_1)\delta^D(\boldsymbol{k}_1 - \boldsymbol{k}_2) = \langle F(\boldsymbol{k}_1)F(\boldsymbol{k}_2)\rangle, \quad (7)$$

where $F(\boldsymbol{k})$ is the Fourier transform of the considered field $F(\boldsymbol{x})$ and δ^D is the Dirac delta. Note that the fields we consider are statistically homogeneous and isotropic, so the power spectrum only depends on the magnitude of the wavenumber, k. We use the publicly available *Pylians3*[9] library to compute the power spectra of the maps. In this section, we use the power spectrum as a summary statistic to quantify the agreement between the reconstructed maps and their unmasked versions.

The results are shown in Figure 5 for the data from the test set, masked using irregular and regular masks. The top panels show the power spectra measured from the masked data (red thick and blue dashed lines), from the reconstructed maps (blue and red dots with corresponding statistical errors), as well as from the ground-truth maps (black thick lines) averaged over the 3000 maps from the test set. The error bars (on red and blue dots) and shaded bands (around the black thick lines) show the errors on the mean of the 3000 estimates (i.e., the standard deviation scaled by $\sqrt{3000}$). The differences between the power spectra from the reconstructed and the ground-truth maps are barely visible in the top panels. We thus show the ratio between these two quantities in the bottom panels of Figure 5. As in the top panels, shaded bands in the bottom panels of Figure 5 correspond to the error on the mean of 3000 estimates.

For irregular-shaped masks, we find that the power spectra of the reconstructed maps agree very well with the reference ones. In particular, for masks covering 15% of the total area, the power spectra from the reconstructed maps show a systematic bias with respect to the reference ones of less than ∼1% up to a wavenumber of $k \sim 20\ h\,\mathrm{Mpc}^{-1}$ (blue dots with error bars in the top left panel, blue line with shaded band in the bottom left panel of Figure 5). The accuracy degrades only marginally for wavenumbers below $k \sim 20\ h\,\mathrm{Mpc}^{-1}$, when extending the analysis to irregular masks that cover 30% of the input maps (red points in top left panel and red line in the bottom left panel). For larger wavenumbers (up to the Nyquist wavenumber of $k_{\mathrm{Nyq.}} \sim 30\,h\,\mathrm{Mpc}^{-1}$), the power spectra estimated from the reconstructed maps stay accurate within ∼5% (∼10%) for irregular masks covering 15% (30%) of the area.

For regular masks that cover a continuous area of the maps, the reconstruction is less accurate than that for the irregular-shaped masks. As already discussed in Section 4.2, this is due to the fact that regular-shaped masks erase entire structures in a single large patch. Furthermore, the learning process is also complicated by the fact that we have only very limited number

[9] https://pylians3.readthedocs.io

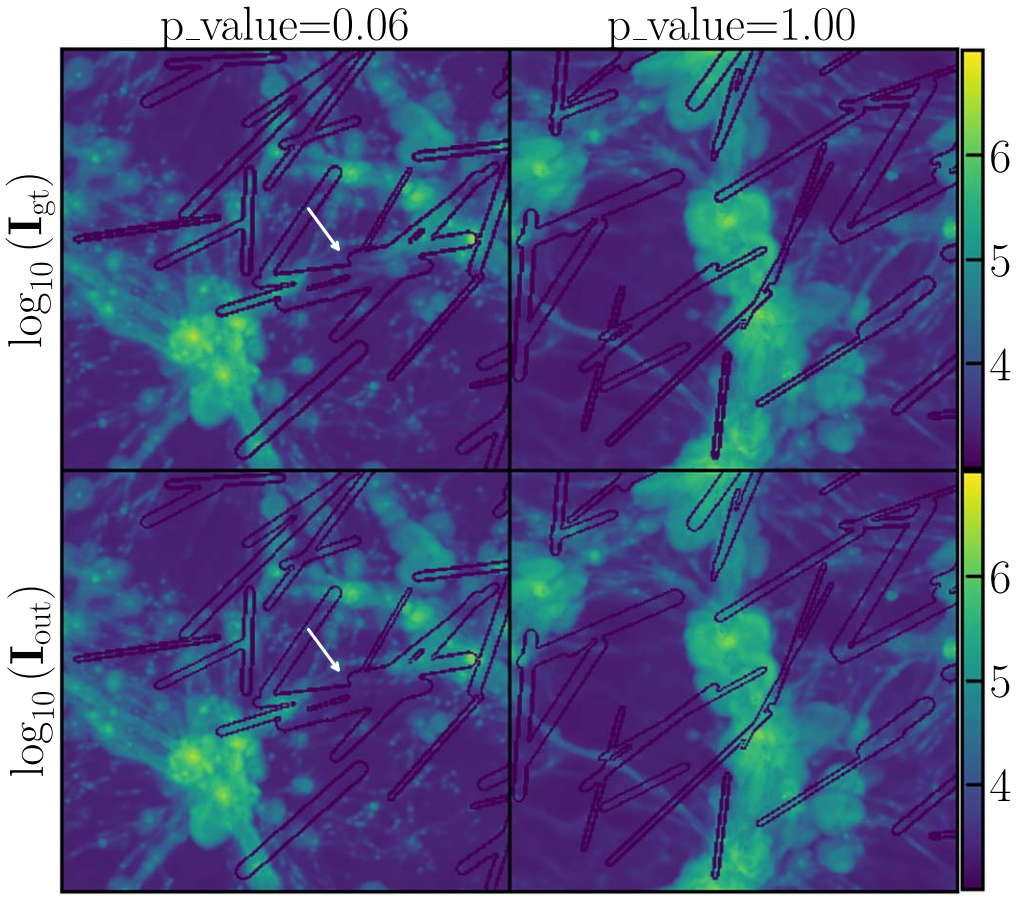

Figure 3. Two cases from the test set where the comparison of the pixel values distributions using the Kolmogorov–Smirnov test results in two extreme *p*-values: very low in the left column and very high in the right column. In both cases, irregular masks that cover 15% of the data are applied. The top row shows the ground-truth maps while the bottom row shows the reconstructed ones. In all panels, dark contours represent the edges of the binary mask. The reconstruction performs poorly in the large contiguous masked patch highlighted by the white arrow in the left column, resulting in a low *p*-value. Such large continuous patches are absent in the right column, and the model returns a near-perfect reconstruction and a high *p*-value.

Table 1
Correlation Coefficients between the Simulation Parameters and the Model Performance.

	Irr. 15%	Irr. 30%	Reg. 15%	Reg. 30%
Ω_m	−0.041	+0.055	−0.040	−0.036
σ_8	+0.075	+0.084	−0.028	−0.029
$A_{\rm SN1}$	−0.058	−0.082	+0.007	−0.002
$A_{\rm AGN1}$	+0.001	+0.007	−0.010	+0.013
$A_{\rm SN2}$	−0.001	−0.000	−0.017	+0.011
$A_{\rm AGN2}$	+0.007	+0.022	−0.007	−0.041

Note. Pearson Correlation Coefficients between the values of the six simulation parameters and the K-S test p-values estimated using the model output for the 3000 Maps from the holdout test set. The different columns report results for irregular (Irr.) and regular (Reg.) shaped masks covering 15% and 30% of the data.

(10) of maps for each set of simulation parameters to train the model, far below the standard size of data sets used to train deep convolutional neural networks. Nevertheless, the network does an excellent job in reconstructing maps where the mask covers 15% of the total area, with the recovered power spectra matching the reference ones within ∼5% up to Nyquist wavenumber of $k_{\rm Nyq.} \sim 30\,h\,{\rm Mpc}^{-1}$. For regular masks that erase 30% of the data, the agreement degrades drastically and becomes strongly scale-dependent.

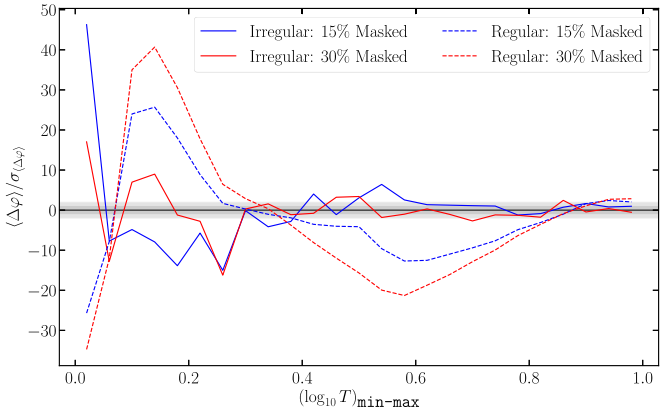

Figure 4. Mean difference between the probability density functions of the `min-max` scaled and $\log_{10}$-transformed temperature maps, estimated from the model output and the corresponding ground-truth map averaged over 3000 maps in the test set, in units of the standard error on the mean. Continuous lines show results when data are masked using irregular masks, while dashed lines correspond to the cases when regular-shaped masks are employed. Blue and red lines correspond to masks covering 15% and 30% of the total area, respectively. Horizontal shaded bands delimit 1σ and 2σ intervals. It is evident from this figure that the difference between the probability density functions of the reconstructed values and their ground-truth counterpart is (i) larger in the low-intensity regime, (ii) larger for regular masks compared to the irregular masks covering the same extent, and (iii) positively correlated with the extent of the masked regions.

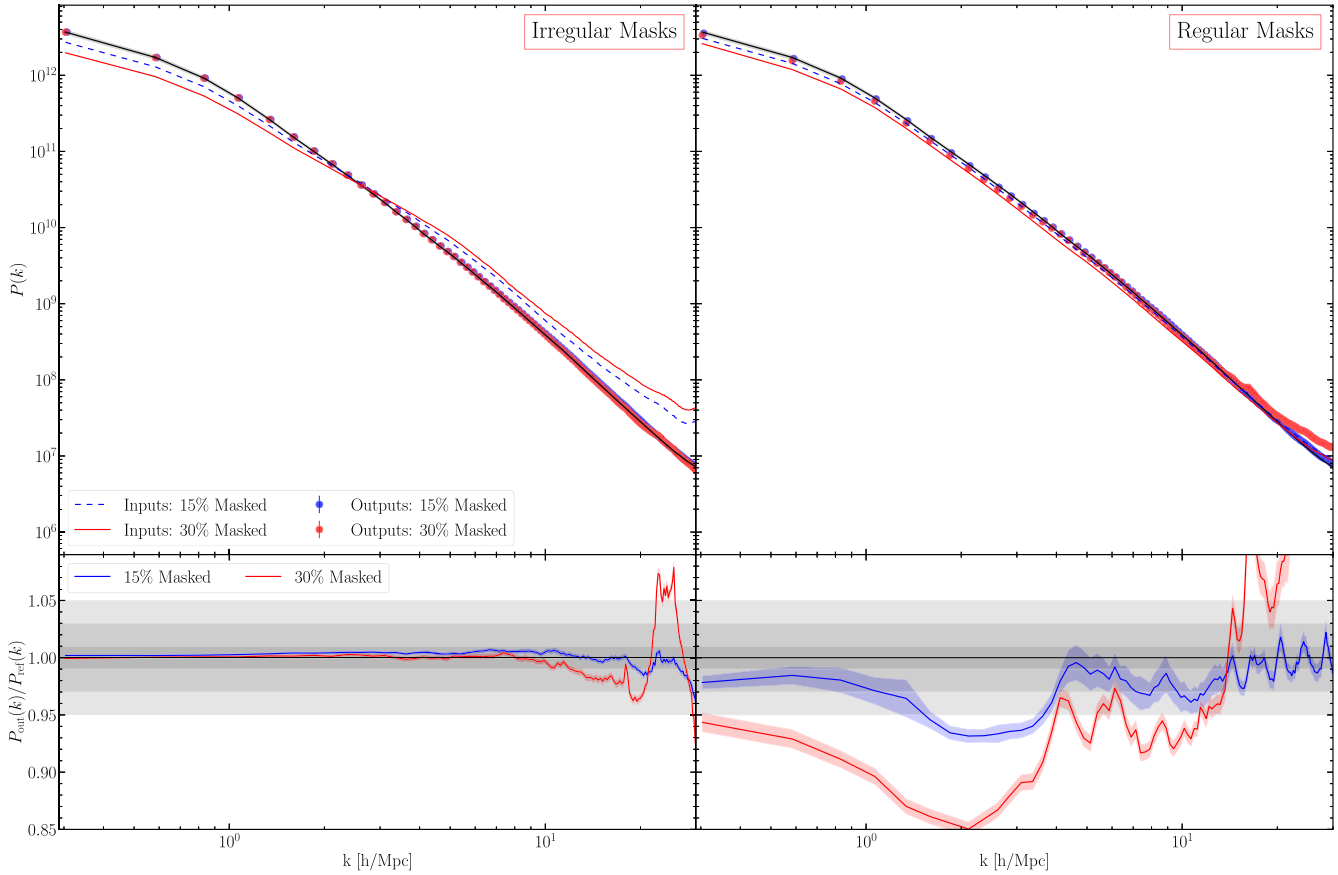

Figure 5. Top panels: Power spectra measured from the ground-truth gas temperature maps averaged over 3000 maps in the test set (black thick line with shaded band), from the input maps masked using irregular-shaped masks (left panels) or regular-shaped masks (right panels) that cover 15% (blue shaded line) and 30% (red thick line) of the total area. Results, after the reconstruction is applied, are shown as blue dots with error bars for masks covering 15% and with red dots with error bars for those covering 30% of the area. Bottom panels: ratio between the power spectra measured from the reconstructed $P_{\rm out}$ and ground-truth maps $P_{\rm ref}$ are shown when masks cover 15% (blue line with shaded band) and 30% (red line with shaded band) of the area. All errors shown as shaded bands or error bars refer to the error on the mean of 3000 estimates.

This analysis, combined with the results in Section 4.2, shows that the neural network breaks down when large portions of the data are erased in a single patch, while results are reliable for the measured power spectrum in other cases explored in this work. One natural way to improve the performance is to train the neural network either on a larger number of simulations for each set of cosmological and astrophysical parameters or on data over an area much larger than the homogeneity scale of the field.

4.4. Performance on Auxiliary Data

So far, we have used the CMD IllustrisTNG-based gas temperature maps, split into the train, validation, and test sets, to both train the model and test its performance on unseen data. In this section, we use this model and try to reconstruct the missing data in a number of different fields that are not used during the model training. In particular, we test the performance of the model to recover missing data in maps from other fields, such as: (1) SIMBA-based gas temperature maps ($T_{\rm SIMBA}$), (2) gas density ($M_{\rm gas}$), (3) total matter density ($M_{\rm tot}$), (4) gas pressure (P), (5) electron density (n_e), and (6) the magnesium-to-iron ratio (Mg/Fe). Here, we limit the analysis to irregular masks that cover 15% of the total area.

The scales of the pixel intensities in these auxiliary fields are significantly different than the gas temperature field used to train the model. In order to feed the neural network with pixel values that cover a range similar to that of the training set, we first rescale each single map to the `min-max` range of the gas temperature maps in the training set and then normalize it using the ($\mu_{\rm train}$, $\sigma_{\rm train}$) values used in Section 3.4.

The visual comparison between the ground truth and the model output is shown in Figure 6, where each row contains results from a different field. Except for the gas pressure (P) maps, the differences between the model output and the ground truth are very subtle and can be noticed only through a direct comparison as shown in the rightmost column in Figure 6. For the gas pressure (P) map, the model completely fails to recover reliable estimates of the field in specific regions where the model predicts negative pressure. We note that this mainly occurs in the areas with low pixel values in the ground-truth maps, in agreement with results shown in Figure 4, i.e., the model struggles to provide accurate estimates of the field in the low-intensity areas. While this effect is unnoticeable in other fields, it is exacerbated for the gas pressure map. It is very interesting to see that, even for fields that have a very different morphology, e.g., Mg/Fe, our model is still able to inpaint features with great success.

We also perform the analysis using the power spectrum of the auxiliary fields and show the results in Figure 7. Although results for all six fields are worse than those seen in Figure 5, we notice that, for the magnesium–iron density field, the model is able to match the reference power spectrum within ∼5% up

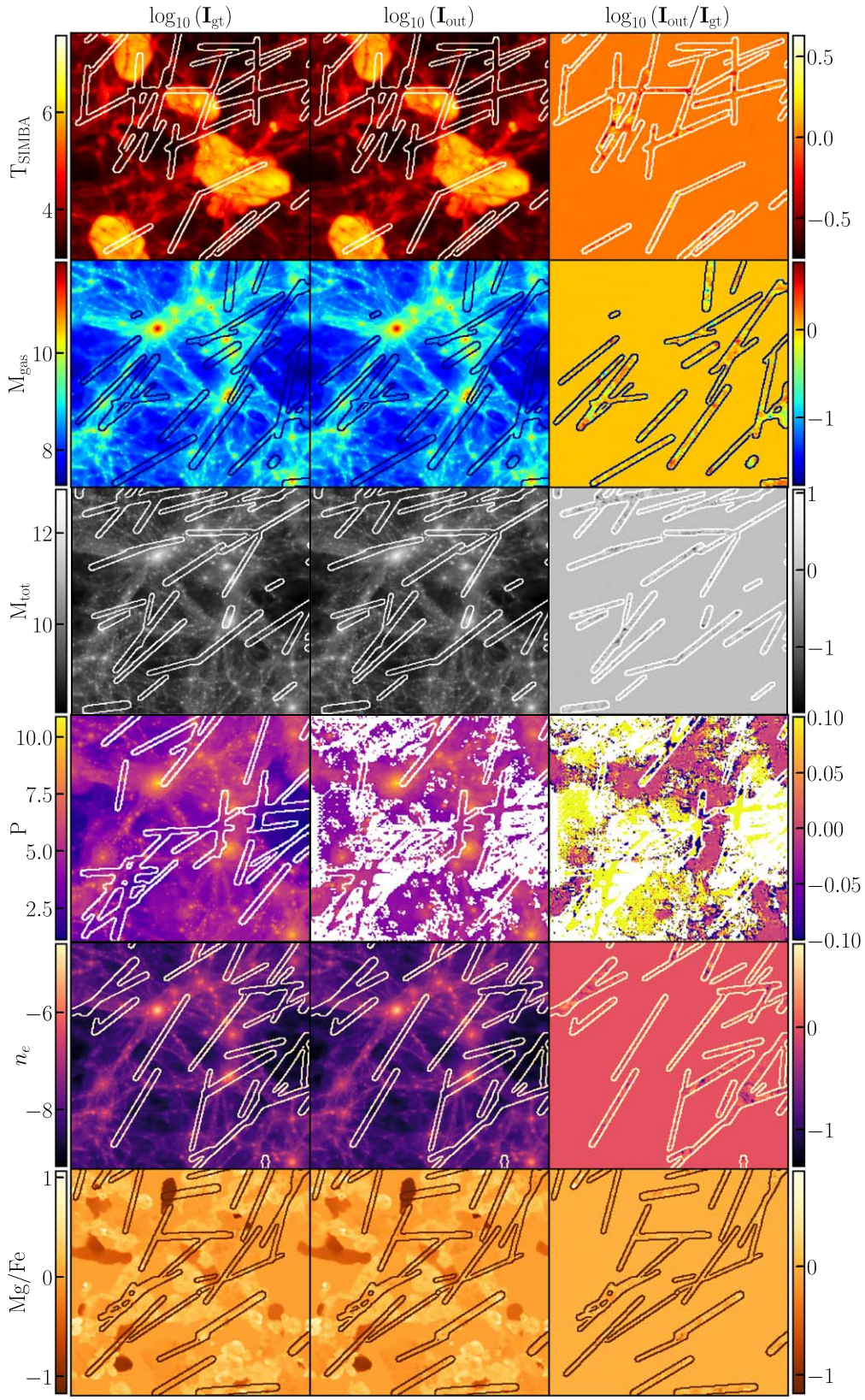

Figure 6. Two-dimensional maps of six different auxiliary fields not used in the network training. From top to bottom row: SIMBA-based gas temperature map ($T_{\rm SIMBA}$), gas density ($M_{\rm gas}$), total matter density ($M_{\rm tot}$), gas pressure (P), electron number density (n_e), and the magnesium-to-iron ratio (Mg/Fe). Column-wise from left to right: ground truth ($\log_{10}(\boldsymbol{I}_{\rm gt})$), model output ($\log_{10}(\boldsymbol{I}_{\rm out})$), and the difference between the model output and the ground-truth maps ($\log_{10}(\boldsymbol{I}_{\rm out}/\boldsymbol{I}_{\rm gt})$). For a fixed row, the left two columns share the same color coding shown in the color bars on the left, while the range of the color map in the rightmost column is adapted to highlight the differences between the model output and the ground-truth map. The color bars on the right show the color coding for the rightmost column.

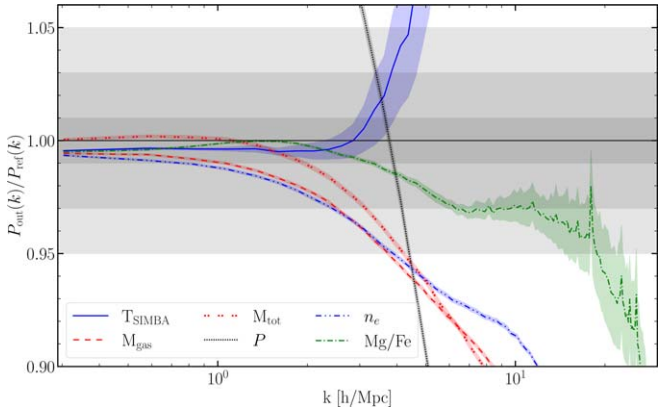

Figure 7. Ratio between the power spectra estimated using the model output and the ground truth averaged over 15,000 maps. Shaded bands show the corresponding errors on the mean. The horizontal bands delimit the 1%, 3%, and 5% intervals around the reference. Results are shown for six different auxiliary fields not used in the network training process. This result indicates that the model trained on the temperature maps fails to accurately reconstruct the masked pixels when applied to a different field. The degree at which it fails also varies from field to field.

to $k \sim 10\,h\,\mathrm{Mpc}^{-1}$. This can be explained by the fact that, as seen in Figure 6, the structures in the Mg/Fe maps are less complex compared to the other fields. These structures also extend well beyond the typical width of the masks. Interestingly, even for the SIMBA-based gas temperature maps, the power spectra of the reconstructed maps are accurate, within the level of 1%, only for $k < 3\,h\,\mathrm{Mpc}^{-1}$, indicating a difference in the small-scale morphological features with respect to the maps based on the IllustrisTNG simulations. The predicted power spectra show a systematic error within $\sim$1% up to $k \sim 1\,h\,\mathrm{Mpc}^{-1}$ for the gas density ($M_{\rm gas}$) and the electron density (n_e) fields. This scale extends to $k \sim 2\,h\,\mathrm{Mpc}^{-1}$ for the total matter density ($M_{\rm tot}$) field. On the other hand, the neural network completely fails to return reliable predictions for the gas pressure (P) maps, resulting in a significantly biased estimates of the power spectra. We do not attempt to provide a physical explanation for these results and leave this for future work.

Our analysis in this section shows that the model does not generalize particularly well to fields it is not exposed to during training. While the model fails to return reliable predictions for some fields, in other cases the validity of the predictions is limited to the largest scales (smallest wavenumbers k). These results highlight the need to train a model specifically for the field under investigation. In other words, our model has learned characteristic features of the gas temperature field that, although very generic due to the large variety of cosmological and astrophysical models present in CMD, are still very distinct to those present in other fields.

5. Summary and Conclusions

In this paper, we test the ability of a state-of-the-art deep convolutional neural network architecture, based on the MADF module, to inpaint masked pixels in 2D maps of the CAMELS CMD. We focus our attention on the gas temperature maps based on the IllustrisTNG simulations; CMD provides 15,000 maps obtained from 1000 state-of-the-art magnetohydrodynamic simulations with different values of the cosmological and astrophysical parameters.

The data set is split into a train set of 10,000 maps, a validation set of 2000 maps and a test set of 3000 maps. We mimic the missing/masked data in the maps by applying two different kinds of binary masks: (1) regular-shaped ones that cover a continuous area of each map in a circular or rectangular patch randomly placed within the map and (2) irregular-shaped masks that are composed of a number of segments of various width and length randomly placed across the map area. For each type of mask, we test the model performance using two different extents, covering 15% and 30% of the total area. We train the model for 130 epochs using a batch size of 16 for a total of 81,250 training iterations.

We check the model performance using the holdout test set of 3000 gas temperature maps and different binary masks. Through a qualitative visual comparison between the model output and the target ground truth, we first show that the model outputs are visually indistinguishable from the ground truth for irregular masks covering either 15% or 30% of the map. The difference becomes more evident for regular-shaped masks. In particular, for regular masks covering 30% of the data in each map, reticular-like artifacts start to appear in correspondence with the masked pixels, indicating a breakdown of the model for such a large masks. We also quantify the statistical agreement between the output of the model and the unmasked maps using two different summary statistics: (i) the probability density functions and (ii) the 2D power spectrum.

We compare the temperature probability density functions of the model output with that of the ground truth using the Kolmogorov–Smirnov test in the masked regions. We find that, for irregular masks, the observed distribution of the K-S test p-values supports the null hypothesis that the reconstructed maps follow the same distribution of the corresponding ground-truth maps. For regular masks, on the other hand, the results of the K-S test indicate that the model fails to match the probability density function of the ground-truth temperature maps. In particular, for regular masks covering 15% of the pixels, a vast majority of the 3000 test maps exhibit a p-value < 0.05 that indicates a rejection of the null hypothesis. For the largest regular masks that occult 30% of the pixels, we find that the K-S test p-values are systematically $\lesssim$0.05, indicating a strong evidence against the hypothesis that the reconstructed field matches the ground truth in distribution. We do not find any correlation between the K-S test p-values and any of the six simulation parameters. We also show that the main sources of such a disagreement are the low-intensity pixels.

Estimates of the 2D power spectra highlight an excellent agreement with a systematic error below 1%–2% up to $k \sim 20\,h\,\mathrm{Mpc}^{-1}$ between the model output and the ground truth when data are masked using irregular masks covering up to 30% of the pixels. The accuracy deteriorates significantly when regular masks are employed, although the systematic offset remains within 5% up to the Nyquist wavenumber $k \sim k_{\rm Nyq.}$ when only 15% of the pixels are masked. The model breaks down when regular masks covering 30% of the total area are used.

The main cause of the model breakdown when data are erased in large patches is the unique nature of the structures being removed combined with a smaller number of maps (for each set of cosmological and astrophysical parameters) used to train the network. On one hand, the neural network is unable to retrieve the statistical properties of the missing data from the unmasked pixels; on the other hand, it fails to learn

the semantic features of the field from the ensemble of the training maps for a fixed set of cosmological and astrophysical parameters. We thus expect an improvement in the model performance by increasing either the size of each map or the number of maps in the training set.

Finally, we use the model that was trained on the CMD gas temperature maps to perform inpainting on CMD maps of different fields like the SIMBA-based gas temperature maps ($T_{\rm SIMBA}$), the total matter density ($M_{\rm tot}$), the gas density ($M_{\rm gas}$), the gas pressure (P), the magnesium-to-iron ratio (Mg/Fe), and the electron density (n_e). We find that, even when using irregular masks that extend over 15% of the pixels, the model performance degrades significantly compared to when it is applied to the same field it is trained on. An even more important result is that the model performance becomes strongly field-dependent, indicating the need to train the model specifically on the field under investigation.

We conclude that the model used in this work is able to recover reliable pixel values distributions when data are missing in irregular-shaped patches. These results hold for gas temperature maps that span 1000 different cosmological and astrophysical models and that exhibit very different morphological aspects, such as halos, filaments, and voids. The power spectra of the inpainted maps exhibit an impressive agreement with their unmasked versions: within 1% for $k_{\rm max} = 20\ h\,{\rm Mpc}^{-1}$ and within 5% all the way to the Nyquist wavenumber at $k \sim 30\ h\,{\rm Mpc}^{-1}$. For regular-shaped masks, our model breaks down in recovering reliable probability density functions for the field in the masked patches, regardless of the extent, while it yields power spectrum estimates accurate at 5% only when 15% of the pixels are masked. This could be a consequence of the very large variety of models seen by the networks; we would expect a higher accuracy also for regular masks if the model was trained on a very large number of images with a fixed cosmological and astrophysical model.

The results presented in this paper have important consequences for cosmological surveys, where missing, masked, and damaged data are very common issues. This paper paves the way to tackle these issues in a novel way. However, more work is needed in order to apply this to real data. We plan to pursue this direction in future work.

This work has made use of the Tiger cluster of Princeton University. The CAMELS Multifield Data set (CMD) is publicly available at https://camels-multifield-dataset.readthedocs.io. Details on the CAMELS simulations can be found at https://www.camel-simulations.org.

D.A.A. was supported in part by NSF grants AST-2009687 and AST-2108944, and by the Flatiron Institute, which is supported by the Simons Foundation.

ORCID iDs

Faizan G. Mohammad https://orcid.org/0000-0001-9243-7434
Francisco Villaescusa-Navarro https://orcid.org/0000-0002-4816-0455
Shy Genel https://orcid.org/0000-0002-3185-1540
Daniel Anglés-Alcázar https://orcid.org/0000-0001-5769-4945
Mark Vogelsberger https://orcid.org/0000-0001-8593-7692

References

Ade, P., Aguirre, J., Zeeshan, A., et al. 2019, JCAP, 2019, 056
Allys, E., Marchand, T., Cardoso, J. F., et al. 2020, PhRvD, 102, 103506
Banerjee, A., & Abel, T. 2021a, MNRAS, 504, 2911
Banerjee, A., & Abel, T. 2021b, MNRAS, 500, 5479
Banerjee, A., Castorina, E., Villaescusa-Navarro, F., Court, T., & Viel, M. 2020, JCAP, 2020, 032
Bayer, A. E., Villaescusa-Navarro, F., Massara, E., et al. 2021, ApJ, 919, 24
Bianchi, D., & Verde, L. 2020, MNRAS, 495, 1511
Dai, J.-P., Verde, L., & Xia, J.-Q. 2020, JCAP, 2020, 007
Davé, R., Anglés-Alcázar, D., Narayanan, D., et al. 2019, MNRAS, 486, 2827
Dawson, K. S., Schlegel, D., Ahn, C., et al. 2013, AJ, 145, 10
de la Bella, L. F., Tessore, N., & Bridle, S. 2021, JCAP, 2021, 001
de la Torre, S., Guzzo, L., Peacock, J., et al. 2013, A&A, 557, A54
Demir, U., & Unal, G. 2018, arXiv:1803.07422
DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016, arXiv:1611.00036
Fluri, J., Kacprzak, T., Lucchi, A., et al. 2019, PhRvD, 100, 063514
Friedrich, O., Uhlemann, C., Villaescusa-Navarro, F., et al. 2020, MNRAS, 498, 464
Gatys, L. A., Ecker, A. S., & Bethge, M. 2015, arXiv:1508.06576
Giri, U., & Smith, K. M. 2022, JCAP, 2022, 028
Gualdi, D., Gil-Marin, H., & Verde, L. 2021a, JCAP, 2021, 008
Gualdi, D., Novell, S., Gil-Marín, H., & Verde, L. 2021b, JCAP, 2021, 015
Gupta, A., Matilla, J. M. Z., Hsu, D., & Haiman, Z. 2018, PhRvD, 97, 103515
Hahn, C., & Villaescusa-Navarro, F. 2021, JCAP, 2021, 029
Hahn, C., Villaescusa-Navarro, F., Castorina, E., & Scoccimarro, R. 2020, JCAP, 2020, 040
Hassan, S., Andrianomena, S., & Doughty, C. 2020, MNRAS, 494, 5761
Hopkins, P. F. 2015, MNRAS, 450, 53
Jeffrey, N., Alsing, J., & Lanusse, F. 2021, MNRAS, 501, 954
Kuruvilla, J., & Aghanim, N. 2021, A&A, 653, A130
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv:1110.3193
Liu, G., Reda, F. A., Shih, K. J., et al. 2018, arXiv:1804.07723
Massara, E., Villaescusa-Navarro, F., Ho, S., Dalal, N., & Spergel, D. N. 2021, PhRvL, 126, 011301
Merloni, A., Predehl, P., Becker, W., et al. 2012, arXiv:1209.3114
Mohammad, F. G., Percival, W. J., Seo, H.-J., et al. 2020, MNRAS, 498, 128
Montefalcone, G., Abitbol, M. H., Kodwani, D., & Grumitt, R. D. P. 2021, JCAP, 2021, 055
Nazeri, K., Ng, E., Joseph, T., Qureshi, F. Z., & Ebrahimi, M. 2019, arXiv:1901.00212
Ntampaka, M., Eisenstein, D. J., Yuan, S., & Garrison, L. H. 2020, ApJ, 889, 151
Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. 2016, arXiv:1604.07379
Pillepich, A., Springel, V., Nelson, D., et al. 2018, MNRAS, 473, 4077
Puglisi, G., & Bai, X. 2020, ApJ, 905, 143
Raghunathan, S., Holder, G. P., Bartlett, J. G., et al. 2019, JCAP, 2019, 037
Ravanbakhsh, S., Oliva, J., Fromenteau, S., et al. 2017, arXiv:1711.02033
Ribli, D., Pataki, B. A. ´, Zorrilla?Matilla, J. M., et al. 2019, MNRAS, 490, 1843
Ross, A. J., Percival, W., Sánchez, A., et al. 2012, MNRAS, 424, 564
Ross, A. J., Bautista, J., Tojeiro, R., et al. 2020, MNRAS, 498, 2354
Samushia, L., Slepian, Z., & Villaescusa-Navarro, F. 2021, MNRAS, 505, 628
Schmelzle, J., Lucchi, A., Kacprzak, T., et al. 2017, arXiv:1707.05167
Sevilla-Noarbe, I., Bechtol, K., Carrasco Kind, M., et al. 2021, ApJS, 254, 24
Simonyan, K., & Zisserman, A. 2014, arXiv:1409.1556
Spergel, D., Gehrels, N., Baltay, C., et al. 2015, arXiv:1503.03757
Square Kilometre Array Cosmology Science Working Group, Bacon, D. J., Battye, R. A., et al. 2020, PASA, 37, e007
Tamura, N., Takato, M., Shimono, A., et al. 2016, Proc. SPIE, 9908, 99081M
The LSST Dark Energy Science Collaboration, Mandelbaum, R., Eifler, T., et al. 2018, arXiv:1809.01669
Uhlemann, C., Friedrich, O., Villaescusa-Navarro, F., Banerjee, A., & Codis, S. 2020, MNRAS, 495, 4006
Vafaei Sadr, A., & Farsian, F. 2021, JCAP, 2021, 012
Valogiannis, G., & Dvorkin, C. 2022, PhRvD, 105, 103534
Villaescusa-Navarro, F., Hahn, C., Massara, E., et al. 2020, ApJS, 250, 2
Villaescusa-Navarro, F., Angles-Alcazar, D., Genel, S., et al. 2021, ApJ, 915, 71
Villaescusa-Navarro, F., Genel, S., Angles-Alcazar, D., et al. 2022a, arXiv:2201.01300

Villaescusa-Navarro, F., Genel, S., Angles-Alcazar, D., et al. 2022b, ApJS, 259, 61
Weinberger, R., Springel, V., & Pakmor, R. 2020, ApJS, 248, 32
Weinberger, R., Springel, V., Hernquist, L., et al. 2017, MNRAS, 465, 3291
Yan, Z., Li, X., Li, M., Zuo, W., & Shan, S. 2018, in Proc. European Conf. on Computer Vision (ECCV) (Berlin: Springer), 1
Yang, C., Lu, X., Lin, Z., et al. 2016, in Computer Vision – ECCV 2018 15th European Conference, ed. V. Ferrari et al. (Berlin: Springer)
Yi, K., Guo, Y., Fan, Y., Hamann, J., & Wang, Y. G. 2020, arXiv:2001.11651
Yu, J., Lin, Z., Yang, J., et al. 2018, in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (Piscataway, NJ: IEEE), 5505
Yu, J., Lin, Z., Yang, J., et al. 2019, in Proc. IEEE/CVF Int. Conf. on Computer Vision (Piscataway, NJ: IEEE), 4471
Zhu, M., He, D., Li, X., et al. 2021, ITIP, 30, 4855
Zorrilla Matilla, J. M., Sharma, M., Hsu, D., & Haiman, Z. 2020, PhRvD, 102, 123506