Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **506,** 1427–1437 (2021) https://doi.org/10.1093/mnras/stab1715
Advance Access publication 2021 June 21

Peculiar velocity estimation from kinetic SZ effect using deep neural networks

Yuyu Wang [1,2]★ Nesar Ramachandra,[3,4]★ Edgar M. Salazar-Canizales,[5,6] Hume A. Feldman,[2] Richard Watkins[7] and Klaus Dolag[8,9]

[1]*Department of Astronomy, School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai 200240, China*
[2]*Department of Physics & Astronomy, University of Kansas, Lawrence, KS 66045, USA*
[3]*High Energy Physics Division, Argonne National Laboratory, Lemont, IL 60439, USA*
[4]*Kavli Institute for Cosmological Physics, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA*
[5]*Departamento de Física, Universidad de Sonora, 83000 Hermosillo, Mexico*
[6]*Department of Physics, University of Arizona, Tucson, AZ 85721, USA*
[7]*Department of Physics, Willamette University, Salem, OR 97301, USA*
[8]*University Observatory Munich, Scheinerstr 1, D-81679 Munich, Germany*
[9]*Max-Planck-Institut für Astrophysik (MPA), Karl-Schwarzschild Strasse 1, D-85748 Garching bei München, Germany*

Accepted 2021 June 11. Received 2021 June 11; in original form 2020 October 10

ABSTRACT

The Sunyaev–Zel'dolvich (SZ) effect is expected to be instrumental in measuring velocities of distant clusters in near future telescope surveys. We simplify the calculation of peculiar velocities of galaxy clusters using deep learning frameworks trained on numerical simulations to avoid the independent estimation of the optical depth. Images of distorted photon backgrounds are generated for idealized observations using one of the largest cosmological hydrodynamical simulations, the Magneticum simulations. The model is tested to determine its ability of estimating peculiar velocities from future kinetic SZ observations under different noise conditions. The deep learning algorithm displays robustness in estimating peculiar velocities from kinetic SZ effect by an improvement in accuracy of about 17 per cent compared to the analytical approach.

Key words: methods: statistical – techniques: radial velocities – cosmic background radiation.

1 INTRODUCTION

The Sunyaev–Zel'dolvich (SZ) effect (Sunyaev & Zeldovich 1970, 1972, 1980) describes the process of cosmic microwave background (CMB) distortion caused by the inverse Compton scattering of CMB photons off by electrons in galaxy clusters. The SZ effect has two contributions: thermal (tSZ) and kinetic SZ (kSZ) effect. The tSZ effect is caused by the random motion of hot electrons in the intra-cluster medium, while the kSZ effect is caused by the bulk motion of galaxy clusters. Therefore, the kSZ effect can be used in estimating peculiar velocities of galaxy clusters (e.g. Rephaeli & Lahav 1991; Bhattacharya & Kosowsky 2008; Zhang et al. 2008; Kashlinsky et al. 2009; Atrio-Barandela et al. 2012; Planck Collaboration XIII 2014; Sayers et al. 2016; Hurier 2017; Soergel et al. 2017; Planck Collaboration LIII 2018; Kirillov & Savelova 2019). However, the weak signal of the kSZ effect makes its detection very difficult. Hand et al. (2012) first detected the kSZ effect from CMB maps with the Atacama Cosmology Telescope (ACT) through pairwise momentum estimator. Using similar methods, several groups have detected the kSZ effect in both real and Fourier spaces (e.g. Soergel et al. 2016; Planck Collaboration XXXVII 2016; Calafut, Bean & Yu 2017; Li et al. 2018; Sugiyama, Okumura & Spergel 2018). In addition, some studies detected the kSZ effect by cross-correlating kSZ temperature map with density, velocity field (e.g. Hill et al. 2016; Schaan et al. 2016; Nguyen et al. 2020) or other observables such as the angular redshift fluctuations (Chaves-Montero et al. 2019; Hernandez-Monteagudo, Chaves-Montero & Angulo 2019). The kSZ effect through measurements of the CMB temperature dispersion was detected by Planck Collaboration LIII (2018). Furthermore, Mittal, de Bernardis & Niemack (2018) discussed the ability of measuring the kSZ effect for individual clusters in the upcoming multifrequency surveys. With the improvements in the kSZ measurement, the estimate of peculiar velocities using kSZ effect for individual clusters may become possible.

The peculiar velocity field is a powerful tracer of density fluctuations, which is generally studied through ensemble statistics such as bulk flows, velocity correlation functions, and the pairwise velocity statistics (e.g. Borgani et al. 2000; Watkins, Feldman & Hudson 2009; Kumar et al. 2015; Wang et al. 2018). The pairwise velocity statistics is the mean value of the peculiar velocity difference of galaxy pairs at separation $\boldsymbol{r}$ and is a widely used approach to study the large-scale velocity field (e.g. Ferreira et al. 1999; Juszkiewicz et al. 2000; Feldman et al. 2003; Zhang et al. 2008; Hand et al. 2012; Planck Collaboration XXXVII 2016). Traditionally, estimating peculiar velocities using the kSZ effect requires information about optical depth, which describes the integration of electron densities. However, the measurement of optical depth has errors and biases that may affect the estimate of peculiar velocities. Lindner et al. (2015) estimates an average uncertainty of the cluster optical depth

★ E-mail: yuyuwang@sjtu.edu.cn (YW); nramachandra@anl.gov (NR)

around 31 per cent and Mittal et al. (2018) forecasts an average uncertainty about 24 per cent in observations. In addition, using emission-weighted temperature, which is not observable, rather than density-weighted temperature in measurements may lead to a biased optical depth estimation (Diaferio et al. 2005; Dolag & Sunyaev 2013). In simulations, the optical depth varies between models with and without star formation and feedback (Flender et al. 2016; Flender, Nagai & McDonald 2017). The weak kSZ signal and optical depth errors make the kSZ peculiar velocity calculation imprecise and difficult. Machine learning algorithms may provide a simpler and more accurate method for estimating kSZ peculiar velocities.

Machine learning algorithms are designed without explicit programming of the physical phenomena, instead perform complex analyses in a data-driven manner. Some machine learning methods, including Gaussian processes, decision trees, nearest-neighbour algorithms, and support vector machines, have been used in astrophysical contexts (see Baron 2019 for a recent overview). Deep neural networks, a class of extremely flexible statistical models, are a subset of machine learning algorithms. These typically have a large number of trainable parameters that can be optimized from abundant quantities of data, while the relevant features are extracted automatically. Utilization of such deep learning methods is rapidly increasing due to the availability of data, advancements of computational architectures (such as the graphic processors, tensor processors, and dedicated accelerators), and the development of accessible software libraries (such as `TensorFlow`, `Keras`, `Torch`, and `JAX`). Specifically, the convolutional neural networks (CNNs), where features of images are extracted hierarchically in various layers of the deep network, are powerful tools in image-based regression, classification, compression, and generation tasks.

However, the model interpretability and explainability of deep learning methods remain to be areas of active research. The overparametrized architecture of deep CNNs results in a difficult uncertainty quantification, and features important assessment and understanding of failure modes. The dependence on hyper parameter searches, optimal architectures, network initialization, and optimization routines also contributes to the cryptic nature of the results achieved from deep learning algorithms. Thus, deep learning algorithms are often characterized as 'black box' inference techniques.

Despite these caveats, deep learning neural networks trained on sufficient amount of data outperform the traditional classification and regression techniques (shown in various comparison studies, for instance Metcalf et al. 2019 in strong lensing detection problem). Specifically in deep CNNs, the low-, mid-, and high-level image features computed in the initial, middle, and final convolutional layers, respectively, are used to correlate inputs and targets in a highly efficient manner. This makes CNNs practical tools in image processing tasks, including astronomical applications.

Learning the intrinsic characteristics of the data set may be accomplished unsupervised where the training is unaccompanied by correct responses, e.g. in generative models (Ravanbakhsh et al. 2016; Morningstar et al. 2018; He et al. 2019). Alternatively, a supervised routine involves learning correct mapping during training. Supervised techniques for object identification have been applied in a broad variety of astrophysical problems including strong lensing image classifications (Petrillo et al. 2017) and parameter estimations (Hezaveh, Perreault Levasseur & Marshall 2017; Levasseur, Hezaveh & Wechsler 2017; Morningstar et al. 2018), which have demonstrated improvements to predictive precision and inference speed compared to traditional inference techniques.

Machine learning applications in cosmological analyses frequently deal with simulated data instead of observational data. This is in part due to the lack of large quantity of observational data. On the other hand, the ability of calibrating the forward model parameters is not robust enough to generate unbiased training data.

In this paper, we use simulation data to test the feasibility of extracting peculiar velocities from kSZ effect by deep learning architectures. In Section 2, we describe the relation between the SZ effect and the peculiar velocity. In Section 3, we introduce the simulation we used for generating training and validation data. In Section 4, we display the CNN structure of the deep learning model. In Section 5, we show predictions of our model and compare it with the analytical method. In Section 6, we exam the model predictions through the pairwise velocity statistics. In Section 7, we test the feasibility of the model to observations under noise conditions. In Section 8, we conclude this paper.

2 SUNYAEV–ZEL'DOVICH EFFECT

The relation between radial motions of galaxy clusters and the observed radiation temperature was first introduced by Sunyaev & Zeldovich (1980) with the equation (1), where v_e indicates the velocity of electron along the line of sight, v_c is the line of sight peculiar velocity of cluster, $\tau = \int \sigma_T N_e \, dl$ is the Thomson Scattering optical depth, σ_T is the Thomson Scattering cross-section, and N_e is the electron density.

$$\frac{\Delta T_{kSZ}}{T_{CMB}} = -\frac{1}{c}\int \sigma_T N_e v_e dl \simeq -\frac{\tau}{c} v_c. \quad (1)$$

On the other hand, the tSZ effect (Sunyaev & Zeldovich 1970) is usually expressed by the Compton y parameter:

$$\frac{\Delta T_{tSZ}}{T_{CMB}} = y f(x), \quad y = \int \frac{k_B T_e}{m_e c^2} \sigma_T N_e \, dl, \quad (2)$$

where $f(x) = x\coth(x/2) - 4$ and x is the dimensionless frequency given by $x = h\nu/(k_B T_{CMB})$.

Since the kSZ signal is independent of the redshift and has a strong suppression on the secondary CMB anisotropy, the kSZ effect can be available up to the era of reionization. However, due to the weakness of the signal and the error in optical depth measurement, the peculiar velocity estimation from kSZ effect is very challenging in real observations.

Alternatively, the potential of utilizing numerical simulations for estimating peculiar velocity from the kSZ effect is being studied extensively. For instance, Soergel et al. (2017) have shown promising results with obtaining pairwise velocity statistics with kSZ effect by applying map filtering to the signals and used tSZ effect to estimate the average optical depth.

For both observations and simulations, the requirement of optical depth estimation is inevitable when using the analytical method to calculate the kSZ peculiar velocity. In addition, the estimation of optical depth in simulations varies between models with and without star formation and feedback. The measurement of optical depth for a single cluster in observation is even more challenging. Therefore, a method that can predict peculiar velocities from kSZ effect without complicated estimation of the optical depth would reduce the difficulty in calculating kSZ peculiar velocities significantly. Deep learning algorithm provides a possible approach to achieve it. A training data set from a numerical simulation with a realistic SZ map-making pipeline may empower the deep learning model to simplify the computation in the estimation of peculiar velocities by avoiding the map filtering and optical depth estimation.

Table 1. Specifications of the training data along with the cosmological parameters of the Magneticum simulation Box0.

Matter density, Ω_m	0.272
Cosmological constant density, Ω_Λ	0.728
Baryon density, Ω_b	0.046
Hubble parameter, h (100 km s^{-1} Mpc^{-1})	0.704
Amplitude of matter density fluctuations, σ_8	0.809
Primordial scalar spectral index, n_s	0.963
Box size (h^{-1} Mpc)	2688
Number of particles	2×4536^3
Mass of dark matter particles, m_{dm} ($10^9\ h^{-1}\ M_\odot$)	13
Mass of gas particles, m_{gas} ($10^9\ h^{-1}\ M_\odot$)	2.6
Softening of particles, f_p (h^{-1} kpc)	10
Softening of stars, f_s (h^{-1} kpc)	5
Redshift range for clusters in slice 1	[1.04, 1.32]
Redshift range for clusters in slice 2	[1.32, 1.59]
Redshift range for clusters in slice 3	[1.59, 1.84]
Redshift range for clusters in slice 4	[1.84, 2.15]
Mass of galaxy clusters	$[1, 70] \times 10^{13}\ M_\odot$
Average mass of galaxy clusters	$10^{14}\ M_\odot$
Number of kSZ maps of each slice	10 000
Number of tSZ maps of each slice	10 000
Size of maps	$2R_{vir}$

3 SIMULATION AND TRAINING DATA

Deep neural networks typically utilize a large amount of training data in order to capture the complexities in the data and optimize the model. Therefore, cosmological simulations that can provide a large number of galaxy cluster samples are necessary. In addition, the simulation data must resemble idealized observations from telescopes, which leads to a light-cone pipeline to generate kSZ and tSZ images.

In this paper, we use the Magneticum simulations[1] to generate kSZ and tSZ cluster images. The Magneticum simulations are a set of cosmological hydrodynamical simulations with a large range of scales and resolutions. The Magneticum simulations are generated by an extended version of the *N*-body/SPH GADGET3 code (Springel, Yoshida & White 2001; Springel 2005; Beck et al. 2016) with *WMAP7* (Larson et al. 2011) cosmological parameters from Komatsu et al. (2011). The dark matter only simulation includes dark matter and dark energy that provide gravity information, while the hydrodynamical simulation uses the hydrodynamic equations to include the baryonic component, which can be described as an ideal fluid. In addition, these simulations follow a wide range of physical processes (for details, see Hirschmann et al. 2014; Teklu et al. 2015), which are important for galaxy formation and the evolution of the intra-galactic and intra-cluster medium (see Biffi, Dolag & Böhringer 2013; Dolag, Komatsu & Sunyaev 2016; Gupta et al. 2017, and accompanying results). With the baryonic particles and temperature information, the SZ signal can be detected by tracking back along the line of sight.

In this paper, we use the largest box, Box0 (see also Bocquet et al. 2016; Soergel et al. 2017; Ragagnin et al. 2019), in the Magneticum simulations. Table 1 shows the cosmological parameters of the simulation box and the parameters of our datasets. We take four redshift slices from the simulation that cover redshift in a range of [1.04, 2.15]. From those four redshift slices, we generated 40 000 kSZ and 40 000 tSZ images (10 000 images from each redshift slice) through SMAC (Dolag et al. 2005), which is a map making utility for idealized observations. The size of a kSZ/tSZ cluster image is set to be twice its Virial radius, which is the radius within where the system obeys the Virial theorem. To reduce the calculation expense, we use the redshift of each slice instead of the redshift of each cluster in calculating the Virial radius, which means the size of the cluster images is not perfectly normalized to the Virial radius. According to our test, the difference is small and its effect on the final results is negligible.

Fig. 1 shows the kSZ and tSZ examples of four clusters generated from the Magneticum simulations. We train the neural network with 80 per cent of the images, which are similarly as the examples shown in figure, and use the rest 20 per cent of them as validation data for testing.

4 THE DEEP LEARNING MODEL

A custom-designed deep learning algorithm is implemented here to predict the peculiar velocity from kSZ effect. CNNs are an obvious choice for such image-based regression analyses due to the following reasons: first, the amount of generated data (40 000 kSZ images) can be efficiently utilized in deep learning neural networks that consist of a large number of trainable model parameters called weights. It can be seen that with respect to the scaling of accuracy with the size of the data set, deep learning neural networks outperform most existing machine learning models. Secondly, despite having characteristic features in the SZ signal (as seen in Fig. 1), the feature-mapping to peculiar velocities is not straightforward due to the optical depth. This makes feature-agnostic training algorithms like CNNs more desirable than feature-specified learning methods for modelling SZ images. The CNNs can extract high- and low-level features from a series of convolutional filters, which are used to train the peculiar velocity prediction.

Numerous deep learning neural network architectures are currently in literature and under active research. However, we do not wish to compare different CNN variants in this work, nor claim to achieve the best possible accuracy in estimating peculiar velocities. our goal in this paper is to demonstrate the feasibility of using deep learning neural networks to estimate peculiar velocities using the direct input of kSZ images and highlight the advantage of such simulation-based training approaches over the analytical calculation techniques on the kSZ peculiar velocity estimation.

Fig. 2 shows our CNN architecture with only the kSZ image as input data. It follows a conventional deep neural network architecture like the CIFAR-10 (LeCun, Bengio & Hinton 2015), with layers stacked sequentially. The kSZ image, the input data, will be addressed through several layer blocks (including convolutional, pooling, and dropout layers) and multiple dense layers to get the peculiar velocity as the output. Short descriptions of each layers are as follows: (1) Convolutional layers consist of numbers of image kernels that extract morphological features of the image. While the high-level features are extracted at the initial convolutional layers, more abstract features are obtained later. (2) The pooling layer operates on each map independently, and progressively reduces the spatial size of the map to reduce the amount of computations in the network. (3) Dropout layers re-initialize a sub-set of neurons of the network at every epoch of the training, which reduces the chances of overtraining. (4) The flatten later converts the 2D matrix to a single 1D vector. (5) Dense layers use fully connected neurons[2] to map this 1D vector to the peculiar velocity corresponding to the input image.

[1] http://www.magneticum.org

[2] For given inputs $\boldsymbol{x}$, the output $\boldsymbol{y}$ of each neuron is expressed in terms of its non-linear activation function ϕ, weights $\boldsymbol{W}$, and biases $\boldsymbol{b}$ as $\boldsymbol{y} =$

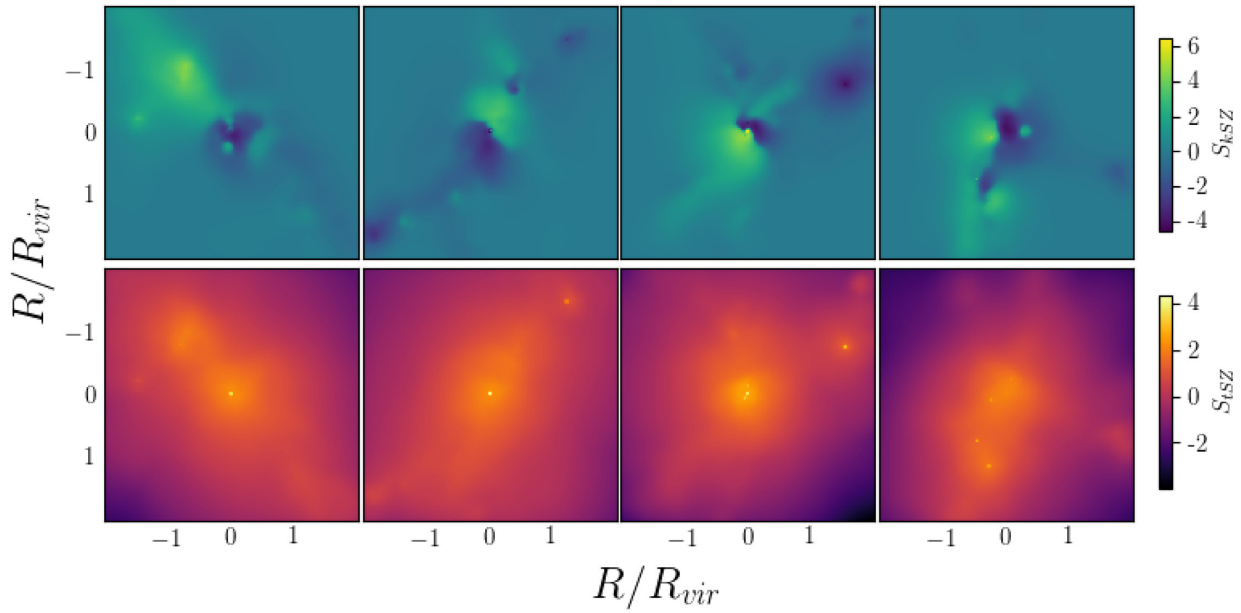

Figure 1. kSZ (upper panels) and tSZ images (lower panels) of four clusters. $S_{\rm kSZ}$ and $S_{\rm tSZ}$ indicate the kSZ and tSZ signals re-scaled to increase the image contrast. For the kSZ signals in the upper panels, the colour corresponding to the kSZ effect is presented via $S_{\rm kSZ} = \sinh^{-1}(\Delta T_{\rm kSZ}/T_{\rm CMB} \times 10^6)$. For the lower panels, the tSZ signal is a function of the Compton y parameter, $S_{\rm tSZ} = \log_{10}(y \times 10^6)$. The width of each image equals to four times of the cluster Virial radius.

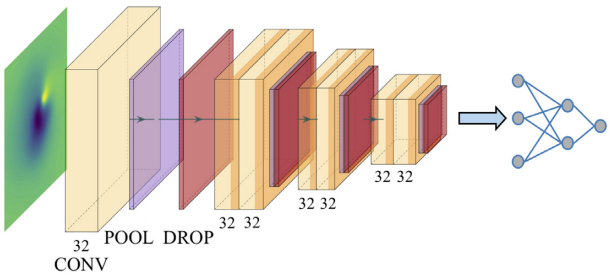

Figure 2. Schematic CNN architecture for regression including the kSZ effect only. The real architecture used in this analyses is multiple blocks of convolutional, pooling, and dropout layers repeated before feeding the dense layers.

Overall, the repeated convolutional layer blocks extract abstract-featured maps from the images, which are then used as inputs in the dense layers towards the end of the network. As opposed to image classifications, this regression pipeline has a linear activation to get point estimation of the peculiar velocity. The loss function is defined by the mean square error value $L = (v - v_{\rm p})^2$, where v is the known peculiar velocity from the simulation, taken as true values (true velocity) for the training, and $v_{\rm p}$ is the predicted peculiar velocity from the deep learning model. By providing enough correct data to learn from, the model can be trained to project the input kSZ image to the output peculiar velocity.

The model of including both kSZ and tSZ images has similar architecture with an independent repeating convolutional structure, shown in Fig. 3. The only difference in the combined kSZ and tSZ image analysis is that the kSZ and tSZ are computed in separate branches. After the flatten layer, the outputs from those two branches

$\phi(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$. The trainable parameters $(\boldsymbol{W}, \boldsymbol{b})$ of the model are optimized during the learning phase.

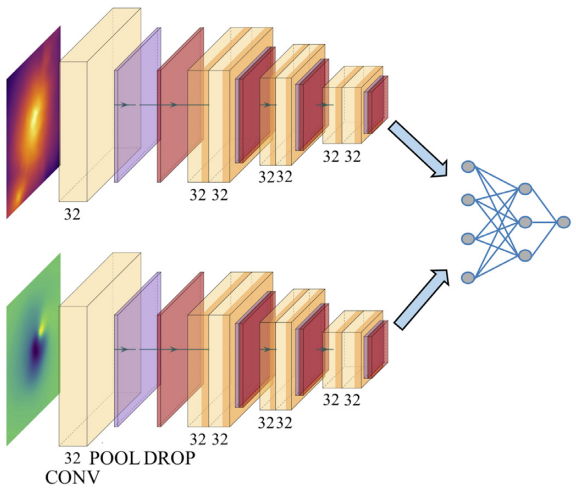

Figure 3. Same as Fig. 2 but for the both kSZ and tSZ effects. Two separate CNN branches process the images with input kSZ and tSZ signals, and the outputs from both the branches are then combined.

are concatenated to a 1D vector, which is then fed into dense layers for predicting the peculiar velocity.

4.1 Uncertainty quantification

One of the shortcomings of a traditional regression analysis with CNNs is that it lacks proper treatment for the uncertainty quantification. This stems from the vast number of trainable parameters in the CNNs, such as the ones shown in Figs 2 and 3. A complete understanding of the posterior (in our case, the probability distribution of target peculiar velocities for given input SZ images) becomes intractable due to the large number of statistical model parameters. Bayesian neural network frameworks using Monte Carlo (MC) or variational inference techniques have been explored for solving

such inference problems, but many of these methods are challenging due to computational expenses, lack of convergence or clear diagnostics.

Alternatively, the MC dropout method (see Gal & Ghahramani 2015, for a detailed review) offers a middle ground for approximating the prediction uncertainty within reasonable computational overload. This is done by the utilization of existing trained deep learning models with dropout layers in prediction of the error bars around the mean estimates.

A dropout layer, as explained previously, is generally used in CNNs to avoid overfitting in the training phase. However, they can also be used in the testing phase as an approximate sampling scheme for model parameters. It was also shown by Gal & Ghahramani (2015) that the MC dropout is a Bayesian approximation of neural networks to Gaussian processes, where the error modelling is formally defined.

The implementation of MC dropout is as follows: We consider an ensemble of neural networks (with ensemble size $N_{\rm tot}$) of the same architecture, but only different from each other by a fraction (prescribed by the dropout rate d) of trained neurons that are re-initialized to a random value (or 'dropped-out'). Using the base architectures shown in Figs 2 and 3 with dropout rate d, we obtain this ensemble of $N_{\rm tot}$ networks. Each of these networks in the ensemble provide a different point-prediction of the peculiar velocity.

When a validation image $\mathbb{I}$ is forward propagated through each network in the ensemble, they provide individual predictions $v_{\rm p}^i(\mathbb{I})$, where $i = 0, 1, \ldots, N_{\rm tot}$. These individual predictions $v_{\rm p}^i(\mathbb{I})$ are different from each other due to the fact that a different fraction of their network parameters are dropped-out. The mean of all the individual predictions is calculated as $\langle v_{\rm p}\rangle = \frac{1}{N_{\rm tot}}\sum_{i=0}^{N_{\rm tot}} v_{\rm p}^i(\mathbb{I})$ and the variance as $\sigma_v^2 = \frac{1}{N_{\rm tot}}\sum_{i=0}^{N_{\rm tot}}[v_{\rm p}^i(\mathbb{I}) - \langle v_{\rm p}\rangle]^2$, respectively. These aggregate mean and variance will be considered as the uncertainty quantified prediction from the ensemble. The theoretical details of this approach are summarized in Appendix B.

Hence, the MC dropout is a simple prediction uncertainty quantification tool without any additional expensive computation tasks while training, unlike the Bayesian neural networks that explicitly define distributions in predictions (Kendall & Gal 2017). In addition to providing uncertainty estimations, such ensemble methods can also monitor failure modes, i.e. the choice of network architecture and training schemes can be compared in terms of robustness of the results.

For our implementation of both kSZ and combined kSZ and tSZ, we utilize an ensemble of $N_{\rm tot} = 100$ networks for our predictions. We also use a large dropout rate $d = 0.5$ to test models for both the consistency and the robustness of our final predictions. For the same $N_{\rm tot}$, we have observed a small decrease in the prediction uncertainty with reducing the dropout rate, but the mean does not vary significantly. Various combinations of dropout rate distribution (such as applying dropouts to different fully connected layers) have been checked to ensure that they do not affect the uncertainty estimates.

5 TRAINING

We build two models respective to the two CNN architectures in Section 4: Model I, kSZ only model shown in Fig. 2; Model II, the combined kSZ and tSZ model shown in Fig. 3. We check the universality of the models by training the models both with data from single redshift slices and with data of multiple redshift slices (all four of the redshift slices at once).

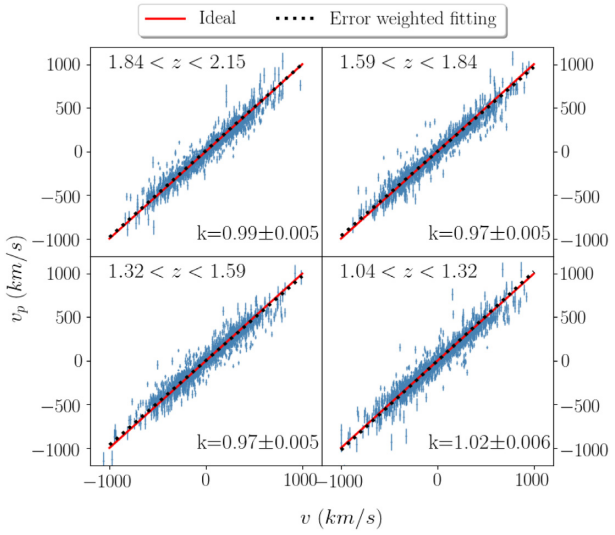

Figure 4. The results of Model I trained by kSZ images in each redshift slice. The x- and y-axes show the true (v) and predicted ($v_{\rm p}$) peculiar velocities, respectively. The red solid line shows the 1:1 ideal relation between the true and predicted velocities, while the black dotted line shows the uncertainty weighted linear fitting of the scatter. The error bars show the uncertainty of the predicted velocity using the MC dropout method.

For Model I, we first train the model with kSZ images of each redshift slice, which means we train the model four times independently, each time using 80 per cent of the 10 000 kSZ images of a single redshift slice; secondly, we train the model with the data of multiple redshift slices, using 80 per cent of the entire set of 40 000 kSZ images as the training set.

Fig. 4 shows the prediction results of Model I trained by kSZ images of single redshift slices. In the figure, the fitting line (black dotted line) is weighted by the uncertainty ($1/\sigma_v$), which represents the predictions accompanied by their error bars. The uncertainty weighted fitting result (black dotted line) agrees with the ideal expectation well, which is also represented by the fitting slope (k value). Although trained by data from different redshift slices, the prediction results of those four training sets have very similar fitting slopes, which means the model for predicting peculiar velocity from kSZ images is fairly stable with different redshifts. This is consistent with equation (1) that the kSZ effect is independent of the redshift. In addition, the similarity of contours (tested but not shown in the figure) of the scatters of different redshift slices proves redshift independence.

Fig. 5 shows the results of the Model I trained by the kSZ images from multiple redshift slices, which covers a larger redshift region. Comparing with Fig. 4, Fig. 5 has larger scatters due to more validation data. However, the fitting slope is similar to the one in Fig. 4. Though the model trained by the full data (from all four redshift slices) may have larger errors, it covers a larger region which makes the model more universal and flexible for applications.

In both Figs 4 and 5, the predictions ($v_{\rm p}$) using kSZ images show good agreements with the true velocities (v). However, the differences between the predictions and expectations result in scatter. Since Model I does not include information about optical depth, we add tSZ information into the training to explore a possible improvement (Model II). However, the results of Model II show no significant difference from the results of Model I, which might be due to the particular values of the optical depth of our simulated

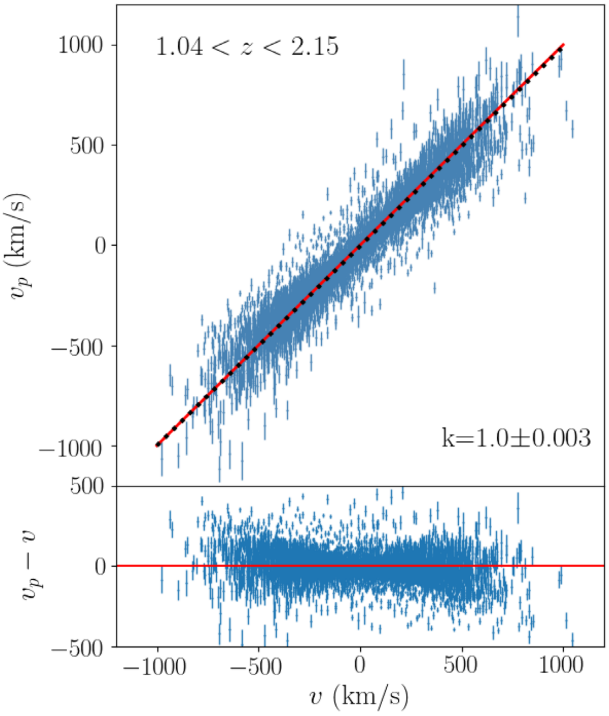

Figure 5. The results of the Model I trained by kSZ images of multiple redshift slices. The *x*- and *y*-axes show the true (v) and predicted (v_p) peculiar velocities, respectively. The red solid line shows the 1:1 ideal relation between the true and predicted velocities, and the black dotted line shows the uncertainty weighted linear fitting of the scatters. The error bars show the uncertainties of the predicted velocity using the MC dropout method. The bottom panel shows the difference between predictions and expectations ($v_p - v$).

images. The distribution of the optical depth in each redshift slice is a highly concentrated Gaussian distribution. For instance, the mean value and the standard deviation of the redshift slice 4 are 0.002 and 0.0003, respectively. This narrow range of the optical depth does not cause large enough variations in the kSZ velocity estimation. Therefore, Model II includes optical depth information, but it does not show significant improvement to the velocity estimation. The result of Model II is presented in Appendix A.

5.1 Error analysis

In this section, we quantify the uncertainty in order to test the performance of our models. Since the difference between the results of Model I and Model II is negligible, we only present the error analysis for Model I in this section. In the Fig. 5, the MC dropout uncertainties (error bars of predictions) increases with the magnitude of the predicted velocity, and the average relative MC dropout uncertainty (σ_v/v_p) is about 25 per cent.

However, the value of the relative MC dropout uncertainty is highly affected by its denominator, the predicted velocity v_p. Though the dropout uncertainty of low v_p is smaller than the dropout uncertainty of high v_p, the smaller denominator will increase the relative uncertainty of low v_p. Therefore errors of low-velocity clusters might bias the estimate of the average dropout uncertainty. We set different velocity limits (v_{limit}) to eliminate the effect from the low velocity, which is shown by the red line in Fig. 6. Eliminating velocities lower than 20 km s^{-1} reduces the average uncertainty

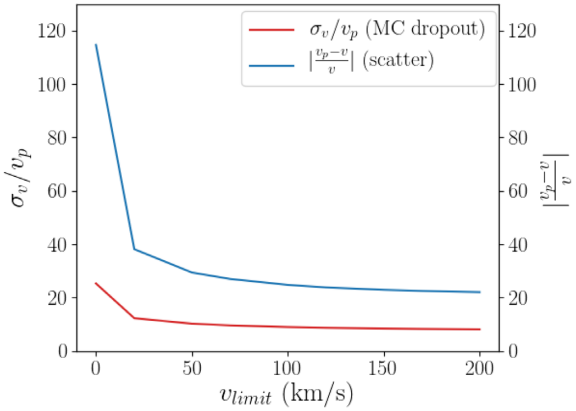

Figure 6. The average MC dropout uncertainty and the average scatter in percentage of Model I trained by data from multiple redshift slices with different velocity limits. The *x*-axis is the velocity limits, for example, the label 20 in *x*-axis means eliminating all the predicted velocities lower than 20 km s^{-1} ($|v_p| < 20$). The *y*-axis shows the average MC dropout uncertainty σ_v/v_p and the average relative scatter $|\frac{v_p-v}{v}|$ in percentage. The red line indicates the average uncertainty in percentage and the blue line shows the average scatter in percentage.

significantly to about 12 per cent. With larger velocity limits, the average uncertainty converges to about 8 per cent.

In addition, the scatter, which is the difference between the true velocity v and the predicted velocity v_p, is another factor that affects the accuracy of the prediction. While the dropout uncertainty is a measure of precision (statistical uncertainty), scatter is a measure of accuracy (systematic uncertainty). In the bottom panel of Fig. 5 (the residual plot), the absolute differences (the scatters) between predictions and expectations are mostly smaller than 200 km s^{-1}. We also see an increasing trend of scatters with the magnitude of the velocity, but it is not a very strong dependence. Using the same method, we calculate the average relative scatter (difference of predictions and expectations over expectations) with different velocity limits, which is shown by the blue line in Fig. 6. After eliminating the velocities lower than 20 km s^{-1}, the average scatter becomes about 38 per cent. With larger velocity limits, the average scatter converges to about 20 per cent.

5.2 Comparison with analytical calculations of peculiar velocity

The analytic calculation for estimating peculiar velocities from the kSZ effect requires information of optical depth of each individual clusters (equation 1). The optical depth of individual clusters in this paper is calculated through equation:

$$\tau_{\rm cluster} = \frac{\int_0^{R_{\rm vir}} \int_{l_-}^{l_+} \sigma_{\rm T} N_{\rm e} \, {\rm d}l \, {\rm d}r}{\pi R_{\rm vir}^2}, \quad (3)$$

where $l_- = -100\,h^{-1}$ Mpc, $l_+ = +100\,h^{-1}$ Mpc, and $R_{\rm vir}$ is the Virial radius of clusters. Therefore, the optical depth of individual clusters is calculated by the averaging electron density within the Virial radius. The integral distance dl for the optical depth is 200 h^{-1} Mpc, which is large enough to get a converged optical depth value. The kSZ value used in the analytical method is calculated by averaging the kSZ signals of each cluster within its Virial radius.

Fig. 7 shows the results of Model I and the analytical method (v_c) for each redshift slice. From the figure, we could conclude that

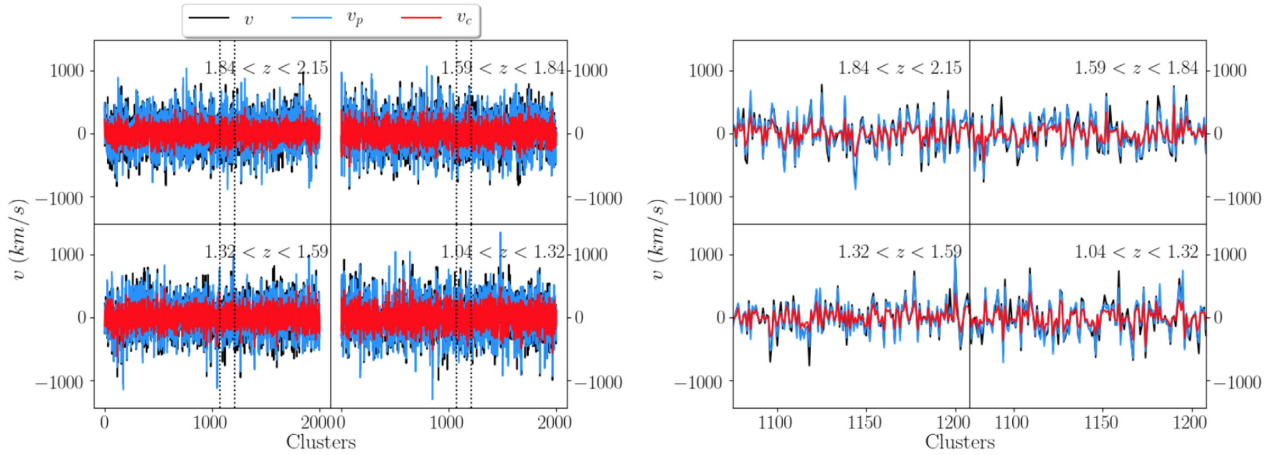

Figure 7. The results of Model I (blue), the analytical method (red), and true velocities from the simulation (black). The four left-hand panels show the result of 2000 testing clusters in each redshift slice. The right-hand panels show the amplification of the selected areas.

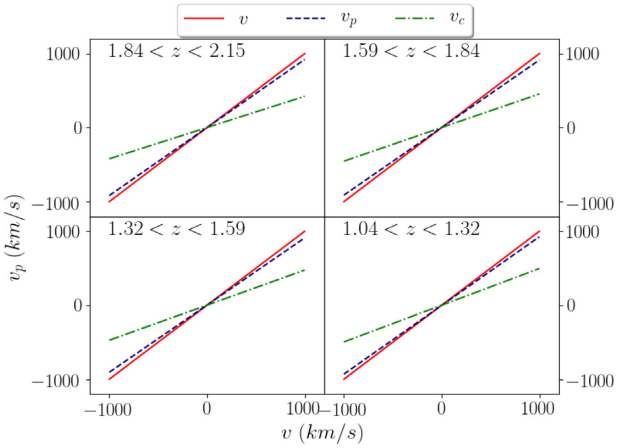

Figure 8. Linear fittings of results of Model I and the analytical method. The *x*- and *y*-axes show the true (v) and predicted (v_p) peculiar velocities, respectively. The red solid line shows the 1:1 ideal relation between the true and predicted velocities, the navy dashed line shows the linear fitting of Model I predictions, and the green dash–dotted line shows the fitting of the analytical result.

both the predictions using the deep learning neural network and the analytical method show strong correlations with the true peculiar velocity from the simulation. However, predictions of the analytical method have smaller magnitude than predictions of the deep learning neural network. The bias caused by the smaller magnitude becomes more obvious in Fig. 8. The fitting slope of the analytical method is around 0.47, which indicates a significant bias from expectations. Since predictions of the analytical method have no uncertainty, the fitting lines in Fig. 8 are not weighted by uncertainties. We found the performance of the analytical method improves when the integral distance dl is reduced, due to the fact that a longer line of sight path leads to more noise in the optical depth calculation. However, the deep learning result is still better than the analytical result even for 4 h^{-1} Mpc integral distance.

For the analytical method, the choice of the averaging area of the cluster is heuristic; therefore, the calculations of kSZ signals and optical depth are affected by the averaging radius or aperture. We set the averaging radius of the calculation to be the Virial radius of each cluster. However, this choice of averaging radius may miss some features of kSZ signals outside that radius. The deep learning algorithm, instead, provides a better approach for dealing with the image that it can extract more details about the cluster pattern from kSZ images with CNNs. Therefore, it provides less biased velocity predictions. In addition, the result of analytical method becomes worse with larger line of sight path due to the noise in the optical depth, while its effect on deep learning is negligible. These facts make the deep learning algorithm a more powerful tool for estimating peculiar velocities in observations.

6 PAIRWISE VELOCITY

Though our model trained by simulation data provides predictions with average uncertainties around 12 per cent, the average scatter from expectations is not ideal (38 per cent). In addition, the uncertainty using observational kSZ signals can be worse due to difficulties in detections. Therefore, ensemble statistics of peculiar velocities, rather than analysis of individual velocities, may be required. In this section, we apply pairwise velocity statistics to our predicted peculiar velocities.

Fig. 9 shows the pairwise velocity statistics of Model I trained by the data from multiple redshift slices. In the pairwise velocity calculation, we use all of the predicted velocities without any velocity limits. Although the uncertainty and scatter without velocity limits are larger, the pairwise velocity of predictions agree with the result of the true velocities with small uncertainties (error bars). The uncertainties of the pairwise velocities are calculated in two different ways: (1) subsampling method and (2) perturbation method that is perturbing the velocity catalogue 100 times by the MC dropout uncertainty and calculating the statistical error through the standard deviation of the 100 perturbed catalogues. In the figure, the error bars of the perturbation method are so small, that they are invisible.

7 ADAPTATION TO OBSERVATIONS

To test the feasibility of our model as applied to observations, we mimic observational kSZ signals by perturbing the simulated kSZ images with noise. We employ three types of noise in the perturbations: (1) Gaussian blur noise, (2) white noise, and (3) residual tSZ signals.

Fig. 10 shows the example images using different noise schemes. For (1) the Gaussian blur noise, we set the smoothing width as

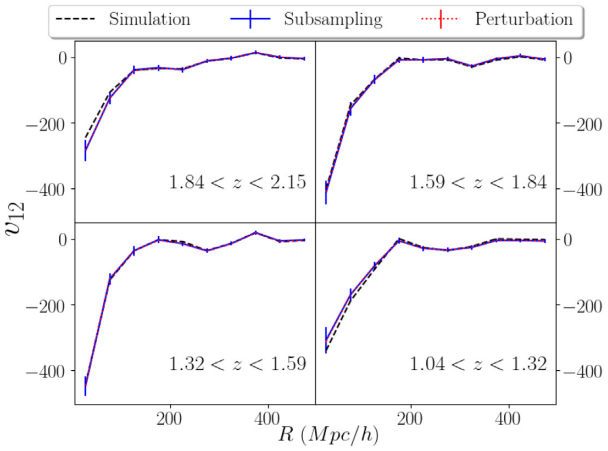

Figure 9. The pairwise velocity estimates of the true and predicted velocities from Model I without velocity limits. The black dashed line shows the result of the true velocities, the blue solid line indicates the result of the predicted velocities with error bars calculated by subsampling method, and the red dotted line indicates the result of predicted velocities with error bars calculated by perturbation method.

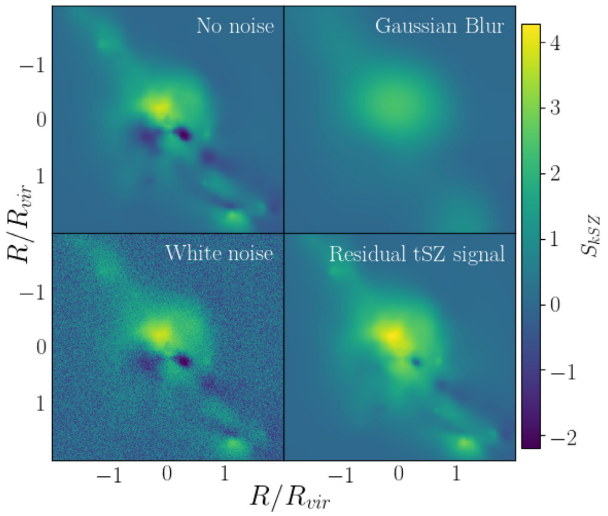

Figure 10. kSZ images with different noise schemes. $S_{\rm kSZ}$ indicates the re-scaled kSZ signal to increase the image contrast. The unit of $S_{\rm kSZ}$ is $\sinh^{-1}(\Delta T_{\rm kSZ}/T_{\rm CMB} \times 10^6)$.

40 per cent of the cluster Virial radius; thus, its width varies between observations. Here, we do not use any observation setting as a reference for the Gaussian blur width, since we are only testing the effect of its noise on the model with simulation data. For (2) the white noise scheme, we use Gaussian noise with standard deviation equal to the average value of the original kSZ signal, which means the signal-to-noise ratio equals one. Again, this ratio is only used for testing. For (3) the residual tSZ signal, which is a source of error in kSZ detections. We added 10 per cent tSZ signals (from the simulation) to the kSZ image to mimic the possible noise caused by the remnant tSZ signals in kSZ observations.

We test our model with these noise schemes and present the results in Fig. 11. One should notice that we implement two methods in the test: (1) the model is trained without noise but tested with the noisy images (blue scatters and navy dotted lines); (2) the model is both

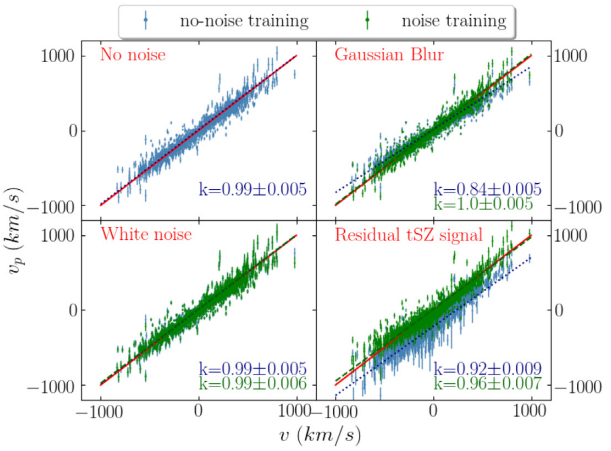

Figure 11. Predictions with different noise schemes. The x- and y-axes show the true (v) and predicted ($v_{\rm p}$) peculiar velocities, respectively. The models are trained by data of multiple redshift slices with (green) and without (blue) noise and tested by the noisy kSZ images. The red line shows the 1:1 ideal relation between the true velocity and the predicted velocity. The blue scatters and navy dotted lines show the results and uncertainty weighted fitting of models trained by kSZ images without noise and tested by the kSZ images with noise. The green scatters and dark green dashed line show the results and uncertainty weighted fitting of models trained and tested by noisy kSZ images.

trained and tested by the noisy images (green scatters and dark green dashed lines).

For method (1), our model shows great compatibility with the white noise. However, the prediction of adding residual tSZ noise shows biased results. We found that the bias caused by tSZ can be regarded as a constant shift from the ideal expectation. The larger the residual tSZ signal is, the larger the shift is. Therefore, the problem caused by the residual tSZ signal might be solved by making a simple correction. In contrast, the bias caused by the Gaussian blur drives the prediction magnitude smaller.

For method (2), the prediction of white noise shows no significant difference from method (1), while biases caused by the Gaussian blur and residual tSZ noise are improved significantly by training the model with noisy images. The improvement shows the capability of our model of dealing with noisy kSZ signals.

Due to the difficulties in kSZ and optical depth measurements, observational detections of kSZ for individual clusters is very rare, and the peculiar velocity estimated from kSZ observations is not accurate enough (Sayers et al. 2019) to train a deep learning model. Therefore, the only possibility for applying a deep learning neural network model to estimating peculiar velocities from kSZ observations is to train the model with perturbed simulation data. In this paper, we only tested three possible sources of uncertainties (Gaussian blur, white noise, and non-cleaned tSZ signal) in observations, and the noise intensities were set only for testing. A real kSZ observation may include different kinds of noise, such as noise from CMB anisotropies (Aghanim, Górski & Puget 2001) and dusty star formation galaxies. According to Mittal et al. (2018), the noise from CMB anisotropies and residual tSZ would not be the dominant sources of uncertainty for the Cerro Chajnantor Atacama Telescope (CCAT). Instead, the kSZ detection will be significantly affected by the image resolution and emission from dusty star formation galaxies. Therefore, to apply the deep learning algorithm to a specific observation (such as CCAT-prime), the simulated training data set would have to include noise that represents the corresponding observational conditions, which

will be studied in future work. Considering the advantages of deep learning neural networks over the analytical method (Section 5.2), estimating peculiar velocities from kSZ effect with deep learning algorithms is very promising. With upcoming kSZ detections, a suitable machine learning model for observational kSZ is foreseeable.

8 CONCLUSION

The analytical method of estimating peculiar velocities from the kSZ effect requires several steps, such as map filtering and optical depth calculation. In addition, the error in optical depth estimates makes it difficult to predict the peculiar velocity accurately.

In this paper, we test the feasibility of using deep learning neural networks to simplify the estimation of peculiar velocities from the kSZ effect. By comparing results of different redshift slices, using simulation data, we find that our deep learning model is redshift independent, which is consistent with theory.

Considering the relation between the tSZ effect and the optical depth, we build models that are trained by kSZ images (Model I) and kSZ + tSZ images (Model II). Those two models have similar predictions and uncertainties. We find that the average uncertainty of Model I is about 12 per cent and the average scatter is about 38 per cent. Although the average scatter is not ideal, the pairwise velocity of the predictions indicates that our model can provide reliable kSZ peculiar velocities for cosmological studies.

The similar results of Model I and Model II are caused by the small variation of the optical depth value. According to the current data and results, Model I provides more precise results than the analytical method with small optical depth variation. However, the Magneticum simulation is the only hydrodynamical simulation that is large enough to provide enough data for the deep learning training. Therefore, we are unable to show the improvement of Model II on the kSZ-velocity estimation with current data. While we believe that Model II should be able to deal with larger optical depth variations, we leave this for future studies.

We tested the feasibility of our model on observations by perturbing the kSZ signals with three different noise schemes: Gaussian blur, white noise, and residual tSZ noise. When using simulation training data and noisy validation data, the prediction with white noise shows few biases, while the biases caused by the Gaussian blur and residual tSZ noise are more significant. However, these biases can be improved by using noisy data for both training and testing. Our results clearly show that deep learning neural network can be used to estimate peculiar velocities from the kSZ effect with both simulations and observations. A possible way for applying deep learning neural network to observations is to train the model with simulated training data sets that include noise types particular to the observations being analysed. However, developing suitable models for observations will require more kSZ detection of individual galaxy clusters in the future.

In conclusion, using deep learning neural networks to estimate peculiar velocities from the kSZ effect is both feasible and promising. This method could simplify the analytical calculation of kSZ peculiar velocities significantly using only SZ input, which avoids the estimation of optical depth as well as map filtering.

ACKNOWLEDGEMENTS

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562.

NR's work at Argonne National Laboratory was supported under the U.S. Department of Energy contract DE-AC02-06CH11357.

HAF and RW were partially supported by NSF grant AST-1907404. An award of computer time was provided by the INCITE programme. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

KD acknowledge the support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2094 – 390783311.

The calculations were carried out at the Leibniz Supercomputer Center (LRZ) under the project 'pr86re'. We are especially grateful for the support by M. Petkova through the Computational Center for Particle and Astrophysics (C2PAP) and the support by N. Hammer at LRZ when carrying out the Box0 simulation within the Extreme Scale-Out Phase on the new SuperMUC Haswell extension system. NR thanks Jonás Chaves-Montero for discussions and help with the manuscript.

DATA AVAILABILITY

The data underlying this article will be shared on reasonable request to the corresponding author.

REFERENCES

Aghanim N., Górski K. M., Puget J.-L., 2001, A&A, 374, 1
Atrio-Barandela F., Kashlinsky A., Ebeling H., Kocevski D., 2012, preprint (arXiv:1211.4345)
Baron D., 2019, preprint (arXiv:1904.07248)
Beck A. M. et al., 2016, MNRAS, 455, 2110
Bhattacharya S., Kosowsky A., 2008, J. Cosmol. Astropart. Phys., 8, 030
Biffi V., Dolag K., Böhringer H., 2013, MNRAS, 428, 1395
Bocquet S., Saro A., Dolag K., Mohr J. J., 2016, MNRAS, 456, 2361
Borgani S., da Costa L. N., Zehavi I., Giovanelli R., Haynes M. P., Freudling W., Wegner G., Salzer J. J., 2000, AJ, 119, 102
Calafut V., Bean R., Yu B., 2017, Phys. Rev. D, 96, 123529
Chaves-Montero J., Hernandez-Monteagudo C., Angulo R. E., Emberson J. D., 2019, preprint (arXiv:1911.10690)
Diaferio A. et al., 2005, MNRAS, 356, 1477
Dolag K., Sunyaev R., 2013, MNRAS, 432, 1600
Dolag K., Hansen F. K., Roncarelli M., Moscardini L., 2005, MNRAS, 363, 29
Dolag K., Komatsu E., Sunyaev R., 2016, MNRAS, 463, 1797
Feldman H. et al., 2003, ApJ, 596, L131
Ferreira P. G., Juszkiewicz R., Feldman H. A., Davis M., Jaffe A. H., 1999, ApJ, 515, L1
Flender S., Bleem L., Finkel H., Habib S., Heitmann K., Holder G., 2016, ApJ, 823, 98
Flender S., Nagai D., McDonald M., 2017, ApJ, 837, 124
Gal Y., Ghahramani Z., 2015, preprint (arXiv:1506.02157)
Gal Y., Ghahramani Z., 2016, in Balcan M. F., Weinberger K. Q., eds, Proc. 33rd International Conference on Machine Learning. PMLR, USA, p. 1050
Gupta N., Saro A., Mohr J. J., Dolag K., Liu J., 2017, MNRAS, 469, 3069
Hand N. et al., 2012, Phys. Rev. Lett., 109, 041101
He S., Li Y., Feng Y., Ho S., Ravanbakhsh S., Chen W., Póczos B., 2019, Proc. Natl. Acad. Sci., 116, 13825
Hernandez-Monteagudo C., Chaves-Montero J., Angulo R. E., 2019, preprint (arXiv:1911.12056)
Hezaveh Y. D., Perreault Levasseur L., Marshall P. J., 2017, Nature, 548, 555
Hill J. C., Ferraro S., Battaglia N., Liu J., Spergel D. N., 2016, Phys. Rev. Lett., 117, 051301
Hirschmann M., Dolag K., Saro A., Bachmann L., Borgani S., Burkert A., 2014, MNRAS, 442, 2304

Hurier G., 2017, preprint (arXiv:1701.09072)
Juszkiewicz R., Ferreira P. G., Feldman H. A., Jaffe A. H., Davis M., 2000, Science, 287, 109
Kashlinsky A., Atrio-Barandela F., Kocevski D., Ebeling H., 2009, ApJ, 691, 1479
Kendall A., Gal Y., 2017, in Guyon I., Luxburg U. V., Bengio S., Wallach H., Fergus R., Vishwanathan S., Garnett R., eds, Advances in Neural Information Processing Systems. Curran Associates, Inc., United States, p. 5574
Kirillov A. A., Savelova E. P., 2019, Ap&SS, 364, 1
Komatsu E. et al., 2011, ApJS, 192, 18
Kumar A., Wang Y., Feldman H. A., Watkins R., 2015, preprint (arXiv:1512.08800)
Larson D. et al., 2011, ApJS, 192, 16
LeCun Y., Bengio Y., Hinton G., 2015, Nature, 521, 436
Levasseur L. P., Hezaveh Y. D., Wechsler R. H., 2017, preprint (arXiv:1708.08843)
Li Y.-C., Ma Y.-Z., Remazeilles M., Moodley K., 2018, Phys. Rev. D, 97, 023514
Lindner R. R., Aguirre P., Baker A. J., Bond J. R., Crichton D., Devlin M. J., Essinger-Hileman, 2015, ApJ, 803, 79
Metcalf R. B. et al., 2019, A&A, 625, A119
Mittal A., de Bernardis F., Niemack M. D., 2018, J. Cosmol. Astropart. Phys., 2, 032
Morningstar W. R., Hezaveh Y. D., Levasseur L. P., Blandford R. D., Marshall P. J., Putzky P., Wechsler R. H., 2018, preprint (arXiv:1808.00011)
Nguyen N.-M., Jasche J., Lavaux G., Schmidt F., 2020, preprint (arXiv:2007.13721)
Petrillo C. E., Tortora C., Chatterjee S., Vernardos G., Koopmans L. V. E., Verdoes Kleijn G., 2017, MNRAS, 472, 1129
Planck Collaboration XIII, 2014, A&A, 561, A97
Planck Collaboration XXXVII, 2016, A&A, 586, A140
Planck Collaboration LIII, 2018, A&A, 617, A48
Ragagnin A., Dolag K., Moscardini L., Biviano A., D'Onofrio M., 2019, MNRAS, 486, 4001
Ravanbakhsh S., Lanusse F., Mandelbaum R., Schneider J., Poczos B., 2016, preprint (arXiv:1609.05796)
Rephaeli Y., Lahav O., 1991, ApJ, 372, 21
Sayers J. et al., 2016, ApJ, 820, 101
Sayers J., Montaña A., Mroczkowski T., Wilson G. W., Zemcov M., Zitrin A., 2019, ApJ, 880, 45
Schaan E. et al., 2016, Phys. Rev. D, 93, 082002
Soergel B. et al., 2016, MNRAS, 461, 3172
Soergel B., Saro A., Giannantonio T., Efstathiou G., Dolag K., 2017, preprint (arXiv:1712.05714)
Springel V., 2005, MNRAS, 364, 1105
Springel V., Yoshida N., White S. D. M., 2001, New Astron., 6, 79
Sugiyama N. S., Okumura T., Spergel D. N., 2018, MNRAS, 475, 3764
Sunyaev R. A., Zeldovich Y. B., 1970, Ap&SS, 7, 3
Sunyaev R. A., Zeldovich Y. B., 1972, Comments Astrophys. Space Phys., 4, 173
Sunyaev R. A., Zeldovich I. B., 1980, MNRAS, 190, 413
Teklu A. F., Remus R.-S., Dolag K., Beck A. M., Burkert A., Schmidt A. S., Schulze F., Steinborn L. K., 2015, ApJ, 812, 29
Wagner-Carena S., Park J. W., Birrer S., Marshall P. J., Roodman A., Wechsler R. H., LSST Dark Energy Science Collaboration, 2021, ApJ, 909, 187
Wang Y., Rooney C., Feldman H. A., Watkins R., 2018, MNRAS, 480, 5332
Watkins R., Feldman H. A., Hudson M. J., 2009, MNRAS, 392, 743
Zhang P., Feldman H. A., Juszkiewicz R., Stebbins A., 2008, MNRAS, 388, 884

APPENDIX A: COMBINED kSZ AND tSZ MODEL

By adding the tSZ signal into deep learning neural network, we test the peculiar velocity predicted by the Model II. The prediction results of Model II trained by both kSZ and tSZ images of single and multiple redshift slices show negligible differences from the results of Model I. Fig. A1 shows the results of Model II using data of multiple redshift slices. Similar to Model I, the results of Model II are redshift independent. However, the prediction is not improved by adding tSZ information into the model. The similar performances between Model I and Model II shows that deep learning neural network could estimate the peculiar velocity accurately with only kSZ input, while simplifying the calculation significantly.

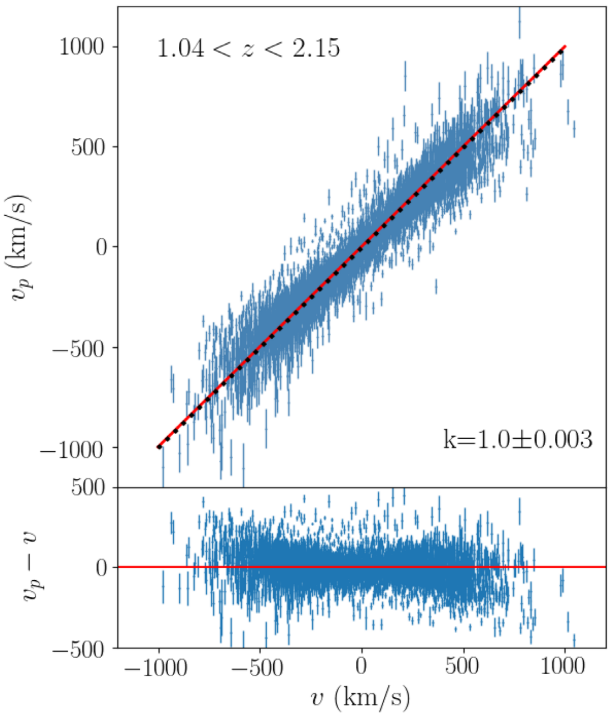

Figure A1. Same as Fig. 5 but for Model II trained by both kSZ and tSZ images.

APPENDIX B: MONTE CARLO DROPOUT UNCERTAINTY

For inputs $\boldsymbol{x}$ and target $\boldsymbol{y}$, a probabilistic neural network trained on data $\boldsymbol{D} \equiv (\boldsymbol{x}_{\rm train}, \boldsymbol{y}_{\rm train})$ predicts a probability distribution function $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{D})$. In order to learn this predictive distribution, sampling over all the model parameters $\boldsymbol{W}$, i.e. all the weights and the biases, are required. Using Bayes' theorem, the predictive distribution of the network shown in equation (B1) can be written in terms of the likelihood of the model $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{W})$.

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{D}) = \int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{W}) p(\boldsymbol{W}|\boldsymbol{D})\, {\rm d}\boldsymbol{W}. \quad \text{(B1)}$$

Here, $p(\boldsymbol{W}|\boldsymbol{D})$ is the distribution over all the network parameters. Due to large number of network parameters in deep neural network, the exact parametric posterior distribution $p(\boldsymbol{W}|\boldsymbol{D})$ is usually intractable. Instead of the determining this, the MC dropout technique Gal & Ghahramani (2016) relies on drawing different configurations of network parameters $\boldsymbol{W}_i$ from an approximate parametric distribution $\boldsymbol{W}_i \sim q(\boldsymbol{W}|\boldsymbol{D})$, as shown in equation (B2). Each dropout configuration is attained by randomly switching off

neurons in a trained neural network.

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{D}) \approx \int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{W}) q(\boldsymbol{W}|\boldsymbol{D})\, \mathrm{d}\boldsymbol{W} \approx \frac{1}{N_{\rm tot}} \sum_{i=1}^{N_{\rm tot}} p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{W}_i). \tag{B2}$$

Hence, different dropout configurations yield different predictive distributions. Each dropout configuration yields a different output by randomly switching neurons off and on with each forward propagation. The mean and variance of the outputs of the ensemble of the resulting neural networks can be computed as well. Hence, the MC dropout technique mitigates the problem of representing uncertainty in deep learning without sacrificing either computational complexity or test accuracy. On the other hand, it has to be noted that the MC dropout technique only provides the epistemic uncertainty in the neural networks (Levasseur et al. 2017), but the emerging distribution does not correspond to the full posteriors from the model. Recent progress by Wagner-Carena et al. (2021) and others explore the problem of using MC dropouts to get full Bayesian posteriors.

This paper has been typeset from a TEX/LATEX file prepared by the author.