

PGNets: planet mass prediction using convolutional neural networks for radio continuum observations of protoplanetary discs

Shangjia Zhang ¹★, Zhaohuan Zhu ¹ and Mingon Kang ²

¹*Department of Physics and Astronomy, University of Nevada, Las Vegas, 4505 S. Maryland Pkwy, Las Vegas, NV 89154, USA*

²*Department of Computer Science, University of Nevada, Las Vegas, 4505 S. Maryland Pkwy, Las Vegas, NV 89154, USA*

Accepted 2021 November 29. Received 2021 November 29; in original form 2021 March 22

ABSTRACT

We developed convolutional neural networks (CNNs) to rapidly and directly infer the planet mass from radio dust continuum images. Substructures induced by young planets in protoplanetary discs can be used to infer the potential young planets' properties. Hydrodynamical simulations have been used to study the relationships between the planet's properties and these disc features. However, these attempts either fine-tuned numerical simulations to fit one protoplanetary disc at a time, which was time consuming, or azimuthally averaged simulation results to derive some linear relationships between the gap width/depth and the planet mass, which lost information on asymmetric features in discs. To cope with these disadvantages, we developed Planet Gap neural Networks (PGNets) to infer the planet mass from two-dimensional images. We first fit the gridded data in Zhang et al. as a classification problem. Then, we quadrupled the data set by running additional simulations with near-randomly sampled parameters, and derived the planet mass and disc viscosity together as a regression problem. The classification approach can reach an accuracy of 92 per cent, whereas the regression approach can reach 1σ as 0.16 dex for planet mass and 0.23 dex for disc viscosity. We can reproduce the degeneracy scaling $\alpha \propto M_p^3$ found in the linear fitting method, which means that the CNN method can even be used to find degeneracy relationship. The gradient-weighted class activation mapping effectively confirms that PGNets use proper disc features to constrain the planet mass. We provide programs for PGNets and the traditional fitting method from Zhang et al., and discuss each method's advantages and disadvantages.

Key words: hydrodynamics – waves – planet–disc interactions – protoplanetary discs.

1 INTRODUCTION

Detecting young planets in protoplanetary discs is essential to infer where and when planets form and how massive they are, putting stringent constraints on planet formation theory. Unfortunately, despite thousands of exoplanets having been discovered, only a few of them are around young stars within 10 million years old. There are even fewer young forming planets found in dusty protoplanetary discs. One notable example is PDS 70 system, where two young planets with several Jupiter mass have been discovered within the dusty cavity of 80 au (Keppler et al. 2018; Müller et al. 2018; Wagner et al. 2018; Christiaens et al. 2019; Haffert et al. 2019; Isella et al. 2019; Hashimoto et al. 2020; Wang et al. 2020). However, such firm detection seems to be rare (e.g. Zurlo et al. 2020), and we are not expecting to directly detect a planet whose mass is less than one Jupiter mass (Ruane et al. 2017). There are other promising methods to detect young planets in discs, such as using disc kinematic features influenced by the planet (Perez et al. 2015; Pinte et al. 2018; Teague et al. 2018; Izquierdo et al. 2021; Rabago & Zhu 2021). However, all these methods can only detect planets that are more massive than Jupiter. The only method that can detect planets less massive than Jupiter is to use gaps in the dust continuum images (e.g. Zhang et al. 2018). Since dust particles drift to the local pressure maximum, even

small gas perturbations by a low-mass planet can lead to observable dust gaps (Paardekooper & Mellema 2006; Zhu et al. 2014; Rosotti et al. 2016; Dipierro & Laibe 2017; Dong et al. 2018).

Hydrodynamical simulations with dust particles have been carried out to infer the planet properties from radio observations. At early times when Atacama Large Millimeter/submillimeter Array (ALMA) high-resolution observations were scarce, numerical simulations were fine-tuned to fit one source at a time (e.g. Dipierro et al. 2015; Dong, Zhu & Whitney 2015; Picogna & Kley 2015; Dipierro et al. 2018). However, when high-angular resolution surveys became available (e.g. Andrews et al. 2018; Long et al. 2018; Cieza et al. 2021), it was impractical to run direct numerical simulations for each source. Thus, relationships between the gap and planet properties have been solved. For gaseous gaps, such relationships have been well characterized (Fung, Shi & Chiang 2014; Kanagawa et al. 2015, 2016). However, the relationships are less clear for dusty gaps mainly because the gap width/depth can vary significantly with different sized particles in discs. Lodato et al. (2019) assumed that the dusty gap width is scaled with the planet Hill radius to derive the masses of planets in the Taurus survey (Long et al. 2018). Rosotti et al. (2016) carried out simulations to derive the relationship between the planet mass and the distance between the planet and the pressure maximum at the outer gap edge. However, Rosotti et al. (2016) did not consider the effects of particle size and disc viscosity, both of which can change the gap–planet relationship significantly. A detailed study was done by Zhang et al. (2018), who carried out a large grid of

* E-mail: shangjia.zhang@unlv.edu

planet–disc interaction simulations with dust particles, generated synthetic observations, and derived relationships between the planet mass, the disc scale height, the disc viscosity, and the particle size in disc.¹ Such relationships have been used to derive the young planet population from the DSHARP survey (Andrews et al. 2018; Zhang et al. 2018).

However, the approach in Zhang et al. (2018) still suffers several shortcomings. First, the procedures to derive the planet mass are relatively complicated. Disc viscosity, scale height, and particle size need to be constrained by other methods and specified beforehand. Based on these parameters, different fitting formulae need to be adopted. To make the method in Zhang et al. (2018) easier to use, parallel to the method presented in this paper, we provide new PYTHON programs for the Zhang et al. (2018) linear fitting method,² which can automatically find the planet mass after the parameters are specified. The second shortcoming, which is intrinsic to the method itself, is that the synthetic images generated from simulations were azimuthally averaged before deriving the relationship between the planet mass and gap widths/depths. Thus, all information from non-axisymmetric features was lost. These features have rich information on disc and planet properties. For example, an eccentric gap suggests a massive planet. A lopsided disc with a large intensity asymmetry indicates a low viscosity or large particles.

To directly extract information from two-dimensional (2D) images, in this work, we adopted machine learning techniques. Machine learning techniques have been widely used across the astronomical community for decades, and here we briefly describe several works related to computer vision image tasks, using convolutional neural networks (CNNs), specifically.

CNNs have been used widely in galactic and extragalactic studies. Dieleman, Willett & Dambre (2015) is one of the first works that use modern CNN networks. It originated from an international competition *Galaxy Challenge* that aimed to build automated tools for galaxy morphology classification based on annotated images from the Galaxy Zoo project. They applied CNNs to the morphological classification of crowd-sourcing annotated images and achieved >99 per cent accuracy, which would benefit analysis in future large galaxy surveys such as Vera C. Rubin Observatory. Hezaveh, Perreault Levasseur & Marshall (2017) used CNNs to automate analysis of strong gravitational lenses. The traditional method with maximum likelihood modelling requires human expertise and is time consuming. The network can quantify image distortions caused by strong gravitational lensing and estimate these structures' corresponding matter distribution with comparable accuracy as the traditional method, but 10 million times faster. Non-experts can quickly obtain lensing parameters for a large sample of data. Hassan et al. (2019) utilized CNNs to identify reionization sources from 21 cm maps. Active galactic nuclei and star-forming galaxies are two leading sources that reionized our Universe. CNNs were trained to distinguish the sources on the 21 cm images. The technique would aid power-spectrum observations and provide extra information to break degeneracies between a broad range of reionization models. The classification accuracy is between 92 and 100 per cent, depending on the redshift and neutral fraction range.

CNNs have also been used in the Solar system study. Lieu et al. (2019) trained CNNs on simulated observations of an upcoming mission *Euclid* for Solar system objects' identification. They used

transfer learning (i.e. training on several established CNN architectures with some modifications) on a relatively small data set. Their best model correctly identified objects with a top accuracy of 94 per cent, successfully separating Solar system objects from other astronomical sources.

In the field of star formation, identifying signatures of stellar feedback in molecular clouds used to mainly rely on visual inspection. Van Oort et al. (2019) ran three-dimensional (3D) magnetohydrodynamic (MHD) simulations with stellar feedback and produced 2D synthetic CO continuum images. They trained CNNs on synthetic data and identified shells in real observations. Later, they extended the work to 3D so that they could make full use of molecular line spectrum datacube. They found stellar feedback bubbles and predicted feedback properties (Xu et al. 2020a), and identified 20 new outflows (Xu et al. 2020b) that were missed by previous visual inspections.

While CNNs have not been applied to the planet–disc interaction study, general machine learning techniques are receiving more attention in the field. Recently, Auddy & Lin (2020) used fully connected neural networks to fit the relationship between the planet mass and parameters such as gap width, aspect ratio, viscosity, dust-to-gas ratio, Stokes number, and density profile. Compared to previous fitting methods, the work is the first to fit the relationship non-linearly. Since a deep neural network is good at fitting problems that are intrinsically non-linear, their estimated planet mass follows closer to the simulation data given a multidimensional input. Nevertheless, users still need to provide inputs that are barely constrained from observations, and asymmetric information is still lost when 2D images are converted to one-dimensional (1D) radial profiles. While we were modifying this paper after the first referee report, Auddy et al. (2021) published a CNN approach for the 2D images, which alleviated some of these shortcomings. Compared with this paper, they used similar neural networks with hydrodynamical simulations. Their models allow additional disk parameters as inputs. However, the models were trained on disk density contours in simulations instead of synthetic observational images.

Our aim in this paper is to infer the planet mass from the 2D observational images directly.³ In Section 2, we briefly introduce the background and the basic glossary of CNNs. In Section 3, we describe the simulation set-up, synthetic observation production, preprocessing, augmentation, and the neural network set-up and training. In Section 4, we analyse the results, apply the networks to several gaps in DSHARP observations, and compare the derived planet masses to those from the previous method. After a short discussion in Section 5, we conclude our paper in Section 6.

2 BACKGROUND

2.1 Regular neural network

The most common subset of deep learning is regular neural networks (or fully connected neural networks). Neural networks receive a vector and transform it through a series of hidden layers. Each hidden layer is made up of a set of neurons, where each neuron is fully connected to all neurons in the previous layer. The last fully connected layer is called the 'output layer'. For classification problems, it gives scores for different classes. For regression problems, it predicts continuous values.

¹We use 'the linear fitting method' to denote this method throughout this paper.

²The code is available at <https://github.com/zhangsj96/DSHARP.VII.git>.

³The code is available at <https://github.com/zhangsj96/PGNets.git>.

Each neuron has some parameters to be tuned, which can be accomplished by training the model. The training is a process of minimizing the loss function and updating parameters through back-propagation. The data are separated as training, validation, and testing sets. The training data are used to feed into the neural network. The validation data are not used in training the model but are used as a metric to monitor the training result at every epoch. The testing data are used to evaluate the model accuracy after the training is completed.

2.2 CNN

For regular neural networks, the input vector could be extremely large if the input is an image that is represented by either a flattened matrix (e.g. a greyscale image) or a flattened tensor (e.g. an RGB-coloured image). There are correlations between neighbouring pixels and different colour channels in an image, but a vector representation loses such correlations. Thus, regular neural networks are not ideal for training image data. The convolutional operation naturally takes the local connections into account. To that end, CNNs are powerful in fitting image data. As a variation of regular neural networks, their primary unit of computation is the convolutional operation instead of simple matrix multiplication. A layer of a convolutional network has neurons arranged in three dimensions: width, height, and depth. A convolutional kernel will be operated on this 3D tensor, and the output becomes the next layer. Usually, the network's width and height become smaller for later layers, while the depth becomes deeper. The process of downsampling the feature map is called pooling. LeCun et al. (1998) introduced LeNet to recognize hand-written digit characters. It reached a very high performance and brought artificial neural networks into popularity.

2.3 Residual neural network

A residual neural network (ResNet) is a kind of CNN that has connections even between skipping (non-neighbouring) layers (He et al. 2016b). It has shortcuts to jump over some layers. Typical ResNet models are implemented with double or triple skips that contain Rectification Linear Unit (ReLU) and batch normalization (Ioffe & Szegedy 2015) in between. Skipping effectively simplifies the networks and reduces the parameters. It also avoids the problem of vanishing gradients so that the network can go deeper than traditional CNNs while still improving the performance.

3 METHOD

A schematic view of our method from input to output is summarized in Fig. 1. We first introduce how we convert the simulations to synthetic observations (Section 3.1), and then discuss the pre-processing (Section 3.2) and augmentation (Section 3.3) steps to make the network robust. We layout the Planet Gap neural network (PGNet) structures (VGG-like or ResNet classification, Section 3.4; regression, Section 3.5) and finally obtain the output prediction.

At the output layer, we tried both classification and regression problems. At first, we treated the fitting of the planet mass as a classification problem, since the data set of Zhang et al. (2018) is too sparse (only five discrete planet masses) to return continuous-valued predictions. They were used to demonstrate that CNNs can successfully predict planet masses on discrete grids. Then, we ran additional 150 simulations to provide more sampling between these grids, and built regression model that predicts continuous planet mass and disc viscosity at the same time. The samples were drawn using the Latin hypercube sampling (LHS; McKay, Beckman & Conover 1979).

3.1 Simulations

For the classification problem, we used the results of the planet-disc interaction simulations in Zhang et al. (2018) on the gridded parameter space (five planet masses, M_p , three aspect ratios, h/r , and three disc viscosities, α). We denote a model with an $\{M_p, h/r, \alpha\}$ pair as a generic model, as this model can be used to generate models with different surface densities and maximum particle sizes. For the regression problem, we added additional 150 simulations with near-randomly generated M_p , h/r , and α . We briefly summarize the simulations here. The simulations were carried out with 2D hydrodynamic code FARGO-ADSG (Baruteau & Masset 2008a, b; Baruteau & Zhu 2016). Dust grains were represented by 200 000 superparticles with different sizes. The Stokes number (St) of the particles at r_p ranged from 1.57×10^{-5} to 1.57. The simulations in the gridded parameter space covered three disc viscosities $\alpha = 10^{-4}$, 10^{-3} , and 10^{-2} , three disc aspect ratios at r_p with $h/r = 0.05, 0.07$, and 0.1, and five planet masses with the planet-star mass ratios (q) of 3.3×10^{-5} , 10^{-4} , 3.3×10^{-4} , 10^{-3} , and 3.3×10^{-3} (which are equivalent to the planet masses of $11 M_\oplus$, $33 M_\oplus$, $0.3 M_J$, $1 M_J$, and $3 M_J$ if the central star is a solar-mass star). The parameters for the LHS were drawn from $\alpha \in [10^{-4}, 10^{-2}]$, $h/r \in [0.05, 0.1]$, and $M_p \in [11 M_\oplus, 3 M_J]$. They were near-uniformly drawn in the interval of h/r , $\log(\alpha)$, and $\log(M_p)$.⁴ We initialized the gas surface density as

$$\Sigma_g(r) = \Sigma_{g,0}(r/r_0)^{-1}, \quad (1)$$

where r_0 is the position of the planet and we set $r_0 = r_p = 1$. Our numerical grid extended from $0.1r_0$ to $10r_0$ in the radial direction and 0 to 2π in the θ direction. The data sets used in this paper are at 1000 planetary orbits. We assumed locally isothermal equation of state. The temperature at radius r follows $T(r) = T_0(r/r_0)^{-1/2}$.

To convert superparticle distributions to optical depth maps, we used a subset of particles and gave them different weights depending on their opacity, sizes, and surface density at their locations. We adopted DSHARP opacity (Birnstiel et al. 2018) and neglected scattering. These particles were interpolated on to 1200×1200 regular grid with physical dimensions as $10r_p \times 10r_p$. Then, we smoothed the gridded data with a Gaussian kernel similar to the resolution of ALMA. Gaussian $\sigma = 2$ pixels, which is $0.03r_p$. If the planet is at 20 au, this is the same as the resolution of a circular beam with full width at half-maximum (FWHM) = 1.4 au. Finally, we calculated the brightness temperature or intensity for each grid cell as

$$T_b(x, y) = T_d(r)(1 - e^{-\tau(x, y)}), \quad (2)$$

and we assumed that the mid-plane temperature follows

$$T_d(r) = T_d(r_0) \left(\frac{r}{r_0} \right)^{-0.5}. \quad (3)$$

For more details, please see Zhang et al. (2018).

In principle, the real synthetic image should be calculated with assumptions of detailed observational set-ups (e.g. antenna array configuration and integration time) and CLEAN methods. However, this is unrealistic since (a) they depend on specific observational set-ups and data reduction methods, which cannot be covered thoroughly, and (b) these time-consuming steps need to be applied on each augmented image (Section 3.3), which cannot be realized within our computational power. Instead, in Section 5.1 we will use one

⁴We used pyDOE (<https://github.com/tisimst/pyDOE>) to generate parameters. The generated parameters can be found at <https://github.com/zhangsj96/PGNets.git>.

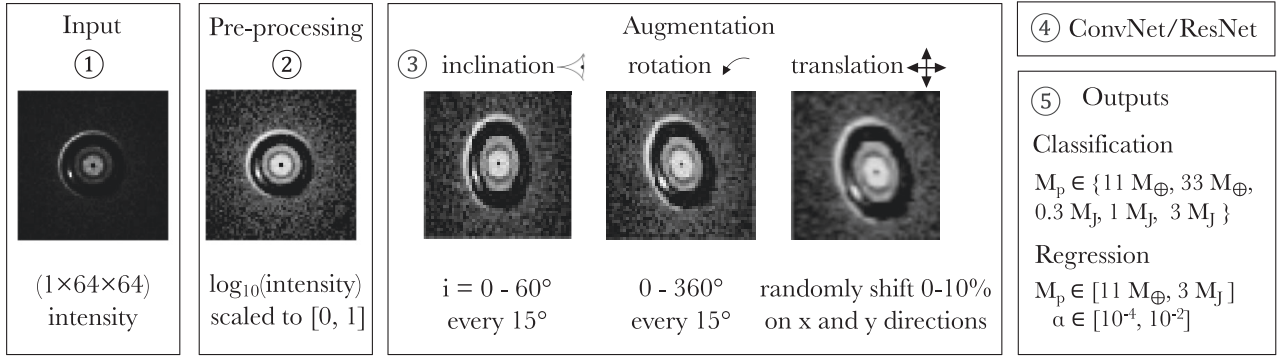


Figure 1. A schematic view of our work from input to output. There are five steps. First, we prepared the input from simulations or observations. Secondly, the image was preprocessed and normalized. Thirdly, the image was augmented for different inclinations, rotations, and translational shifts. Then, it flowed into the neural network. Finally, we obtained class scores and chose the label with the highest score for the classification problem. A planet mass, viscosity pair was returned for the regression problem.

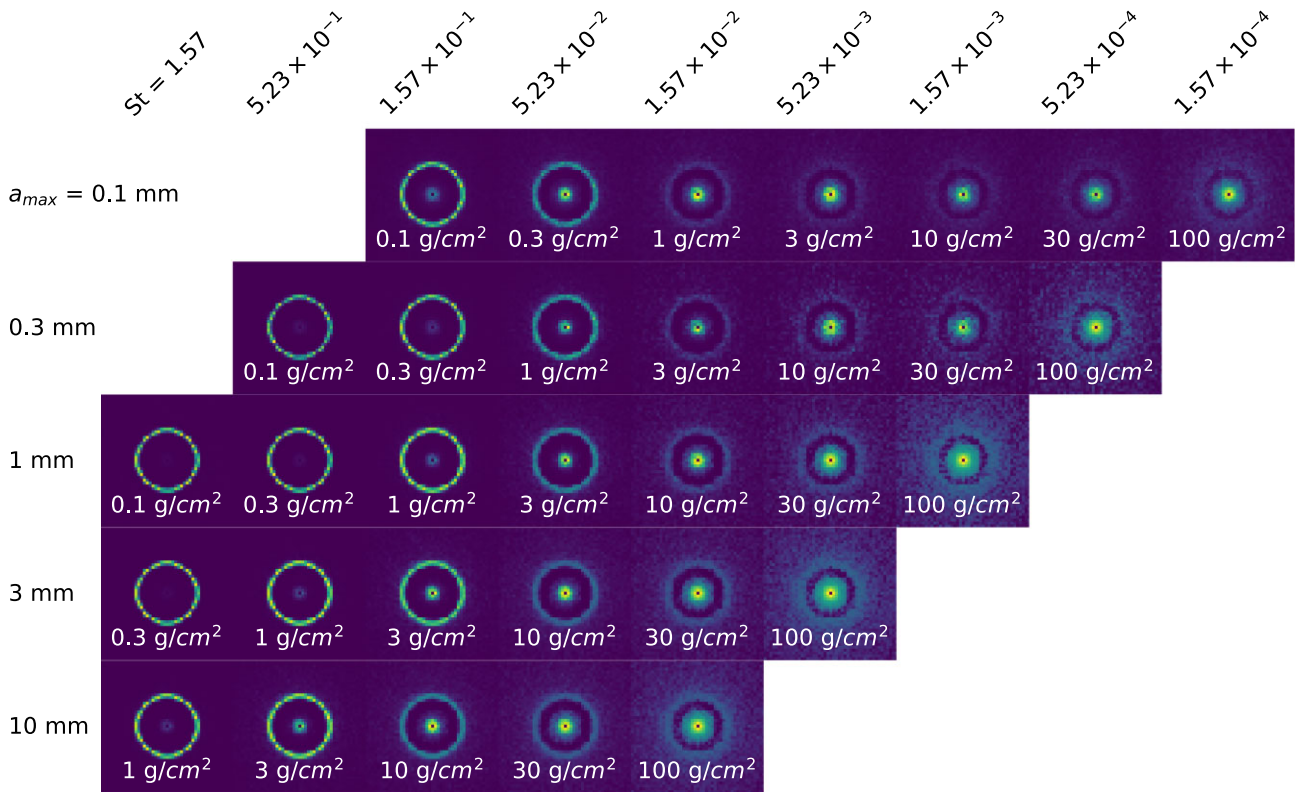


Figure 2. 1.3 mm dust continuum intensity maps for the $M_p = 1 M_J$, $\alpha = 10^{-3}$, and $h/r = 0.07$ case with different a_{\max} and $\Sigma_{g,0}$. In each row, a_{\max} is the same with increasing surface density from left to right. The Stokes number at the planet's location is the same at each column and decreasing from left to right.

case to demonstrate that neglecting these steps does not affect the correctness of the prediction.

To generate synthetic radio continuum images of discs having various dust size distributions and disc surface densities, we chose five different maximum particle sizes $a_{\max} = 0.1 \text{ mm}, 0.3 \text{ mm}, 1 \text{ mm}, 3 \text{ mm},$ and 1 cm and seven different gas surface densities $\Sigma_{g,0} = 0.1, 0.3, 1, 3, 10, 30,$ and 100 g cm^{-2} . These combinations of a_{\max} and $\Sigma_{g,0}$ correspond to nine different characteristic Stokes numbers ($St \propto a_{\max}/\Sigma_{g,0}$). Fig. 2 shows dust continuum intensity at 1.3 mm for a case with $M_p = M_J$, $\alpha = 10^{-3}$, and $h/r = 0.07$, with all combinations of a_{\max} and $\Sigma_{g,0}$. Note that for $a_{\max} = 3 \text{ mm}$ (and 1 cm), the lowest (two) surface density case(s) exceeds the upper

limit of the particles' Stokes number in our simulation. In total, different a_{\max} and $\Sigma_{g,0}$ lead to 32 combinations ($7 + 7 + 7 + 6 + 5$) for each generic model. Since no dust growth and back-reaction are included, the dust drift velocity only depends on the Stokes number. This is why the gaps look similar for a given Stokes number. They are still different in that (a) the opacity is dependent on the maximum particle size and (b) the dust surface density is different. Thus, the radial profile of the optical depth is different. When the disc changes from the optically thin to thick regime (from top to bottom panels), the intensity maps become smoother.

We explored different dust size distributions by choosing three different power-law indices [p ; $n(a) \propto a^{-p}$] being 3.5, 3, and 2,

but their intensity maps are very similar. This could lead to some duplication between training and testing sets. For this reason, we only used one power-law index $p = 3.5$.

An image with too many pixels takes a large amount of memory, does not help the training, and even leads to overfitting. Thus, we downgraded the image size and found that a 64×64 image is good enough to preserve features and help the training. Note that the real observation image will also be fed into the network using this resolution. To convert an observation to the input format, a simple linear interpolation suffices. For our models, we selected the central squared $3r_p \times 3r_p$ region. Outside of this region, there is little emission from our simulations. The data are 2D (grey colour, $1 \times 64 \times 64$) images of disc emission. There are $5 \times 3 \times 3 \times 32 = 1440$ models for the classification problem and $(45 + 150) \times 32 = 6240$ models for the regression problem.

3.2 Pre-processing

We found that adding some noise to images helped the training, so we added 10 per cent RMS noise to each image. The initial value of the intensity in the image can be different by orders of magnitude. In neural networks, it is essential to normalize the input data. We found that taking the logarithm of the value works best for the training. We scaled the image values from 0 to 1. We also set an emission floor as 1 per cent of the maximum emission intensity, so that values below that became zero. In expression,

$$\text{pixel value} = \begin{cases} \frac{\log_{10}(I) - \log_{10}(I_{\max}) + 2}{2} & \text{if } \log_{10}(I) > \log_{10}(I_{\max} - 2), \\ 0 & \text{elsewhere.} \end{cases} \quad (4)$$

This choice of the floor value can also be justified for observational data since the sensitivity limit for these high-resolution ALMA observations is usually between 1 and 0.1 per cent. The gap region has low values, and the ring region has high values. Considering that previous studies used gap shapes to infer the planet mass, we reversed the value as (1-value) so that the gap had high values. It turned out that this procedure had little effect on the results. Since this procedure is well defined, it is also simple to apply it to the real observation and rapidly convert that to the network input. Note that our main focus is on the gap shape (width and depth, and asymmetric features). The normalization process we adopted removed the information of the absolute value of the emission. Ideally, the absolute value of the emission can provide extra information and can be considered in future works.

3.3 Data augmentation

In real observations, discs are hardly to be exactly face-on (e.g. the configuration in Fig. 2). As long as the gap is spatially resolved and the disc is not too edge-on, we should still obtain information from the image. Thus, we inclined the disc from the initial data from 0° to 60° , spacing every 15° . This was simply done by stretching the y -direction of the optical depth map, $\tau(x, y)$, while fixing the x -direction size. The optical depth is increased accordingly. This can be justified considering that mm dust is highly settled in protoplanetary discs (Pinte et al. 2016). With this data augmentation, the observational data for discs with any inclinations can be directly used as an input for our networks.

To explore the rotational symmetry (Dieleman et al. 2015), we rotated the disc every 15° , from 0° to 360° (0° and 360° are only counted once). The degree of rotation is also called position angle. Note that the rotation and inclination variations were done on the

original 1200×1200 ($10r_p \times 10r_p$) data to avoid the sharp edges due to missing values, but whether including this step did not affect the results.

We also explored translational symmetry by randomly shifting the image in x - and y -directions from 0 to 10 per cent. Even though CNNs should conserve translational symmetry by themselves, the fitting result was slightly improved. This is also helpful for the preprocessing of observational data since the disc does not need to be perfectly centred when they are fed into the networks.

Every one of the face-on images can generate thousands of augmented data. However, we still separated the training, validation, and testing sets with those original data. Otherwise, if we separate all the data after the augmentation, the fitting accuracy will actually increase since different augmented data generated from a single model can look very similar. However, this increase of the accuracy is unreal due to the similar data in training, validation, and tests. Thus, we separated the data sets before the augmentation.

3.4 Classification models

We first tested out CNN models as a classification problem on the gridded data set in Zhang et al. (2018). We set up the network as a classification problem instead of a regression problem since our original simulations only have five discrete planet masses. We built two neural networks for comparison, one architecture similar to VGG-16 (Simonyan & Zisserman 2014) and the other adapted from ResNet (He et al. 2016b). We chose not to use the original VGG-16 architecture since it introduced more parameters but did not improve the performance.

Fig. 3 shows the architecture of VGG-like network. As mentioned in the previous section, the input data are grey colour images ($1 \times 64 \times 64$), with values ranging from 0 to 1. Then, it follows two convolutional layers with 16 channels (depth = 16). In this specific neural network, the convolutional filter size is always 3×3 with 1 stride. The padding is on image edges so that the tensor keeps the same number of points (64) in both x - and y -directions. We also used ReLU (i.e. $\max[0, x]$), as the activation function for all layers. After a 2×2 max-pooling operation, the image dimension becomes 32×32 . Then, it follows two convolutional layers with 64 channels and another max-pooling. The last two convolutional layers have 256 channels before a max-pooling. Then, it connects to three fully connected layers, two with 512 nodes and one with 128 nodes. These fully connected layers all have 20 per cent dropout rates to avoid overfitting. Finally, the output layer has five nodes that stand for five different planet masses. There are 9504 565 trainable parameters in total. We found that adding more layers did not increase the accuracy.

The learning rate is a hyperparameter that controls the step size in an optimization process while it is minimizing the loss function. An analogy of it is the step size in solving differential equations. We adopted the initial learning rate as 0.001. We chose the sparse categorical cross-entropy as the loss function and ADAM as the optimizer (Kingma & Ba 2015). We defined accuracy of the method as

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (5)$$

where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively.

To balance the trade-off between rate of convergence and over-shooting, we adopted an adaptive learning rate. The learning rate varying as the training epoch is a cosine function so that the learning

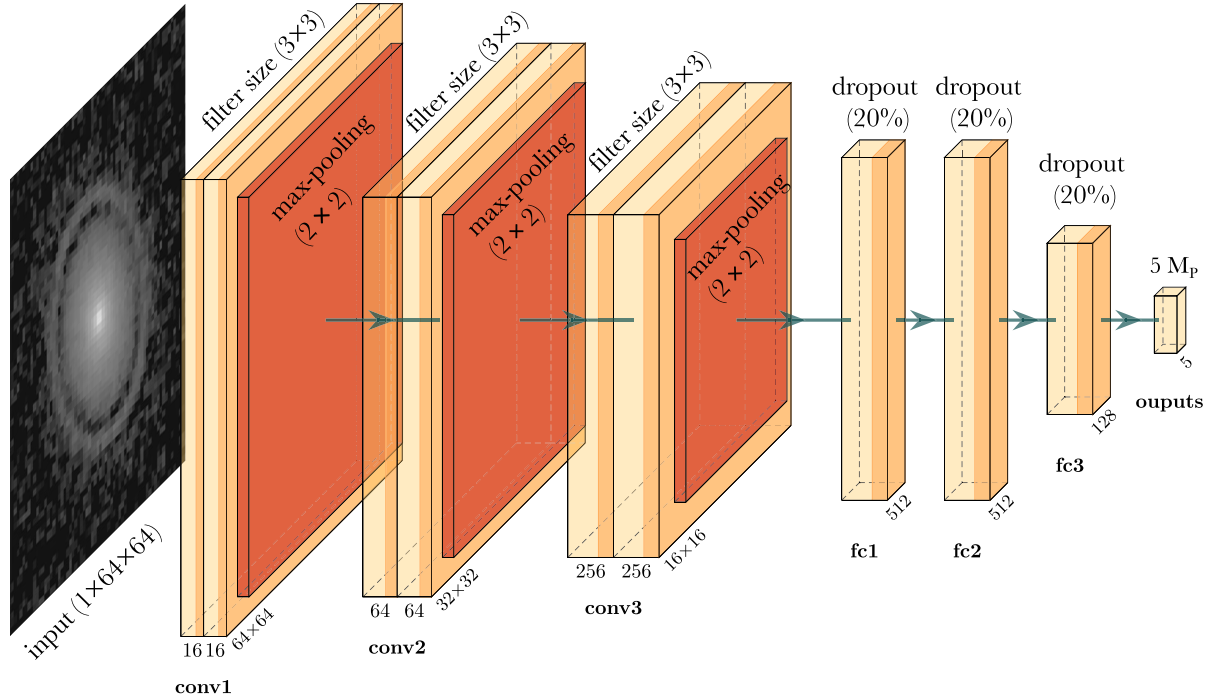


Figure 3. The architecture of VGG-like network. The image as an input is the \log_{10} of the intensity, scaled between zero and one with only one channel. The example image is pre-processed image of GW Lup (Andrews et al. 2018). Then, there are six convolutional layers and three fully connected layers.

rate becomes much smaller at later training stages. The batch size is 128. The training, validation, and testing sets were split as 60 per cent, 20 per cent, and 20 per cent of the 1440 models that were randomly shuffled. We also included different inclinations and rotations for the testing set, so there are total $1440 \times 0.2 \times 5 \times 24 = 34\,560$ testing images. The training was done on a single NVIDIA GPU GeForce RTX 2080 Ti 12GB. We used `Tensorflow v2` (Abadi et al. 2015) to build the network. The VGG-like network took less than 1 h, while the ResNet model took 2 h.

For the residual network, our network was adopted from ResNet v2 (He et al. 2016a) with three skips. It has a bottleneck layer with stacks of batch normalization, activation (ReLU) and convolutional layers. The batch normalization can make ANN faster and more stable. The first shortcut connection per layer is 1×1 convolution and the second and onward shortcut connection is identity. At the beginning of each stage, the feature map size is halved by a convolutional layer with 2 strides, while the number of filter maps is doubled. Within each stage, the layers have the same number of filters and the same filter map sizes. There are 22 convolutional layers in total, but with only 576 357 parameters, an order of magnitude smaller than the VGG-like model. The initial learning rate is 0.005. Other set-ups are the same as the VGG-like model.

3.5 Regression

After carrying out additional simulations with the randomly generated parameters from the LHS, we were able to fit the planet mass in a continuous space. Thus, we fitted the combined data set as a regression problem. We used mean square error (MSE) as the loss function. We made the output layer as an M_p - α pair, as MSE can evaluate vectors (e.g. Alibert & Venturini 2019). In this way, the planet mass and disc viscosity can be predicted together. We made viscosity as an output since a direct measurement of gas

turbulence is difficult (e.g. Flaherty et al. 2018). We adopted the ResNet architecture as mentioned in the previous subsection.

Slightly different from Section 3.4, we split 195 generic models into 175 and 20. The selection was random but can be reproduced in our repository. We then separated the first 175 models, together with their generated models with combinations of surface densities and maximum particle sizes into 60 per cent, 20 per cent, and 20 per cent splits as training, validation, and test data. Finally, we used the rest 20 generic models as a ‘genuine test’, as the input parameters of these models are completely outside the parameter space of $\{M_p, h/r, \alpha\}$ pairs in the training. This additional step can test whether the overfitting happens in the $\{M_p, h/r, \alpha\}$ space. This is a more rigorous test than that in Section 3.4 since we separate the data even before generating the face-on images with different a_{\max} and surface densities. We used callback function `ReduceLROnPlateau` to reduce learning rate when the loss had stopped improving. We kept other hyperparameters the same as what were mentioned in previous subsections.

4 RESULTS

4.1 Classification

The accuracy of VGG-like model reached a plateau after 10 training epochs but still slowly increased up to 40 epochs. The accuracy of the ResNet model reached a plateau after 90 epochs. The ResNet has slightly higher accuracy than the VGG-like model on the validation set. This is also the case for the testing set. The VGG-like model can reach an accuracy of 89 per cent, and the ResNet model can reach an accuracy of 92 per cent. Note that the accuracy reported is the micro accuracy [i.e. applying equation (5) for the whole sample, instead of calculating accuracies for each class and averaging them, which is the macro accuracy]. Even though each class (planet mass) has the same amount of sample in the whole data set, in each subset (training,

ground truth	11 M_{\oplus}	6708 87%	830 11%	139 2%	3 0%	0 0%	7680 22.2%
	33 M_{\oplus}	280 5%	4939 89%	274 5%	26 0%	1 0%	5520 16.0%
	0.3 M_J	190 3%	355 5%	6851 91%	162 2%	2 0%	7560 21.9%
	1 M_J	2 0%	4 0%	106 2%	6561 96%	167 2%	6840 19.8%
	3 M_J	0 0%	0 0%	12 0%	176 3%	6772 97%	6960 20.1%
		prediction					
		11 M_{\oplus}	33 M_{\oplus}	0.3 M_J	1 M_J	3 M_J	

Figure 4. The confusion matrix for the ResNet model. The x-axis is the prediction from the neural network. The ground truth is on the y-axis. The upper number in a box shows the counts of planets with certain prediction and ground truth. The lower percentage shows the fraction of the prediction over the total number of the sample with certain ground truth (sum of a given row). The rightmost numbers are total counts of testing data with certain class labels (sum of a given row) and the percentage among all testing data (sum of the rightmost column).

validation, and testing sets), the data are slightly imbalanced due to the shuffling process to separate them. Nevertheless, macro and micro metrics are similar here.

Fig. 4 shows the confusion matrix of planet mass prediction, and their ground truth for all testing data set applied to the ResNet network. The y-axis shows the true planet masses, whereas the x-axis shows the most likely planet masses predicted by the network. If the data point falls on the matrix diagonal, the prediction is correct. The accuracy is the sum of diagonal counts over the total counts. The planet mass is underestimated if it is on the left and overestimated if on the right. The upper number in each box shows the counts of a certain prediction given a ground truth. The lower percentage shows the fraction of it among all the images within that class. When the planet mass is small, i.e. the gap is narrow, the prediction accuracy is low. If the planet is 11 M_{\oplus} , only 87 per cent of the samples account for correct predictions within that class. The planet mass is always overestimated given this set-up. However, we caution that the planet mass can also be underestimated in reality, since this neural network cannot find planet with mass lower than this limit. The prediction accuracy increases with higher planet masses. If the planet mass is 3 M_J , 97 per cent of the predictions within that class are correct. Likewise, 3 M_J is the upper mass limit in this network, so any real planet mass with higher value will be underestimated.

4.2 Regression

The loss of the regression model reached a floor after 70 training epochs. Fig. 5 shows histograms for the deviation of the predicted planet masses and disc viscosities from the true values in dex. 2D histograms of the joint distribution between M_p and α are plotted at centre, whereas the 1D distributions of the planet mass and viscosity differences are shown on the top and right. Panel (a) shows the result of the test set among the 175/195 generic models that were used in the training. The distributions of M_p and α are all centred close to zero, with uncertainties of 0.16 and 0.23 dex, respectively. They are symmetric with almost zero means. Planet masses and the disc

viscosities tend to be over(under)estimated at the same time and follow $\alpha \propto M_p^3$, which has also been found in Zhang et al. (2018). Panel (b) shows the result of 20/195 generic models that are not used in the training. The overall distributions are similar to those in panel (a). The exception is that the deviation of α is skewed towards positive values (i.e. α tends to be overpredicted). This is because most of the viscosities among these randomly generated 20 generic models have low α (close to 10^{-4}), where it is more likely to be overestimated (see Fig. 6). The similarity between panels (a) and (b) demonstrates that the fitting is robust and can also be applied to simulations not used in the training.

Fig. 6 shows distributions of the difference of the prediction and ground truth in different mass and viscosity regimes. The data are from the test set of 175/195 generic models. The distributions are normalized to have the same height, and the inner box follows the convention of the box plot. The standard deviation of each distribution is listed on the right of each violin plot. The planet masses are divided into five bins in comparison with the classification problem (Fig. 4). Similar to the classification problem, the error of the fitting becomes smaller as the planet mass increases. A small fraction of samples have large errors except at the highest mass bin. At the lower mass end, the planet mass is more likely to be overestimated. At the higher mass end, the uncertainty of the estimate can be as low as 0.1 dex (a factor of 1.3). The viscosity is divided into four bins. When α is less than 3×10^{-3} , it tends to be overestimated. When it is large and close to 10^{-2} , it is more likely to be underestimated. In any viscosity regimes, a small fraction of the predictions would have large deviations from their true values.

4.3 Grad-CAM

While deep neural networks are difficult to interpret, some visualization tools help make sense of them. Gradient-weighted Class Activation Mapping (Grad-CAM; Selvaraju et al. 2017) is a technique for making CNN-based models more transparent by visualizing the regions of input that are important for predictions from these models. It uses the class-specific gradient information flowing into the final convolutional layer of a CNN to provide a coarse localization map of the important regions in the image. It is a generalization of the Class Activation Mapping (Zhou et al. 2016), but requires no retraining.

Fig. 7 shows both images and the activation map derived from the Grad-CAM of the ResNet regression network. The disc is inclined and rotated. Here, the true planet mass is 2.28 M_J , and $\alpha = 1.4 \times 10^{-3}$. The gap region has high values in the activation map, which means that it is important in predicting the planet mass. This lends confidence to the network's reliability since the traditional method also focuses on the gap's properties. This is why we named these neural networks as PGNets. One should be cautious that not every activation map of testing data shows such a good correspondence, and the activation map is a great tool if any result is in doubt.

4.4 Application to observational data

After testing against the synthetic observations, we applied our trained PGNets to real DSHARP observations (AS 209, Andrews et al. 2018; Guzmán et al. 2018; Elias 24, GW Lup, and Sz 114, Andrews et al. 2018) to infer planet masses and compared them with what was found in Zhang et al. (2018). The designation follows Huang et al. (2018), where the integer of gap location in au follows 'D' (for dark gaps). There are several degeneracies if one wants to infer the planet mass. In the previous fitting of Zhang et al. (2018), the

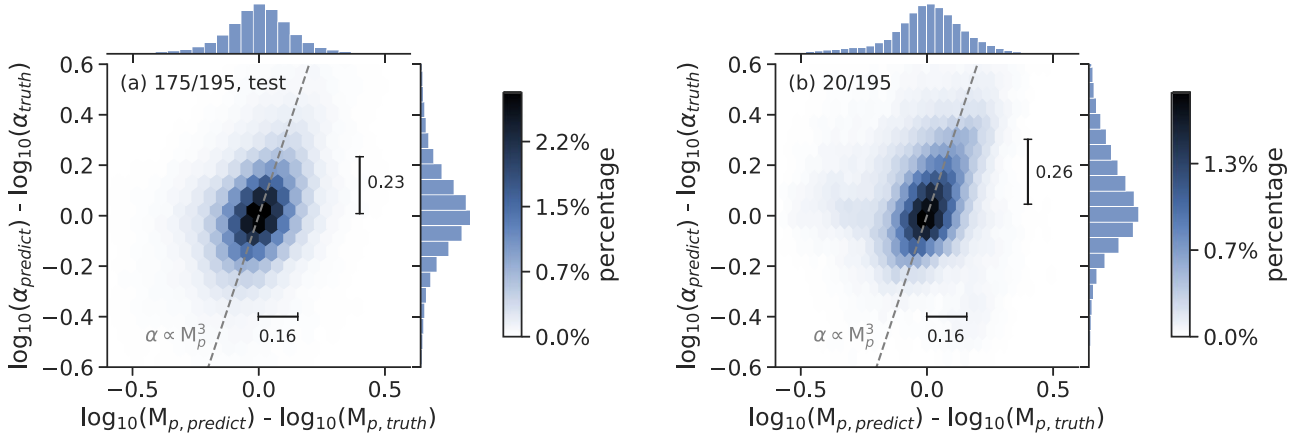


Figure 5. 1D and 2D histograms for the differences between predicted and true values of M_p (horizontal) and α (vertical). The 2D joint distributions are colour coded. (a) The test data set among 175/195 of the generic models. (b) The 20/195 data set that has not been used in the training process. Most of the predictions have small deviations. The horizontal and vertical bars represent the standard deviation of each distribution. Dashed lines represent $\alpha \propto M_p^3$.

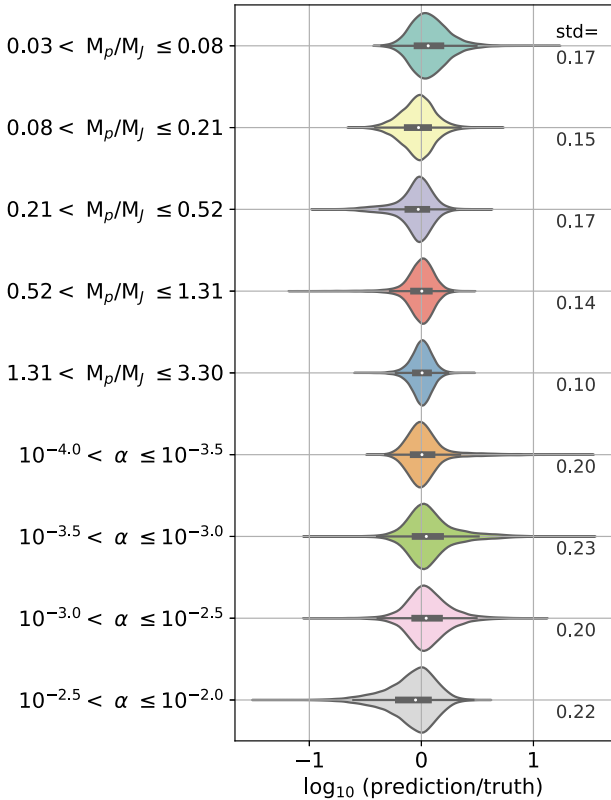


Figure 6. The violin plots of logarithmic-scaled deviation between predicted and true values of M_p and α in different mass and viscosity regimes. The data are from the test set of 175/195 generic models. The colour-shaded regions show the distribution with normalized height, whereas the inner box follows the convention of a box plot, which shows distribution's 25 percentile, 50 percentile (white dot), and 75 percentile. The decimals marked on the right are the standard deviation of each distribution.

gap width depends on not only the planet mass but also the viscosity α and Stokes number St (or a_{\max} and Σ_g). Thus, in comparison with Zhang et al. (2018), we picked the median value as a reference point. The median value is the case with $\alpha = 10^{-3}$ and $a_{\max} = 1$ mm. It is difficult to estimate the uncertainty in these models. For the linear fitting method, two known uncertainties come from α and a_{\max} . A

difference of 10 in α leads to around 0.33 dex change of M_p , whereas a change of 10 in a_{\max} leads to around 0.2 dex change of M_p . For classification models, an error of reference can be 0.4 dex, since two neighbouring planet masses are spaced at that value. For the regression model, the uncertainty of M_p and α can be estimated as 0.16 and 0.23 dex, respectively, as shown in Fig. 5, or read in specific regimes as shown in Fig. 6. For AS 209 and Elias 24, fine-tuned planet masses are available from detailed case-by-case modellings. While the central stars of these discs have different masses, what matters is the planet–star mass ratio M_p/M_* . Thus, we will report M_p/M_* in units of the M_J/M_\odot in the next paragraph. Notice that the radius of the semiminor axis of the gap should be placed around one-sixth of the image size. However, the disc does not need to be centred nor scaled perfectly, since the images were randomly shifted during the training process.

AS 209 is a disc with many gaps. With the linear fitting method, the median value of AS 209 D99⁵ is $0.45 M_J/M_\odot$ (Zhang et al. 2018). The detailed modelling infers $0.1 M_J/M_\odot$. The VGG-like model predicts $0.1 M_J/M_\odot$, whereas the ResNet classification model predicts $0.3 M_J/M_\odot$. The regression model predicts $0.52 M_J/M_\odot$ and $\alpha = 2 \times 10^{-4}$. With 64×64 resolution, gaps inside 40 au have been smoothed out. The secondary gap (a shallower gap inside the major gap) has already been considered since there are also many low-viscosity simulations with secondary gaps in our training data set.

With the traditional fitting method, Elias 24 D57 is $0.51 M_J/M_\odot$. It is $0.2 M_J/M_\odot$ using the fine-tuned model. Both VGG-like and ResNet classification models predict $0.1 M_J/M_\odot$. The regression model predicts $M_p/M_* = 0.21 M_J/M_\odot$, and $\alpha = 2 \times 10^{-4}$.

The inferred planet mass from the traditional fitting method for GW Lup D74 is $0.065 M_J/M_\odot$. Both VGG-like and ResNet classification models predict $0.03 M_J/M_\odot$. The regression model predicts $M_p/M_* = 0.05 M_J/M_\odot$, and $\alpha = 1 \times 10^{-4}$.

Sz 114 is a disc with a very narrow gap. It is too narrow to use the linear fitting method. The mass was obtained by directly comparing with simulations (Zhang et al. 2018). That value for Sz 114 D39 is $0.12 M_J/M_\odot$. Both VGG-like and ResNet models predict 0.03

⁵It is actually B99 ('B' stands for bright rings). However, that bright ring is just a shallow peak inside a wide gap.

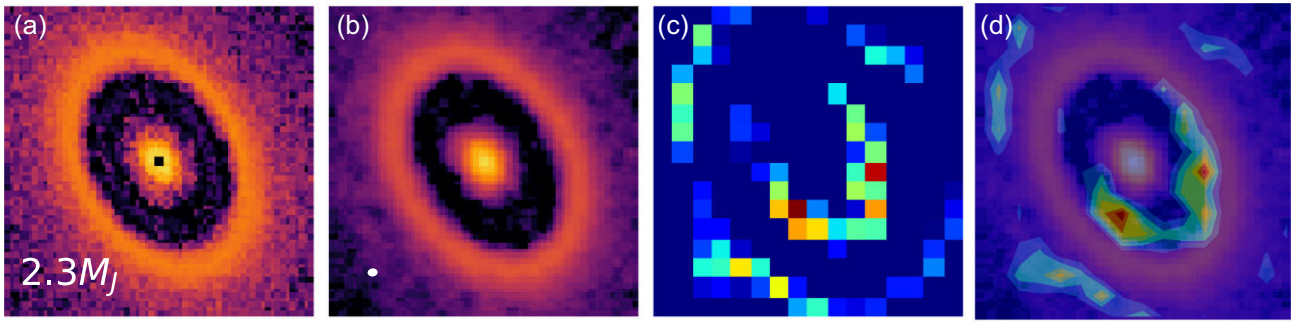


Figure 7. Grad-CAM for the ResNet regression problem. The image is among 20 generic models not in the training set. The true planet mass is $2.28 M_J$ and $\alpha = 1.4 \times 10^{-3}$. (a) Log-scaled image without putting into CASA, (b) log-scaled synthetic image using CASA, (c) activation map of the synthetic image (b), and (d) synthetic image overlaid by the activation map (filled contour). In this case, the activation map successfully focuses on the gap region. The beam size of panel (b) is shown as an ellipse in the bottom left.

M_J/M_\odot . The regression model predicts $M_p/M_* = 0.11 M_J/M_\odot$, and $\alpha = 6 \times 10^{-4}$.

Overall, the VGG-like and ResNet classification models predict similar planet masses. The regression model predicts higher planet masses. It predicts lower viscosities for these discs.

There are many other gaps that can be compared between the linear fitting method and PGNNets. Instead of doing extensive coverage of all available gaps, we made our code available online so that users can apply these methods to any new or archival data. With a superresolution technique, Jennings et al. (2021) even found many more gaps inside 30 au for DSHARP observations. Those images and radial profiles can also be inputs for our models using either the PGNNets or the linear fitting tools.

5 DISCUSSION

5.1 Synthetic observation

To test whether predictions from CNN models are affected by the way we produce synthetic observations, we picked the case in Fig. 7(a), put it into CASA (McMullin et al. 2007) version 6.1.0, and used the `simobserve` task to generate observations with angular resolutions and sensitivities comparable to those of the DSHARP observations. We assumed that the planet is at 40 au, $a_{\max} = 1$ cm, $\Sigma_{g,0} = 30$ g cm $^{-2}$, and $L_* = 2 L_\odot$. The synthetic observations consist of 12 min of on-source integration time with the Cycle 5 C43-5 antenna configuration, 35 min on-source in the C43-8 configuration, and 35 min on-source in the C43-9 configuration. A precipitable water vapor level of 1.0 mm was adopted. The synthetic visibility was imaged in the same manner as the DSHARP sources using the `tclean` task, as described in Andrews et al. (2018).

In Fig. 7, panel (a) shows the intensity map converted from the simulation and panel (b) shows the intensity from the synthetic ALMA observation. The scales of them were normalized as discussed in Section 3.2. The angular resolution is ~ 5 au in FWHM and is marked in the lower left corner of the panel. As shown in panels (c) and (d), the activated regions correctly focus on the gap. The predicted masses are $1.92 M_J$ and $2.39 M_J$, with and without CASA operations, all comparable to the true value, $2.28 M_J$. The predicted α viscosities are 1.3×10^{-3} and 1.2×10^{-3} , with and without CASA operations, also comparable to the true value, 1.4×10^{-3} . Thus, we conclude that our models can be used to predict real ALMA observations even though they were trained on the data without doing `simobserve` tasks.

5.2 Advantages

Compared to the traditional method, CNNs have many advantages. It is quick and convenient. The 2D images do not need to be converted to 1D profiles. Deriving 1D profiles seems to be a simple task but it takes time since one needs to run Markov chain Monte Carlo to find a disc's centroid, inclination, and position angle. Once the CNN model is trained, one can obtain a prediction within a second directly from the image plane. For this reason, it can be applied to a large disc sample. The training is one time and only takes 1 h. Using a traditional method, one needs to search for the best-fitting parameters. For instance, one parameter combining viscosity, h/r , and planet mass is enough to fit the gaseous gap depth and width (Kanagawa et al. 2015, 2016). However, if the dust is included, we also need to find several fitting formulae for different a_{\max} and gas surface density (Zhang et al. 2018). This is not the case for CNNs. With more data or new physics included, the network can be retrained quickly by providing more training data.

Asymmetric information is lost when a 2D image is converted to a 1D radial profile. CNNs can preserve this information. For instance, a low-mass planet in an inviscid disc can have the same gap width as a high-mass planet in a viscous disc. The traditional method cannot break this degeneracy. However, the viscosity of the disc can be inferred from a 2D image. If the disc looks more asymmetric, it should be more inviscid. The planet mass can be better constrained accordingly.

In a regression problem, the model can provide planet mass and disc viscosity at the same time. With the linear fitting method, one can only get a planet mass by assuming a disc viscosity, since both a higher planet mass and a more inviscid disc help open a wider gap. The CNN regression model partially breaks the degeneracy between them.

The CNN methods can handle very shallow or narrow gaps, which is of great difficulty for the traditional fitting method. In Zhang et al. (2018), to make sure most of the data points fall on to the fitting line, narrow gaps with Δ below 0.15 were treated as outliers. One thus needs to compare with individual simulations to estimate a planet mass (e.g. inferring the potential planet in Sz 114's narrow gap), which is extremely time consuming.

5.3 Limitations

The first limitation of the CNN models is that training with the same data set but different architectures would lead to different predictions, even though their accuracy is similar statistically. Even given the

same architecture, different random seeds or data augmentation would also result in different predictions. On the contrary, the traditional linear fitting method provides a definite planet mass as long as a gap width and other disc parameters are provided. Many details of fitting are empirical, but they are transparent to the user.

Compared to the linear fitting method, it is difficult to understand how CNN methods derive the results, even though the Grad-CAM can qualitatively inform us the important features the network uses. In the traditional fitting method, the gap width is proportional to some powers of planet mass, disc viscosity, and h/r at a given St . We can use some other ways to constrain some parameters and then narrow down the planet mass. In classification or single-output regression CNN models, we can only get a single prediction for these parameters when we apply them to observations. Deep neural networks are intrinsically highly non-linear. For an input image, these models are more of methods that help pick up an image in a training set with the most similar feature than methods of finding intrinsic relationships. The user can use their own knowledge to interpret why the planet mass returned by a CNN model is as such (e.g. if the planet mass is low, the user can make sense of it by noticing that the gap is narrow), but the network itself cannot inform this to the user. Instead of using a linear fitting with many assumptions, it helps the user to find the most closed-match model for an input. In some sense, the classification and single-output regression models can be seen as another way to present all the simulations in Zhang et al. (2018) by providing an automated tool to find a closed match. Surprisingly, multi-output regression models can possibly solve this problem. For instance, the multi-output regression PGNNets can help us find the exact relation of the degeneracy between planet mass and disc viscosity among simulations (Fig. 5). There are hopes that we can learn valuable insights from CNN models.

Finally, compared to a detailed modelling, the prediction can only be as good as the physics included in the simulations. While our simulations span a large parameter space, they cannot cover every possible situation. For instance, we cannot predict planet masses below $11 M_{\oplus}$ or above $3 M_J$. A planet with a lower mass than $11 M_{\oplus}$ can only be predicted as massive as that. Only one planet is put into the simulation, and its orbit is fixed. Thus, the model cannot be used to study multiple planets carving a common gap. However, if planets lead to several gaps and they do not influence each other, we can treat them as individual single gaps by masking others. Note that if we assume that one planet can carve two gaps (Bae & Zhu 2018a, b; Dong et al. 2018) and the secondary gap is at $0.5\text{--}0.7 r_p$, we can input all the gaps into CNN models since our CNN models have been trained with data that have these secondary gaps generated by a single planet. The model does not consider migration, which can lead to a different gap shape (Nazari et al. 2019; Kanagawa et al. 2020). The gap substructure can also change with time. The discs are at 1000 orbits, which is 1 Myr for a planet at 100 au or 0.1 Myr for a planet at 20 au. We used 2D simulations, but the situation in 3D might be different. We neglected the self-gravity of the disc and the thermodynamic processes. These effects will also change the shape of the gap (Miranda & Rafikov 2020a, b; Rowther et al. 2020; Zhang & Zhu 2020; Ziampras, Kley & Dullemond 2020). The dust in our simulations was treated as passive test particles. In reality, the dust's back-reaction on the gas is important (Kanagawa et al. 2018; Hsieh & Lin 2020; Huang et al. 2020; Yang & Zhu 2020). When producing the synthetic image, we neglected scattering, but it can affect the images when the disc is optically thick (Liu 2019; Zhu et al. 2019). The dust was also assumed to be settled. Not to mention that it is very likely that some substructures can have non-planet

origins (e.g. snowline and MHD effects). On the other hand, with more physical processes understood and included in simulations, CNNs can be used to make predictions.

5.4 Future perspectives

This line of work can also be applied to infer the Stokes number St of particles in the discs. The Stokes number is highly correlated to the gap width and depth. Measuring St can help us understand the particle size and dust settling.

The gas component is more massive and has a larger radial and vertical extent than the dust in protoplanetary discs. The (sub)mm molecular line observations also contain velocity information at different disc positions, leading to a 3D datacube. One can infer planet mass from kinematic features in channel maps, such as the kink and the deviation from Keplerian velocity (Pinte et al. 2018; Teague et al. 2018). In star formation, CNN models have already been used on the whole 3D datacube (Xu et al. 2020a, b). This can possibly be applicable in protoplanetary discs as well.

6 CONCLUSIONS

Substructures are found to be ubiquitous in protoplanetary discs. If some of them are induced by planets, the increasing number of high-resolution protoplanetary disc observations is revealing the population of young forming planets.

The properties of these substructures (e.g. the width and depth of the gap) are related to the planet mass. Previous works used either fined-tuned models or linear fitting on a large parameter space to infer the planet mass. Instead, we used CNNs to predict the planet mass directly from radio dust continuum images. To train the CNNs, we use data from synthetic observations in Zhang et al. (2018) and some new simulations. We built both classification and regression models. The classification models can predict five planet masses ranging from $11 M_{\oplus}$ to $3 M_J$. The VGG-like model can reach 89 percent accuracy, whereas the ResNet model can reach 92 percent accuracy. The accuracies for less massive planets are lower. They are higher for more massive planets. The regression model can predict planet mass and disc viscosity at the same time. Similar to the classification model, it predicts more massive planets with higher accuracy. The standard deviation of the prediction is around 0.16 dex. The prediction of α has uncertainty around 0.23 dex, and can be used to constrain the disc turbulence.

CNN models cannot fully break the degeneracy between the planet mass and disc viscosity. It tends to over(under)predict them at the same time. However, it is surprising that this degeneracy relationship can be easily found in CNNs without the need of theoretical knowledge (Kanagawa et al. 2016) and detailed fitting and tuning (Zhang et al. 2018). This shows the potential of CNNs as diagnostic tools.

Using Grad-CAM, we showed that the networks indeed catch the important feature, i.e. gaps, in predicting the planet mass. We also applied the networks to several DSHARP gaps and found that the predictions are reasonable compared with the traditional method in Zhang et al. (2018). The code, along with that of the traditional method, is also provided.

The CNN methods are fast compared to fined-tuned models. It is more convenient than the linear fitting method since one can get a prediction instantaneously as long as an image is provided. The network can also be easily updated with more data or physical processes. It preserves the 2D information that should help break the degeneracy between planet mass and disc properties. The regression

model can predict several quantities at the same time. Unlike the traditional method, predicting shallow and narrow gaps takes the same amount of effort using CNNs, even though the prediction still has higher uncertainty in this regime.

There are also several shortcomings for CNN models. Different architectures or training procedures might lead to different predictions. The CNNs are not transparent, and it is difficult to know how exactly the networks work. Lastly, in contrast to the detailed modelling, the robustness of our CNNs ultimately is limited by our training data (simulations), i.e. the physical processes included.

The methods are more suitable for a large disc sample to obtain statistical trends of a young planet population. For individual disc, we can use this method to narrow down the parameter space for detailed simulations. Overall, the traditional linear fitting method (e.g. Zhang et al. 2018) still provides users more control on the input disc parameters, while the wide choice of tools in CNNs (e.g. classification and regression) and diagnostic tools (Grad-CAM) start to make CNNs more robust.

ACKNOWLEDGEMENTS

We thank the anonymous reviewer who raised insightful questions that significantly improved the quality of this manuscript. SZ thanks Yuan-Sen Ting, Duo Xu, and Yunzhong Hou for helpful discussions. SZ thanks Ruobing Dong for providing a CASA task template to generate sythetic ALMA observations. SZ acknowledges support from the University of Nevada, Las Vegas Barrick Fellowship. ZZ acknowledges support from the National Science Foundation under CAREER Grant No. AST-1753168.

Software used in this work are DUSTY FARGO-ADSG (Baruteau & Masset 2008a, b; Baruteau & Zhu 2016), CASA (McMullin et al. 2007), TENSORFLOW (Abadi et al. 2015), MATPLOTLIB (Hunter 2007), NUMPY (Van Der Walt, Colbert & Varoquaux 2011), SCIPY (Virtanen et al. 2020), SEABORN (Waskom 2021), and ASTROPY (Astropy Collaboration 2013).

DATA AVAILABILITY

The data underlying this article are publicly available on GitHub at <https://github.com/zhangsj96/DSHARP.VII.git> for the linear fitting method and at <https://github.com/zhangsj96/PGNets.git> for CNNs.

REFERENCES

- Abadi M. et al., 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, available at <https://www.tensorflow.org/>
- Alibert Y., Venturini J., 2019, *A&A*, 626, A21
- Andrews S. M. et al., 2018, *ApJ*, 869, L41
- Astropy Collaboration, 2013, *A&A*, 558, A33
- Auddy S., Lin M.-K., 2020, *ApJ*, 900, 62
- Auddy S., Dey R., Lin M.-K., Hall C., 2021, *ApJ*, 920, 3
- Bae J., Zhu Z., 2018a, *ApJ*, 859, 118
- Bae J., Zhu Z., 2018b, *ApJ*, 859, 119
- Baruteau C., Masset F., 2008a, *ApJ*, 672, 1054
- Baruteau C., Masset F., 2008b, *ApJ*, 678, 483
- Baruteau C., Zhu Z., 2016, *MNRAS*, 458, 3927
- Birnstiel T. et al., 2018, *ApJ*, 869, L45
- Christiaens V. et al., 2019, *MNRAS*, 486, 5819
- Cieza L. A. et al., 2021, *MNRAS*, 501, 2934
- Dieleman S., Willett K. W., Dambre J., 2015, *MNRAS*, 450, 1441
- Dipierro G., Laibe G., 2017, *MNRAS*, 469, 1932
- Dipierro G., Price D., Laibe G., Hirsh K., Cerioli A., Lodato G., 2015, *MNRAS*, 453, L73
- Dipierro G. et al., 2018, *MNRAS*, 475, 5296
- Dong R., Zhu Z., Whitney B., 2015, *ApJ*, 809, 93
- Dong R., Li S., Chiang E., Li H., 2018, *ApJ*, 866, 110
- Flaherty K. M., Hughes A. M., Teague R., Simon J. B., Andrews S. M., Wilner D. J., 2018, *ApJ*, 856, 117
- Fung J., Shi J.-M., Chiang E., 2014, *ApJ*, 782, 88
- Guzmán V. V. et al., 2018, *ApJ*, 869, L48
- Haffert S. Y., Bohn A. J., de Boer J., Snellen I. A. G., Brinchmann J., Girard J. H., Keller C. U., Bacon R., 2019, *Nat. Astron.*, 3, 749
- Hashimoto J., Aoyama Y., Konishi M., Uyama T., Takasao S., Ikoma M., Tanigawa T., 2020, *AJ*, 159, 222
- Hassan S., Liu A., Kohn S., La Plante P., 2019, *MNRAS*, 483, 2524
- He K., Zhang X., Ren S., Sun J., 2016a, in Leibe B., Matas J., Sebe N., Welling M., eds, Computer Vision – ECCV 2016. Springer Int. Publ., Cham, p. 630
- He K., Zhang X., Ren S., Sun J., 2016b, 2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Deep Residual Learning for Image Recognition. IEEE, Las Vegas, NV, p. 770
- Hezaveh Y. D., Perreault Levasseur L., Marshall P. J., 2017, *Nature*, 548, 555
- Hsieh H.-F., Lin M.-K., 2020, *MNRAS*, 497, 2425
- Huang J. et al., 2018, *ApJ*, 869, L42
- Huang P., Li H., Isella A., Miranda R., Li S., Ji J., 2020, *ApJ*, 893, 89
- Hunter J. D., 2007, *Comput. Sci. Eng.*, 9, 90
- Ioffe S., Szegedy C., 2015, Proc. 32nd Int. Conf. Mach. Learn., PMLR, Vol. 37, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. PMLR, p. 448, available at: <http://proceedings.mlr.press/v37/loff15.html>
- Isella A., Benisty M., Teague R., Bae J., Keppler M., Facchini S., Pérez L., 2019, *ApJ*, 879, L25
- Izquierdo A. F., Testi L., Facchini S., Rosotti G. P., van Dishoeck E. F., 2021, *A&A*, 650, A179
- Jennings J., Booth R. A., Tazzari M., Clarke C. J., Rosotti G. P., 2021, *MNRAS*, in press, preprint (arXiv:2103.02392)
- Kanagawa K. D., Muto T., Tanaka H., Tanigawa T., Takeuchi T., Tsukagoshi T., Momose M., 2015, *ApJ*, 806, L15
- Kanagawa K. D., Muto T., Tanaka H., Tanigawa T., Takeuchi T., Tsukagoshi T., Momose M., 2016, *PASJ*, 68, 43
- Kanagawa K. D., Muto T., Okuzumi S., Tanigawa T., Taki T., Shibaike Y., 2018, *ApJ*, 868, 48
- Kanagawa K. D., Nomura H., Tsukagoshi T., Muto T., Kawabe R., 2020, *ApJ*, 892, 83
- Keppler M. et al., 2018, *A&A*, 617, A44
- Kingma D. P., Ba J., 2015, in Bengio Y., LeCun Y., eds, 3rd Int. Conf. Learn. Represent., ICLR 2015. Conf. Track Proc., San Diego, CA
- LeCun Y., Bottou L., Bengio Y., Haffner P., 1998, *Proc. IEEE*, 86, 2278
- Lieu M., Conversi L., Altieri B., Carry B., 2019, *MNRAS*, 485, 5831
- Liu H. B., 2019, *ApJ*, 877, L22
- Lodato G. et al., 2019, *MNRAS*, 486, 453
- Long F. et al., 2018, *ApJ*, 869, 17
- McKay M. D., Beckman R. J., Conover W. J., 1979, *Technometrics*, 21, 239
- McMullin J. P., Waters B., Schiebel D., Young W., Golap K., 2007, in Shaw R. A., Hill F., Bell D. J., eds, ASP Conf. Ser. Vol. 376, Astronomical Data Analysis Software and Systems XVI. Astron. Soc. Pac., San Francisco, p. 127
- Miranda R., Rafikov R. R., 2020a, *ApJ*, 892, 65
- Miranda R., Rafikov R. R., 2020b, *ApJ*, 904, 121
- Müller A. et al., 2018, *A&A*, 617, L2
- Nazari P., Booth R. A., Clarke C. J., Rosotti G. P., Tazzari M., Juhasz A., Meru F., 2019, *MNRAS*, 485, 5914
- Paardekooper S.-J., Mellema G., 2006, *A&A*, 453, 1129
- Perez S., Dunhill A., Casassus S., Roman P., Szulágyi J., Flores C., Marino S., Montesinos M., 2015, *ApJ*, 811, L5
- Picogna G., Kley W., 2015, *A&A*, 584, A110
- Pinte C., Dent W. R. F., Ménard F., Hales A., Hill T., Cortes P., de Gregorio-Monsalvo I., 2016, *ApJ*, 816, 25
- Pinte C. et al., 2018, *ApJ*, 860, L13
- Rabago I., Zhu Z., 2021, *MNRAS*, 502, 5325

- Rosotti G. P., Juhasz A., Booth R. A., Clarke C. J., 2016, *MNRAS*, 459, 2790
- Rowther S., Meru F., Kennedy G. M., Nealon R., Pinte C., 2020, *ApJ*, 904, L18
- Ruane G. et al., 2017, *AJ*, 154, 73
- Selvaraju R. R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D., 2017, 2017 IEEE Int. Conf. Comput. Vis. (ICCV), Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. IEEE, Venice, p. 618
- Simonyan K., Zisserman A., 2014, preprint ([arXiv:1409.1556](https://arxiv.org/abs/1409.1556))
- Teague R., Bae J., Bergin E. A., Birnstiel T., Foreman-Mackey D., 2018, *ApJ*, 860, L12
- Van Der Walt S., Colbert S. C., Varoquaux G., 2011, *Comput. Sci. Eng.*, 13, 22
- Van Oort C. M., Xu D., Offner S. S. R., Gutermuth R. A., 2019, *ApJ*, 880, 83
- Virtanen P. et al., 2020, *Nat. Methods*, 17, 261
- Wagner K. et al., 2018, *ApJ*, 863, L8
- Wang J. J. et al., 2020, *AJ*, 159, 263
- Waskom M. L., 2021, *J. Open Source Softw.*, 6, 3021
- Xu D., Offner S. S. R., Gutermuth R., Oort C. V., 2020a, *ApJ*, 890, 64
- Xu D., Offner S. S. R., Gutermuth R., Oort C. V., 2020b, *ApJ*, 905, 172
- Yang C.-C., Zhu Z., 2020, *MNRAS*, 491, 4702
- Zhang S., Zhu Z., 2020, *MNRAS*, 493, 2287
- Zhang S. et al., 2018, *ApJ*, 869, L47
- Zhou B., Khosla A., Lapedriza A., Oliva A., Torralba A., 2016, 2016, IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Learning Deep Features for Discriminative Localization. IEEE, Las Vegas, NV, p. 2921
- Zhu Z., Stone J. M., Rafikov R. R., Bai X.-n., 2014, *ApJ*, 785, 122
- Zhu Z. et al., 2019, *ApJ*, 877, L18
- Ziampras A., Kley W., Dullemond C. P., 2020, *A&A*, 637, A50
- Zurlo A. et al., 2020, *A&A*, 633, A119

This paper has been typeset from a $\text{\TeX}/\text{\LaTeX}$ file prepared by the author.