Draft version February 11, 2021
Typeset using LaTeX **twocolumn** style in AASTeX63

Weak-lensing Mass Reconstruction of Galaxy Clusters with Convolutional Neural Network

Sungwook E. Hong (홍성욱) [1,2] Sangnam Park,[1] M. James Jee [3,4] Dongsu Bak [1,5] and Sangjun Cha[3]

[1] *Natural Science Research Institute, University of Seoul, 163 Seoulsiripdaero, Dongdaemun-gu, Seoul 02504, Republic of Korea*
[2] *Korea Astronomy and Space Science Institute, 776 Daedeok-daero, Yuseong-gu, Daejeon 34055, Republic of Korea*
[3] *Department of Astronomy, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea*
[4] *Department of Physics, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA*
[5] *Department of Physics, University of Seoul, 163 Seoulsiripdaero, Dongdaemun-gu, Seoul 02504, Republic of Korea*

(Received; Revised; Accepted)

Submitted to ApJ

ABSTRACT

We introduce a novel method for reconstructing the projected matter distributions of galaxy clusters with weak-lensing (WL) data based on convolutional neural network (CNN). Training datasets are generated with ray-tracing through cosmological simulations. We control the noise level of the galaxy shear catalog such that it mimics the typical properties of the existing ground-based WL observations of galaxy clusters. We find that the mass reconstruction by our multi-layered CNN with the architecture of alternating convolution and trans-convolution filters significantly outperforms the traditional reconstruction methods. The CNN method provides better pixel-to-pixel correlations with the truth, restores more accurate positions of the mass peaks, and more efficiently suppresses artifacts near the field edges. In addition, the CNN mass reconstruction lifts the mass-sheet degeneracy when applied to sufficiently large fields. This implies that this CNN algorithm can be used to measure cluster masses in a model independent way for future wide-field WL surveys.

1. INTRODUCTION

Weak-lensing (WL) is now firmly established as the most direct method to measure the mass of an astrophysical object ranging from a galaxy (galaxy-galaxy lensing) to the cosmological large scale structure (cosmic shear). Many on-going and future WL surveys happening on massive scales reflect the elevated level of interest and confidence in this technique (e.g., Hikage et al. 2019; Troxel et al. 2018; Ivezić et al. 2019; Laureijs et al. 2011; Spergel et al. 2015). Without question, in order to maximize the scientific return from these huge data volume, our highest priority is to understand and control systematics. A number of issues on WL systematics have been identified, including shear calibration, photometric redshift degeneracy, model bias, mass-sheet degeneracy, astrophysical processes, and so on (e.g. Gorenstein et al. 1988; Seitz & Schneider 1996; Squires & Kaiser 1996; High et al. 2007; Jarvis et al. 2008; Meyers & Burchat 2015; Mandelbaum et al. 2015).

In this study, we focus on systematics arising in galaxy cluster mass reconstruction from WL source catalogs. Although galaxy cluster mass reconstruction is one of the earliest WL application and demonstration of its power, the main utility of the two-dimensional mass reconstruction has been rather qualitative investigation of the relative mass distribution of the target field. Very few studies employed the mass reconstruction for quantitative analysis (e.g., derivation of galaxy cluster masses). This is because the current mass reconstruction algorithms suffer from various artifacts. For example, the so-called mass-sheet degeneracy (invariant of the observable shear under a certain linear rescaling of the mass) is one of the major obstacles that prevent us from interpreting the result in absolute terms. Severe nonlinearity arising from the transformation of the reduced shear to the convergence is also a crucial contributing factor. Other critical issues include finite-field effect, ill-posed mathematical inversion, smoothing artifact, field edge systematics, and so on (e.g., Bartelmann 1995; Seitz & Schneider 1996)

The most popular method to estimate the cluster mass so far has been to fit an analytic profile to the observed

Corresponding author: M. James Jee
mkjee@yonsei.ac.kr

shear. This assumes that the cluster mass distribution is spherically symmetric and follows a particular halo model favored by numerical simulations such as an Navarro-Frenk-White (Navarro et al. 1996) profile. Although this provides a method to overcome the aforementioned drawbacks of the mass reconstruction, the obvious weakness is that individual galaxy clusters do not exactly follow the analytical description, much less are consistent with the assumption of spherical symmetry.

In this paper, we introduce a novel method for mass reconstruction based on convolutional neural network (CNN). CNN is a branch of deep learning, which has been considered to be a promising tool in many fields of astronomy in recent years such as photometric redshift (e.g., Schaefer et al. 2018), strong-lensing finding (e.g., Pasquet et al. 2019), image deconvolution (e.g., Flamary 2017), star-galaxy separation (e.g., Kim & Brunner 2017), morphological classification (e.g., Mittal et al. 2019), etc. The current study is the first endeavor to apply CNN to WL mass reconstruction of galaxy clusters. Since there is a rapid growth in data size and complexity from future WL surveys, the approach introduced here will find many useful applications if our CNN algorithm can significantly reduce aforementioned systematics found in the traditional mass reconstruction.

This paper is organized as follows. In §2, we describe the basic theory, CNN architecture, and training data sets. The performance of our CNN mass reconstruction is presented in §3 and discussed in §4 before we conclude in §5. Throughout the paper, we assume a flat ΛCDM cosmology with $H_0 = 70$ km s^{-1} Mpc^{-1}, $\Omega_\Lambda = 0.7$, and $\Omega_\mathrm{m} = 0.3$.

2. METHODS

2.1. *Basic Weak-lensing Theory*

WL formalism is valid in the regime where the source galaxy is much smaller than the characteristic scale of the gravitational potential variation. In this regime, the transformation matrix $\mathbf{A}$ relating the source plane position $\mathbf{x}$ to the image plane position $\mathbf{x}'$ via $\mathbf{x}' = \mathbf{A}\mathbf{x}$ is described by:

$$\mathbf{A} = (1-\kappa)\begin{pmatrix} 1-g_1 & -g_2 \\ -g_2 & 1+g_1 \end{pmatrix}, \tag{1}$$

where $g_{1(2)}$ is the first (second) component of the reduced shear $g = (g_1^2 + g_2^2)^{1/2}$ and κ is the convergence. The reduced shear g is related to shear γ and convergence κ via

$$g = \gamma/(1-\kappa). \tag{2}$$

The convergence κ is the unitless surface mass density:

$$\kappa = \frac{\Sigma}{\Sigma_c}, \tag{3}$$

where Σ_c is the critical surface mass density:

$$\Sigma_c = \frac{c^2 D_s}{4\pi G D_l D_{ls}}. \tag{4}$$

In equation (4), c is the speed of light, G is the gravitational constant, D_l is the angular diameter distance to the cluster (lens), D_{ls} is the angular diameter distance from lens to source, and D_s is the angular diameter distance to source.

The transformation matrix $\mathbf{A}$ in equation (1) converts a circle into an ellipse. There are multiple ways to define the ellipticity of the resulting ellipse, which has been a source of confusion. If we let its semi-major and -minor axes be a and b, respectively, one can show that the reduced shear g in equation (1) becomes

$$g = \frac{a-b}{a+b}. \tag{5}$$

Therefore, it is convenient to use equation (5) to define the ellipticity in WL, which we also adopt in this paper. Since g alone cannot express the orientation of the ellipse, the WL community often uses the complex notation:

$$\mathbf{g} = g_1 + \mathbf{i}g_2, \tag{6}$$

which provides both magnitude $g = (g_1^2 + g_2^2)^{1/2}$ and orientation $\phi = 0.5\tan^{-1}(g_2/g_1)$ of the elongation.

Under the assumption that we can assign a unique ellipticity to every galaxy, the same complex notation (equation 6) can also be used to express its intrinsic ellipticity $\mathbf{e} = e_1 + \mathbf{i}e_2$ prior to WL distortion. Then, the transformation of the intrinsic ellipticity $\mathbf{e}$ to the lensed (distorted) ellipticity $\boldsymbol{\epsilon}$ by the reduced shear $\mathbf{g}$ is given by:

$$\boldsymbol{\epsilon} = \frac{\mathbf{e} + \mathbf{g}}{1 + \mathbf{g}^*\mathbf{e}}, \tag{7}$$

where $\mathbf{g}^*$ denotes the complex conjugate of $\mathbf{g}$.

Inspection of equation (7) reveals that in general each galaxy's lensed ellipticity $\boldsymbol{\epsilon}$ is only slightly different from its intrinsic ellipticity $\mathbf{e}$ in the WL regime where $\mathbf{g}$ is small. When we disregard measurement systematics and assume that the ellipticity distribution of the source population is isotropic, one can show that the unbiased estimator for $\mathbf{g}$ is $\langle\boldsymbol{\epsilon}\rangle$.

2.2. *Conventional Mass Reconstruction and its Limitation*

The mathematical relation between $\boldsymbol{\gamma}$ shear and convergence κ at the position $\mathbf{x}$ is:

$$\boldsymbol{\gamma}(\mathbf{x}) = \frac{1}{\pi} \int \mathbf{D}(\mathbf{x}-\mathbf{x}')\kappa(\mathbf{x}')\mathrm{d}\mathbf{x}', \tag{8}$$

where the kernel $\mathbf{D}$ is:

$$\mathbf{D} = -\frac{1}{(x_1 - \mathbf{i}x_2)^2}. \tag{9}$$

The well-known Kaiser & Squires (1993, KS93) mass reconstruction is based on the straightforward inversion of equation (10):

$$\kappa(\mathbf{x}) = \frac{1}{\pi} \int \mathbf{D}^*(\mathbf{x}-\mathbf{x}')\boldsymbol{\gamma}(\mathbf{x}')\mathrm{d}\mathbf{x}'. \tag{10}$$

KS93 evaluate this convolution in Fourier space while Fischer & Tyson (1997) develop an inversion method in real space. Alternatively, some authors propose to reconstruct the convergence field using equation (8) through the maximum likelihood method (e.g., Seitz et al. 1998; Bradač et al. 2004; Jee et al. 2007).

Inspection of equations (8) and (10) shows that several artifacts may be introduced from the KS93 mass reconstruction. First, the evaluation of the convolution suffers from the so-called finite-inversion problem because in principle κ requires the information of the shear $\boldsymbol{\gamma}$ over an infinite area. Second, equation (10) uses $\boldsymbol{\gamma}$ for its input whereas the directly attainable information from averaging over many galaxy shapes is only $\mathbf{g}$. Third, the solution is not unique as the same equation holds under the transformation: $\kappa \rightarrow \lambda\kappa + (1-\lambda)$, where λ is arbitrary. This ambiguity is often termed the "mass-sheet degeneracy" because, although mathematically somewhat misleading, one can view the transformation as an addition of a thin sheet of mass when $\lambda \approx 1$. Fourth, in the central region of massive clusters where the WL assumption no longer holds, the above equations lose their validity. Fifth, to suppress noise amplification in the inversion, we can only obtain a smoothed convergence field, which gives a biased mass estimate even in the ideal situation where all other issues are carefully accounted for.

2.3. *Mass Reconstruction with Convolutional Neural Network*

2.3.1. *Generation of Training Dataset*

Generation of our training dataset starts from the convergence (κ) maps created from cosmological simulations via ray tracing. We utilize the publicly available data MassiveNuS (Liu et al. 2018). The simulation was originally designed to investigate the impact of massive neutrinos on the large-scale structure. For the current investigation, we chose to retrieve[1] the dataset corresponding to the $\sum m_\nu = 0.177$ eV, $\Omega_\mathrm{m} = 0.2485$, and $A_s = 2.0644 \times 10^{-9}$ setting. However, we emphasize that the details in the simulation parameters are not important within the scope of the current study because, as we shall demonstrate later, our CNN algorithm is designed to learn the rule, which maps the reduced shear field to the convergence field according to general relativity and thus is independent of the above parameter values.

The dataset consists of a total of 50,000 convergence images at five (z = 0.5, 1.0, 1.5, 2.0, and 2.5) different source redshifts (10,000 convergence fields per source redshift) each simulating an area of 3.5° × 3.5°, which matches the field of view (FOV) of the Vera C. Rubin Observatory (Ivezić et al. 2019). For the same field, a higher source redshift convergence image is richer in substructure as more line-of-sight (LOS) structure is included and also the lensing efficiency becomes higher.

We identified clusters by running SExtractor (Bertin & Arnouts 1996) on the convergence field image and cropped a $32' \times 32'$ region approximately centered on each cluster. We randomly generated positions of 25,000 sources within this subfield. The distribution matches the typical source density of ~25 per sq. arcmin in our previous Subaru WL studies (e.g., Finner et al. 2017; Kim et al. 2019; Yoon et al. 2020). Shears $\boldsymbol{\gamma}$ at the position $\mathbf{x}$ were computed using equation (8). This shear $\boldsymbol{\gamma}$ is then converted to the reduced shear $\mathbf{g}$ through $\mathbf{g} = \boldsymbol{\gamma}/(1-\kappa)$. Here we do not consider dispersions in both lens and source redshift and assume that all lensing masses and sources are confined to $z_l = 0.5$ and $z_s = 1.5$, respectively.

We set the intrinsic shape noise per component to $\sigma_e = 0.24$, which is approximately the empirical value from *Hubble Space Telescope* (HST) image analysis. In addition to this shape noise, there is a measurement error due to pixel noise. Assuming that the measurement error is independent of the shape noise, we produced the total ellipticity error as a sum of two Gaussian random numbers. The measurement error depends on galaxy properties (e.g., magnitude, size, profile shape, etc.) and signal-to-noise ratios (S/N). We adopted the source magnitude distribution of the Kim et al. (2019) study, which starts at ~21.5^{th} mag, peaks at ~25.5^{th} mag, and truncates at ~27.5^{th} mag in V-band. The observed relation between magnitude and ellipticity measurement error in Kim et al. (2019) is employed to generate the ellipticity measurement error for our sources.

[1] http://columbialensing.org

Table 1. Outline of our convolutional neural network. See text and Figure 1 for description of each layer.

Layer	Filter Size	Multiplied to	Output Size
Input	-	-	(2, 386, 386)
Conv2D-1	(3, 3)	-	(8, 384, 384)
AvgPool	(3, 3)	-	(8, 128, 128)
TransConv2D-1	(49, 49)	-	(8, 176, 176)
Conv2D-2	(49, 49)	-	(8, 128, 128)
Multiply-1	-	AvgPool	(8, 128, 128)
TransConv2D-2	(49, 49)	-	(8, 176, 176)
Conv2D-3	(49, 49)	-	(8, 128, 128)
Multiply-2	-	Multiply-1	(8, 128, 128)
TransConv2D-3	(49, 49)	-	(8, 176, 176)
Conv2D-4	(49, 49)	-	(8, 128, 128)
Multiply-3	-	Multiply-2	(8, 128, 128)
TransConv2D-4	(49, 49)	-	(8, 176, 176)
Conv2D-5	(49, 49)	-	(8, 128, 128)
Multiply-4	-	Multiply-1	(8, 128, 128)
TransConv2D-5	(49, 49)	-	(16, 176, 176)
Output	(49, 49)	-	(1, 128, 128)

We select 7,000 convergence fields and divide them into 5,000 training, 1,800 validation, and 200 test samples.

2.3.2. *Architecture of CNN*

Figure 1 and Table 1 summarize the architecture of our CNN model that we use to predict the convergence map from the WL shear datasets. Our CNN model takes two-dimensional (2D) arrays of $\epsilon_1(\mathbf{x})$, $\epsilon_2(\mathbf{x})$, and $\Delta\epsilon(\mathbf{x})$ as a three-channel input, where $\mathbf{x}$ denotes 2D pixel coordinates, $\epsilon_i(\mathbf{x})(i=1,2)$ and $\Delta\epsilon(\mathbf{x})$ are the i-th average ellipticity (reduced shear) component and its error at the position $\mathbf{x}$. Because we randomly positioned source galaxies (§2.3.1), these (regularly spaced) input grids were constructed by smoothing the (irregularly spaced) source catalogs with a FWHM$=7''$ Gaussian kernel. For each cluster, the full area of the initial field is $32' \times 32'$, which is represented by 2D arrays of 500×500. We performed data augmentation by subsampling $24\farcm7 \times 24\farcm7$ (386×386) regions 1,444 times. In addition, we applied four rotations (0°, 90°, 180°, and 270°) and two axis flips. The total number of the resulting subfields for each cluster is 1,444×(4 + 2) =8,664. The subsampling scheme also prevents CNN from learning that the position of the cluster is always at the field center.

The main part of our CNN architecture includes the repeated combinations of 2D convolution (Conv2D-# in Table 1) and transposed-convolution layers (TransConv2D-#) with the identical filter size. We tested various choices of filter sizes and found that the 49×49 filter gives the best overall performance. Readers are referred to Appendix A for performance comparisons among different CNN architectures with various choices of filter sizes, input layers, and loss functions. The Conv2D-# and TransConv2D-# operations are activated by the hyperbolic tangent function (tanh). Inspired by the residual neural network (ResNet; He et al. 2015), we use skip-connections between the output of each Conv2D-# layer and that of the previous layer with the matching output size by applying a multiplication operation (Multiply-#). We apply batch normalization to the output of each Multiply-# layer to avoid the so-called gradient vanishing problem (Ioffe & Szegedy 2015). These repeated operations of Conv2D-# and TransConv2D-# are designed to extract features while preserving the size of the output layer (without introducing any padding). Also, this architecture outperforms the other architectures that have only convolution layers when it comes to the prediction of the mass peak positions (Appendix A). Although we did not use any arbitrary padding, the convergence estimates near the field boundary can easily be influenced by the non-vanishing filter size. Therefore, the values within the 14 boundary pixels were not used during our training. The pixel scale of the final κ map is $0\farcs192$ pixel^{-1}.

We performed our CNN training with Tensorflow (Abadi et al. 2015). During the training, we used the Adam optimizer (Kingma & Ba 2014) with a learning rate 10^{-5}, 50 mini-batches per each step, and 20 steps per each epoch. In this study, we introduce the following weighted mean square error inspired by the focal loss (Lin et al. 2017):

$$\mathcal{L} = \sum_{\mathbf{x}} \omega_{\rm f}(\mathbf{x}) \left[\kappa_{\rm pred}(\mathbf{x}) - \kappa_{\rm truth}(\mathbf{x})\right]^2 , \qquad (11)$$

where the weight $\omega_{\rm f}(\mathbf{x})$ at each pixel is determined by the value of truth convergence:

$$\omega_{\rm f}(\mathbf{x}) = 1 + \frac{|\kappa_{\rm truth}(\mathbf{x})|}{\max(\kappa_{\rm truth})} . \qquad (12)$$

This loss function is chosen so that our CNN model is mostly constrained by the high-density regions of clusters, where our scientific interests lie (see Appendix A for performance comparisons). We ran 200 epochs with the NVIDIA V100 GPU, which take about one hour per each training. For the convergence test of our CNN training, we executed 10 independent runs with the same CNN architecture. For each run, we adopted the model that minimized the validation loss. In our presentation of the results (§3), the standard deviations from the 10 runs are used as error estimates.

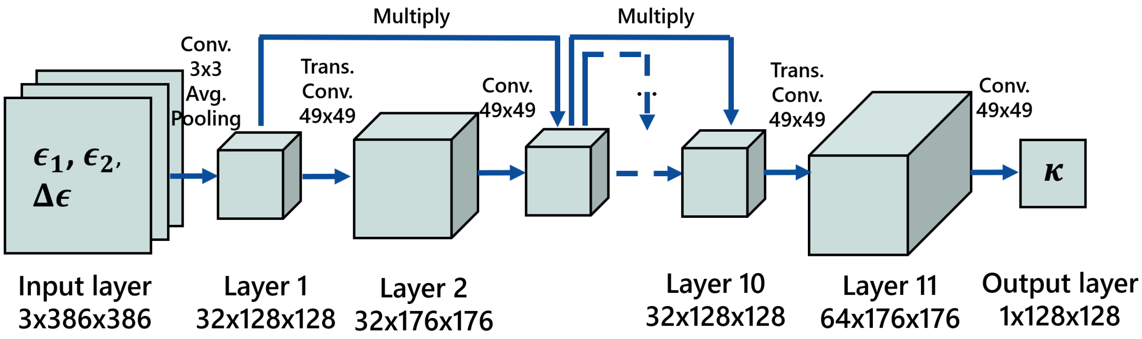

Figure 1. Schematic diagram showing the architecture of our convolutional neural network. The input channel consists of three (ϵ_1, ϵ_2, and $\Delta\epsilon$) layers of 2D (386 × 386) arrays. The main part of the CNN architecture is the repeated combination of convolution and transposed-convolution with 49×49-size filters. We use skip-connections between the output of each convolution layer and that of the previous layer that has the matching size. See Table 1 and text for details.

3. RESULTS

In this section, we compare our CNN results with those of KS93 for the test sample comprised of 200 cluster fields. Because of the data augmentation procedure (§2.3), multiple (subsampled) mass maps are produced for each cluster. Thus, we created one mosaic convergence image for each cluster by taking average of the multiple mass maps. The cluster is approximately located at the center in this mosaic and we use the mosaic for comparison with the truth and KS93 results. Readers are reminded that since these multiple mass maps are generated from the same source catalog, this mosaicking procedure does not benefit by reducing the statistical noise[2]. In §3.1, we use visual inspection to qualitatively compare the reconstruction results. In §3.2, we contrast the values of reconstructed convergence with those of the truth by pixel-by-pixel comparison and by evaluating their probability distributions. In §3.3 and §3.4, we investigate the reconstructed cluster masses and the positional accuracy of their density peaks, respectively. Finally, we examine performances of our CNN method in the presence of bright stars in §3.5. Table 2 summarizes our comparison between the KS93 and CNN results.

3.1. *Qualitative Comparison Based on Visual Inspection*

Figure 2 displays an example of our CNN mass reconstruction. The comparison with the truth and KS93 results illustrates that our CNN mass reconstruction is superior to KS93 in terms of 1) the recovery of the true κ range, 2) the representation of the large-scale structure around the cluster, and 3) the suppression of the noise in the cluster outskirts. Although here only one case is illustrated, these advantages are present for the rest of the test sample.

1) **Recovery of the κ range:** The truth map shows that κ ranges from ~0.05 to ~0.25, where the maximum value is found at the cluster center. The KS93 reconstruction fails to recover this convergence range in the high end. The convergence value at the cluster center is only ~0.05 while the global maximum is found at a different location (see the location of the "X" symbol). On the other hand, the CNN mass reconstruction gives a much higher value $\kappa \sim 0.15$ at the cluster center. Given the inevitable smoothing effect arising from the sparse sampling (25 sources per sq. arcmin), we believe that the improvement over the KS93 is remarkable. We discuss this issue more quantitatively in §3.2.

2) **Reconstruction of the large scale structure:** Inspection of the truth map (Figure 2) indicates that the cluster is not isolated, but located in the high density environment. While it is difficult to trace this large-scale structure surrounding the cluster in the KS93 result, the feature, albeit somewhat smoothed, clearly stands out as an overdense region in our CNN mass reconstruction.

3) **Suppression of noise:** Since the S/N value depends on the local strength of WL signal given the same number density of sources, an ideal mass reconstruction method should employ a smoothing scheme where the kernel size matches the local S/N value. However, in general, it is nontrivial to implement such an "adaptive smoothing" scheme in practice because the S/N information is only obtained after a high-quality convergence field is reconstructed. Therefore, a common practice in the

[2] That is, one can apply the procedure to real observations.

Table 2. Summary of the performances of our CNN and the KS93 mass reconstructions for the 200 test datasets. See text for the definitions of each metric.

Method	$\mathcal{R}(\kappa_{\rm pred}, \kappa_{\rm truth})$	$\mathcal{D}(\widetilde{\kappa}_{\rm pred}, \widetilde{\kappa}_{\rm truth})$	$M^{\rm cl}_{\rm pred}/M^{\rm cl}_{\rm truth}$	$\Delta_{\rm peak}$ [pixel]
KS93	0.253 ± 0.131	6.26 ± 4.57	$0.484^{+0.218}_{-0.149}$	$22.26^{+44.94}_{-20.55}$
CNN	0.407 ± 0.288	4.36 ± 3.77	$0.867^{+0.327}_{-0.296}$	$3.10^{+25.64}_{-2.00}$
CNN-BS	0.276 ± 0.179	3.66 ± 3.22	$0.554^{+0.257}_{-0.211}$	$7.95^{+21.94}_{-5.94}$

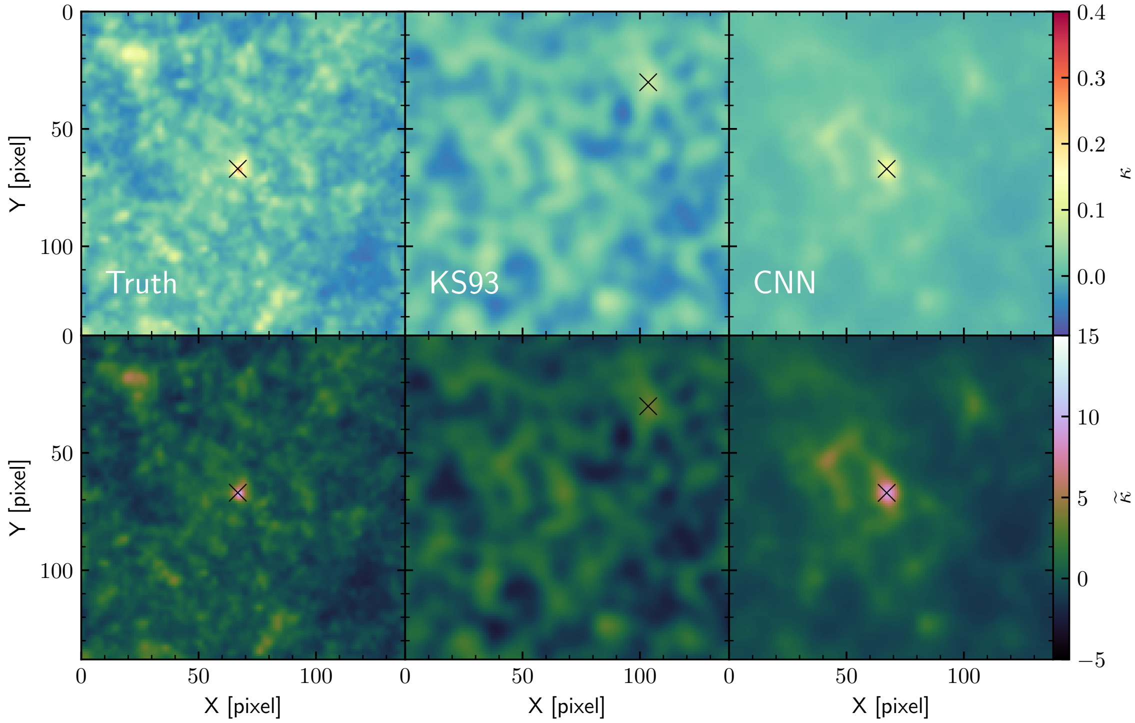

Figure 2. Example of our CNN mass reconstruction. We show the truth convergence map (left) and the reconstructions with the KS93 (middle) and our CNN methods (right). The top panel displays the convergence values $\kappa(\mathbf{x})$ as are whereas the bottom panel shows the rescaled versions using the transformation $\widetilde{\kappa}(\mathbf{x}) \equiv (\kappa(\mathbf{x}) - \langle\kappa\rangle)/\Delta\kappa$, where $\langle\kappa\rangle$ and $\Delta\kappa$ are the average and standard deviation, respectively, evaluated within the field. The "X" symbol denotes the location of the highest value within each convergence field. Here we only display the central $26'.6 \times 26'.6$ region. Visual inspection shows that our CNN reconstruction significantly outperforms the KS93 method in terms of the dynamical range restoration, noise suppression, and large-scale structure representation.

WL community is to perform mass reconstruction with a fixed-size smoothing kernel often optimized for the central region of the cluster (e.g., van Waerbeke 2000). The inevitable artifact is the production of many spurious mass peaks in the cluster outskirt where the S/N value is low. The comparison between the KS93 and our CNN results shows that our CNN result nicely suppresses the noise fluctuation in the outskirt region while still detecting substructures if they are significant (see the bottom panel of Figure 2).

3.2. *Convergence Distribution*

The literature has shown that the distribution of the convergence field can be well-approximated by a log-normal distribution characterized by an extended high-end tail (e.g., Jain et al. 2000; Hilbert et al. 2011; Clerkin et al. 2017). Figure 3 compares the convergence distributions between the KS93 and our CNN reconstructions for the entire test sample and shows that the CNN distribution follows the log-normal trend of the truth (see the left panel). On the other hand, the convergence distribution in the KS93 result is symmetric around zero without any sign of an extended tail at the high end.

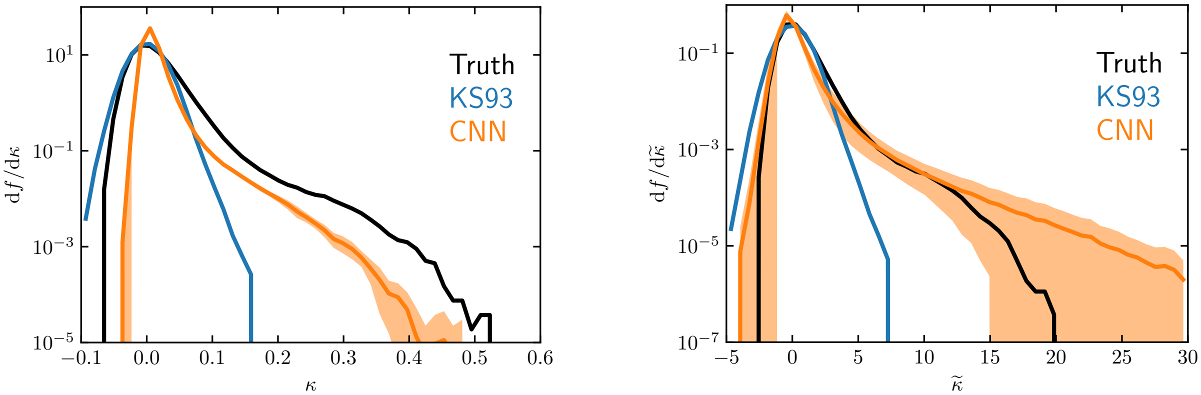

Figure 3. Comparison of convergence (κ) distributions. We measure the κ distribution from the entire test sample (left). The right panel is the same except that the distribution is obtained for $\widetilde{\kappa}$. The orange shade represents the standard deviation measured from our 10 independent runs. The CNN reconstruction provides an extended tail at the high end, mimicking the feature in the truth whereas the KS93 distribution is nearly symmetric around zero. When the convergence is rescaled with its standard deviation, the agreement improves as can be seen in the right panel.

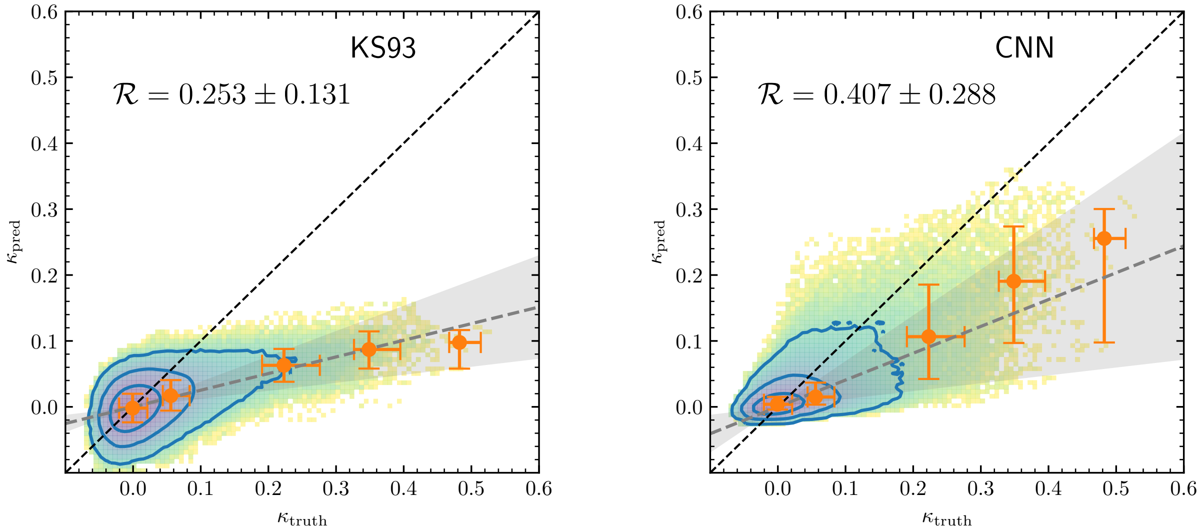

Figure 4. Joint distribution of κ between reconstructed and truth convergence fields. Contours show the 1σ, 2σ, and 3σ confidence levels. We divide the $\kappa_{\rm truth}$ range into five bins in such a way that all bins contain the same number of κ pixels. Orange filled circles are the medians of these bins and their error bars represent the 1σ certainties in $\kappa_{\rm truth}$ and $\kappa_{\rm pred}$. Gray dashed lines and shaded area are the average and standard deviation of the $\mathcal{R}$ metric (the weighted version of the ratio $\kappa_{\rm pred}/\kappa_{\rm truth}$). The CNN reconstruction shows an improved correlation with the truth, although it is also clear that the κ values are underestimated.

The comparison between the CNN result and the truth shows that the CNN distribution is somewhat narrower. This happens because the CNN mass map based on a finite number of source galaxies (25 galaxies per sq. arcmin) is inevitably smoother than the truth. In order to compensate for this smoothing effect, we propose the following normalization:

$$\widetilde{\kappa}(\mathbf{x}) \equiv (\kappa(\mathbf{x}) - \langle\kappa\rangle)/\Delta\kappa, \tag{13}$$

where $\langle\kappa\rangle$ and $\Delta\kappa$ are the average and standard deviation, respectively. The rescaling of the convergence through this normalization takes into account the reduction of $\Delta\kappa$ in mass reconstruction. However, as mentioned in §3.1, the smoothing kernel is not uniform in the CNN mass reconstruction; effectively, the cluster outskirts smoothed with larger kernels than the cores. Therefore, the proposed normalization (equation 13) does not completely resolve the issue. Nevertheless, the right panel of Figure 3 shows that the agreement between the CNN and truth distributions improves dramatically after the normalization.

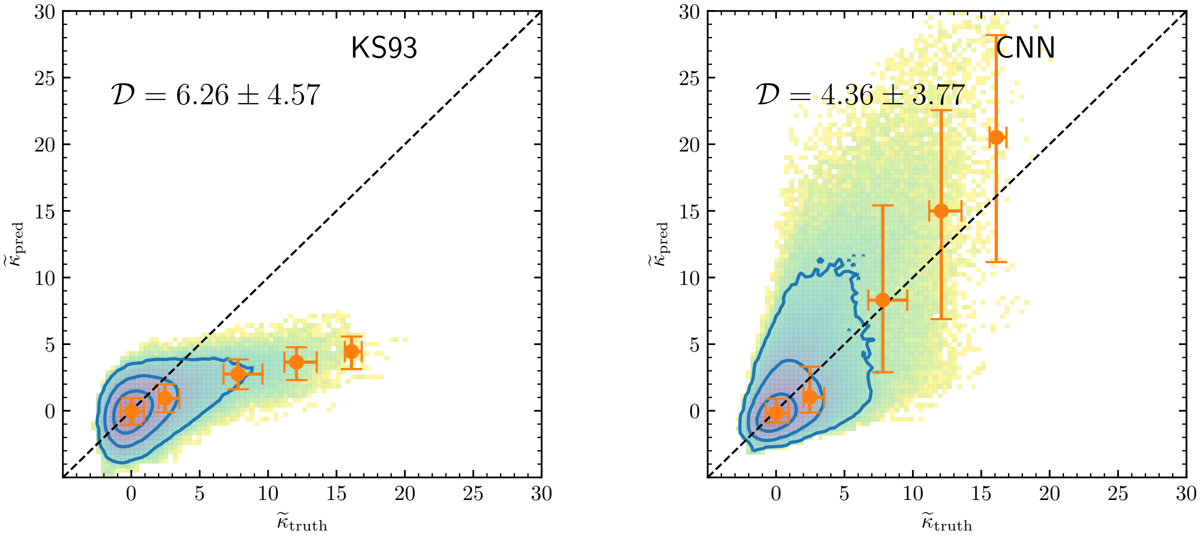

Figure 5. Similar to Figure 4 except that the plots are drawn with the normalized convergence ($\widetilde{\kappa}$).

The joint distribution shown in Figure 4 also confirms that the CNN mass reconstruction provides significantly better pixel-to-pixel correlations with the truth. Also, similarly to the previous case, the normalization significantly strengthens the correlation with the truth (Figure 5).

To quantify the similarity of the reconstruction to the truth, one can suggest the ratio $\kappa_{\rm pred}(\mathbf{x})/\kappa_{\rm truth}(\mathbf{x})$ and the absolute deviation $|\widetilde{\kappa}_{\rm pred}(\mathbf{x}) - \widetilde{\kappa}_{\rm truth}(\mathbf{x})|$ as potential metrics. However, these metrics, if used as they are, would be dominated by the statistics of the convergence pixels near zero. Therefore, we introduce the weighted versions as follows:

$$\mathcal{R} = \frac{\sum_{\mathbf{x}} \omega_{\rm p}(\mathbf{x})[\kappa_{\rm pred}(\mathbf{x})/\kappa_{\rm truth}(\mathbf{x})]}{\sum_{\mathbf{x}} \omega_{\rm p}(\mathbf{x})} \tag{14}$$

$$\mathcal{D} = \frac{\sum_{\mathbf{x}} \widetilde{\omega}_{\rm p}(\mathbf{x})\,|\widetilde{\kappa}_{\rm pred}(\mathbf{x}) - \widetilde{\kappa}_{\rm truth}(\mathbf{x})|}{\sum_{\mathbf{x}} \widetilde{\omega}_{\rm p}(\mathbf{x})}, \tag{15}$$

where the weights $\omega_{\rm p}(\mathbf{x})$ and $\widetilde{\omega}_{\rm p}(\mathbf{x})$ are inversely proportional to the probability distribution:

$$\frac{1}{\omega_{\rm p}(\mathbf{x})} = \left.\frac{{\rm d}f}{{\rm d}\kappa_{\rm truth}}\right|_{\kappa_{\rm truth}(\mathbf{x})} \tag{16}$$

$$\frac{1}{\widetilde{\omega}_{\rm p}(\mathbf{x})} = \left.\frac{{\rm d}f}{{\rm d}\widetilde{\kappa}_{\rm truth}}\right|_{\widetilde{\kappa}_{\rm truth}(\mathbf{x})}. \tag{17}$$

In practice, the weights can diverge when noise makes the derivatives close to zero. To prevent this, we use discrete histograms of the truth with 50 bins for the estimation of $\omega_{\rm p}$ and $\widetilde{\omega}_{\rm p}$. In addition, for the evaluation of equation (14), we iteratively update the weighted average and standard deviation from the 3σ-clipping of $\kappa_{\rm pred}(\mathbf{x})/\kappa_{\rm truth}(\mathbf{x})$ using the mean of the previous step. The $\mathcal{R}$ (better if closer to unity) and $\mathcal{D}$ (better if closer to zero) metrics from the CNN mass reconstruction are 0.407 ± 0.288 and 4.36 ± 3.77, respectively. On the other hand, the KS93 result gives $\mathcal{R} = 0.253 \pm 0.131$ and $\mathcal{D} = 6.26 \pm 4.57$. Both metrics indicate that the κ statistics from the CNN reconstruction better match those from the truth.

3.3. *Projected Cluster Mass*

The pixel-to-pixel comparison in §3.2 shows that although our CNN mass reconstruction better recovers the convergence statistics of the truth than the KS93 method, the distribution is somewhat narrower because of the smoothing implicitly applied to the reconstructed convergence field via CNN. It is our premise that this smoothing artifact is of a less concern when one's interest is to estimate the integrated convergence within a reasonably large aperture. We define the projected cluster mass $M^{\rm cl}_{\rm truth}$ to be the sum of the convergence values within the $r = 1\farcm92$ (10 convergence pixels) radius aperture. At the cluster redshift of 0.5, the radius corresponds to 0.72 Mpc with the adopted cosmology.

Figure 6 shows the comparison of $M^{\rm cl}_{\rm truth}$ between the reconstructed and the truth values. As seen in the pixel-to-pixel comparison, the CNN mass reconstruction also outperforms the KS93 result in the cluster mass estimation. In addition, it is remarkable that the agreement with the truth is significantly better than the one in the convergence pixel-to-pixel comparison. The slope $M^{\rm cl}_{\rm CNN}/M^{\rm cl}_{\rm truth} = 0.867^{+0.327}_{-0.296}$ is consistent with unity. On the other hand, we obtain $M^{\rm cl}_{\rm KS}/M^{\rm cl}_{\rm truth} = 0.484^{+0.218}_{-0.149}$ for the KS93 reconstruction, which is a $\gtrsim 2\sigma$ departure from unity. For the case of CNN, the data points at $M^{\rm cl}_{\rm truth} \gtrsim 40$ hint at the possibility that the estimated masses may be systematically lower. Although

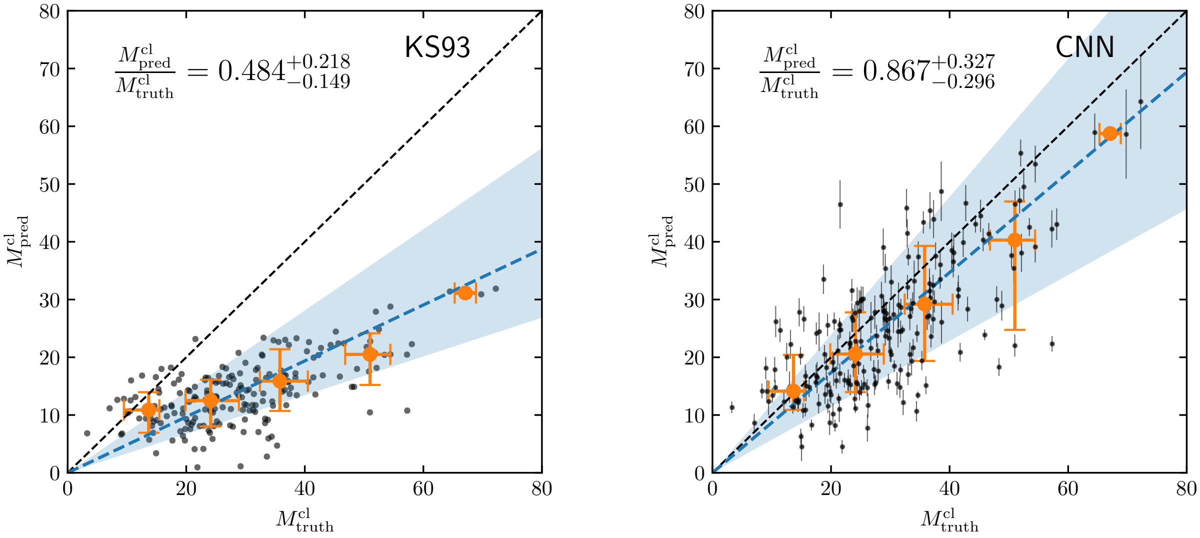

Figure 6. Comparison of projected cluster masses between prediction and truth. The projected cluster mass $M^{\rm cl}$ is defined to be the sum of convergence values within a $r = 1\rlap{.}'92$ circle aperture centered on the truth mass peak. Filled orange circles are the median values in the evenly distributed $M^{\rm cl}_{\rm truth}$ bins, and their error bars represent the 1σ certainties of $M^{\rm cl}_{\rm truth}$ ad $M^{\rm cl}_{\rm pred}$ within the bins. The errors on individual data points (filled black circles) in the right panel are the standard deviations of CNN results from 10 independent runs. Blue dashed lines and filled area are the median and 1σ confidence levels of the ratio $M^{\rm cl}_{\rm pred}/M^{\rm cl}_{\rm truth}$. The ratio of the CNN masses to the truth is consistent with unity ($0.867^{+0.327}_{-0.296}$) while the ratio is significantly lower ($0.484^{+0.218}_{-0.149}$) when the KS93 masses are used.

the sample size is small in this regime, we speculate that this may happen because the employed aperture radius ($r = 1\rlap{.}'92$) is not sufficiently large for these very massive clusters.

3.4. *Cluster Centroid*

Robust estimation of centroids is an important issue in cluster WL studies (e.g., von der Linden et al. 2014; Randall et al. 2008). The centroid serves as a reference to characterize the properties of the cluster. Also, in merging galaxy clusters, the position of the mass clump with respect to other cluster components is critical in our reconstruction of their merging scenarios. Here we compare the performance in centroid recovery between the KS93 and our CNN methods.

We measured the centroid in two steps. First, we located the pixel that has the largest convergence value. Then, we applied a 21 pixel×21 pixel ($4\rlap{.}'03\times4\rlap{.}'03$) square top-hat window and evaluated the first moments. Occasionally, negative convergence values are present within the window in the KS93 mass reconstruction. To prevent the centroid from leaving the window in this case, we rescaled the mass map in such a way that the minimum value within the window becomes zero. The application of the top-hat window is to include the contribution from the large-scale structure around the peak in our estimation of the centroid.

Figures 7 displays the deviations of the reconstructed mass centroids with respect to the truth. The CNN and KS93 results give similarly small ($1 \sim 3$ pixels) centroid deviations for massive clusters ($M^{\rm cl}_{\rm truth} \gtrsim 35$). Remarkably, we find striking differences in the low-mass ($M^{\rm cl}_{\rm truth} \lesssim 35$) regime. The CNN centroid deviations gradually increase for decreasing masses, reaching ~10 pixels at $M^{\rm cl}_{\rm truth} \sim 10$. On the other hand, the KS93 result shows many catastrophic errors ($\gtrsim 50$ pixels) in this regime. This contrast is seen more clearly in Figure 8, where we directly compare the deviations for the same clusters.

We attribute the large difference in the centroid deviations for low mass clusters to the uncontrolled noise fluctuation in the KS93 mass reconstruction discussed in §3.1. As shown by the example in Figure 2, sometimes the highest convergence values are found not within the cluster region. Also, even in the case where the highest convergence value is not catastrophically far from the truth, the lack of the contrast against the neighboring background substructures makes the centroid measurement highly uncertain.

3.5. *Influence of Masking*

Up to now, we have tested our our CNN method while assuming that no masked regions are present within the reconstruction field. In real observations, however, we need to mask out the regions affected by bright stars.

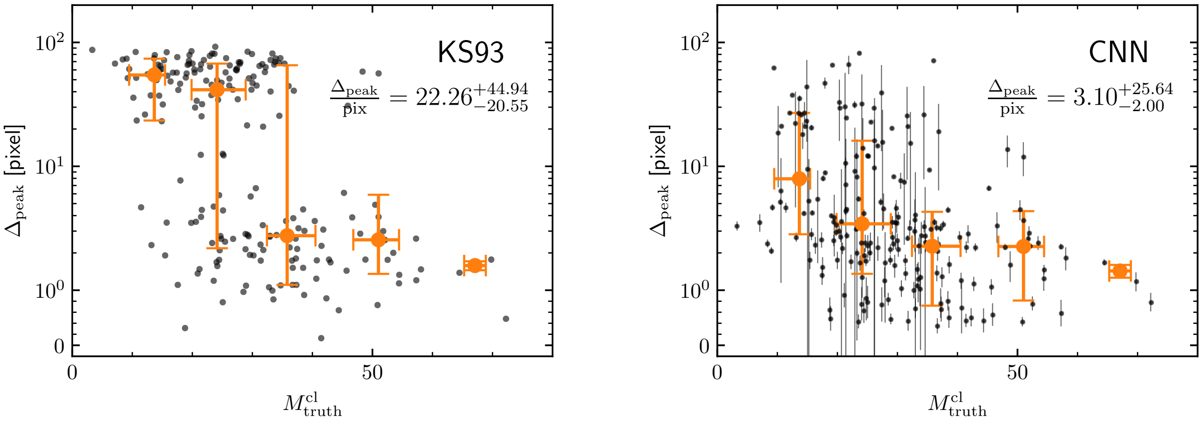

Figure 7. Cluster centroid deviation ($\Delta_{\rm peak}$) as a function of the truth cluster mass ($M^{\rm cl}_{\rm truth}$). Both CNN and KS93 perform well for massive clusters ($M^{\rm cl}_{\rm truth} \gtrsim 35$). However, the KS93 method produces many catastrophic errors for $M^{\rm cl}_{\rm truth} \lesssim 35$.

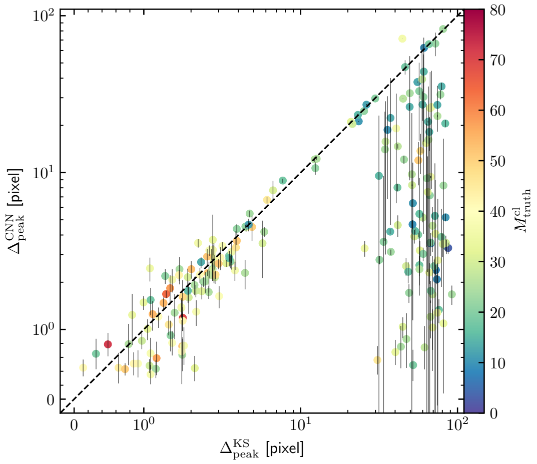

Figure 8. Comparison of the centroid errors between the KS93 and CNN results. The data points are color-coded with the truth mass. There present many catastrophic errors in the KS93 results.

Several methods have been suggested to minimize some artifacts due to the missing information (e.g., Starck et al. 2003; Pires et al. 2009). In this paper, without taking any explicit measure to minimize the influence, we simply investigate the impact of large maskings on our CNN mass reconstruction performance.

The expected number density of bright stars depends on the galactic latitude. And the exact size of the masking for a given magnitude star varies according to specific reduction/analysis methods. Reviewing our previous WL studies with Subaru/Suprime-Cam imaging data, we find that within the typical $20' \times 20'$ WL analysis area, $1 \sim 2$ bright-star maskings were needed with the masking radius ranging from ~$0.5'$ to ~$2'$ (e.g., Finner et al. 2017; Kim et al. 2019; Yoon et al. 2020). To mimic such conditions, we applied bright-star masking to our source catalogs with these number density and size distributions. We ensured that every cluster has at least one masking near the mass peak because we are interested in investigating the effect at its maximum.

Our visual inspection of the result shows that in most cases the CNN method can still detect the cluster mass clumps in the presence of masks. In Figure 9 we display one such example. Although the central mask completely removes source galaxies within the $r = 2'$ (~0.73 Mpc) circular mask placed near the mass peak, the reconstruction can still reveal the cluster nearly at the truth position. However, we find that because of the missing data the convergence values are slightly underestimated.

In order to examine the masking impact quantitatively, we measured the joint distribution, cluster mass comparison, and centroid distribution for the entire test sample as in §3.2, §3.3, and §3.4, respectively. The joint κ distribution displayed in the left panel of Figure 10 clearly indicates that the correlation with the truth in the high convergence regime is significantly weakened. Compared with the non-masking case (right panel of Figure 4), the slope is reduced by a factor of $1.5 \sim 2$ at $0.2 \lesssim \kappa_{\rm truth} \lesssim 0.5$, which is consistent with our expectation from the visual inspection of the convergence map.

Since this weakened correlation in κ is primarily due to the underestimation of the κ values within the mask placed near the cluster center, we can expect that the correlation in cluster mass also suffers in a similar fashion. The slope of the reconstructed mass to the truth becomes $0.554^{+0.257}_{-0.211}$ (see middle panel of Figure 10), which is substantially smaller than the non-masking case $0.867^{+0.327}_{-0.296}$.

Finally, in terms of the centroid deviation, we find that the fraction of the catastrophic errors increases because the convergence values within the masked region are

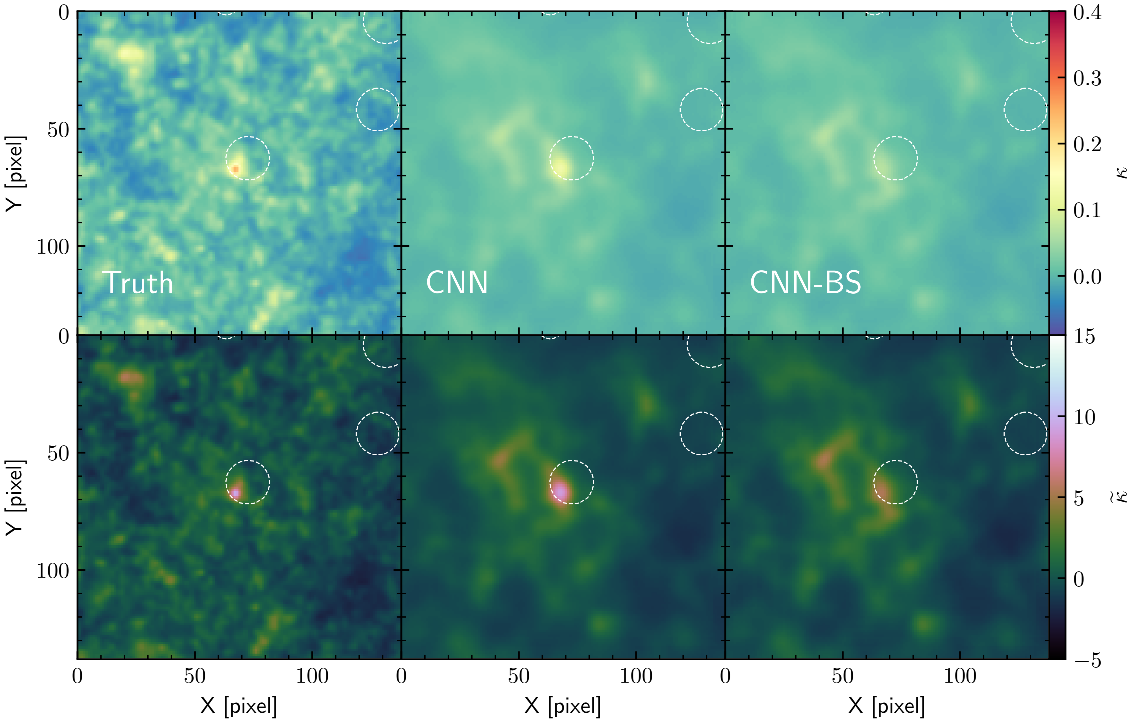

Figure 9. Impact of the bright-star masking on the CNN mass reconstruction. We use the same cluster presented in Figure 2 as an example. The middle (right) panel shows the reconstruction without (with) bright-star masking. The color scale of the bottom panel is based on the normalized convergence ($\widetilde{\kappa}$) as in Figure 2. White dashed circles mark the locations of the bright-star masks. Although the central mask crops out the most significant region of the cluster, the result (CNN-BS) shows that the cluster is still clearly detected near the truth position. We note that the missing information leads to slight underestimation of the peak convergence values.

underestimated and this makes the largest convergence peak within the reconstructed field sometimes lie outside the masked area. The right panel of Figure 10 displays the comparison of the centroid deviation with the KS reconstruction result *performed without any masking*. Even in the low deviation regime ($\Delta^{\rm KS}_{\rm peak} \lesssim 10$ pixels), the CNN method sometimes produces catastrophic errors. As mentioned earlier, this happens because of the in-mask underestimation. However, interestingly, in the regime where KS produces catastrophic errors ($\Delta^{\rm KS}_{\rm peak} \gtrsim 30$ pixels), the CNN performance is sometimes significantly better. This may happen for the cases where the in-mask underestimation is less severe than the KS93 artifacts including noise amplification and inadequate κ-scale recovery (see the discussion in §3.1).

4. DISCUSSION

4.1. *Why Does Our CNN algorithm Outperform the KS93 method?*

The comparison of our CNN mass reconstruction with the KS93 result has shown that the CNN performance is significantly better in several aspects (§3). To name a few, the bias in the projected cluster mass estimation based on the convergence map is greatly reduced. And the fraction of catastrophic errors in the centroid measurement becomes much smaller especially in the low mass regime. Moreover, the convergence map is adaptively smoothed in such a way that a larger kernel is used in regions where the lensing S/N is lower, which leads to effective noise suppression in the cluster outskirts. Here we present our discussion on the reason behind the outperformance.

The main cause for the improvement can be understood if we review some of the key issues in the conventional mass reconstruction (§2.2). The mass-sheet degeneracy is the most fundamental problem because the shear γ remains unchanged under the transformation of the convergence field: $\kappa \rightarrow \lambda\kappa + (1-\lambda)$. This degeneracy can be lifted only by imposing some specific κ value somewhere in the reconstruction field. One reasonable assumption is that the mean convergence is close to zero (although it should not be exactly zero) near the field boundary for a wide field mass reconstruction. This allows us to determine the λ value and thus break the degeneracy. Because our training data sets are drawn

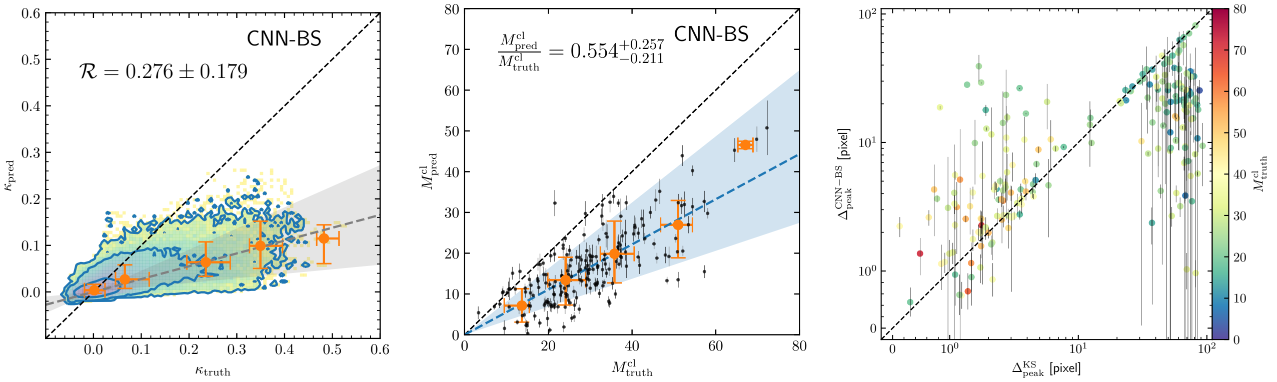

Figure 10. CNN performance under the influence of bright-star masking. Left panel: joint probability measured from the convergence pixels within the masks. Middle panel: comparison of projected cluster mass estimates with the truth. Right panel: comparison of centroid deviations with the KS93 ones performed without bright-star masking.

from cosmological simulation data, we believe that our CNN learns to utilize the information.

Another critical issue is the nonlinearity in the $g \rightarrow \kappa$ mapping. The fact that while the average ellipticity $\langle e \rangle$ provides a reduced shear $g = \gamma/(1-\kappa)$, the convergence is a function of a shear γ (equation 10) is ignored in the original KS93 formalism under the assumption that $g \sim \gamma$ (i.e., $\kappa \ll 1$) in the very weak gravitational lensing regime. Obviously, the condition $\kappa \ll 1$ is invalid in the typical cluster environment. Several suggestions are present in the literature to implement the nonlinearity. For example, Seitz & Schneider (1996) suggest an iterative procedure by updating γ in equation (10) with the information on κ in the previous step. One drawback in this approach might be noise amplification through the iteration. Therefore, some authors propose maximum likelihood-based methods with some regularization constraints (e.g., Seitz et al. 1998; Bradač et al. 2004; Jee et al. 2007). However, the fundamental limitation is that one needs κ on an absolute scale in order to correctly address the nonlinearity $g = \gamma/(1-\kappa)$. That is, the nonlinearity problem cannot be addressed in isolation. In CNN-based deep learning algorithms, these nonlinear issues are routinely addressed, and many applications turn out to be promising (see, e.g., Lucas et al. 2018; McCann et al. 2017; Rivenson et al. 2017, and references therein). In fact, the development of the CNN algorithm is motivated to address nonlinear problems such as denoising, image restoration, deconvolution, super-resolution, medical image reconstruction, holographic image reconstruction and so on. Therefore, it is not surprising to observe that combined with the mass-sheet degeneracy lifting capability, our CNN mass reconstruction significantly outperforms the original KS93 method.

4.2. *Test with Real Observations: Application to the El Gordo Cluster Data*

We have demonstrated that our CNN method can successfully reconstruct the projected mass maps from WL galaxy shears with the overall performance significantly better than that of the classical KS93 algorithm. Now one of the important questions is how well the current CNN method would work given real observational data, where a number of additional issues such as shear calibration errors, instrumental signatures, photometric redshift systematics, etc. are present. With further development in deep learning and astronomical image generation tools, these issues may become tractable through end-to-end WL simulations in the future. Here we apply our CNN method to the *HST* WL data for the high-redshift merging cluster "El Gordo" (Menanteau et al. 2014; Jee et al. 2014). Within the current scope, we are interested in investigating the performance of our CNN method given the difference between the training data set and the real data in the following three aspects. First, our training was performed with the specific $32' \times 32'$ field size whereas the *HST* field size (~$9' \times 9'$) of the El Gordo data is smaller by a few factors. Second, the training is done by assuming that every source galaxy has an identical source redshift. Obviously, the source population in the El Gordo field comes from a wide range of redshifts and more importantly contains a significant fraction of non-background (contamination from cluster members and foreground objects) galaxies. Third, the source density in the training data set is 25 per sq. arcmin, approximately a factor of four lower than the source density (~100 per sq. arcmin) in the *HST* observation of El Gordo.

Our HST catalog for El Gordo is provided by J. Kim *et al.* in prep., who studied the cluster with a new wide-field HST imaging data set (PROP ID: 14153, PI. Hughes). Readers are referred to J. Kim *et al.* in prep. for details in the observation setup and reduction methods. In brief, the cluster was observed in four different

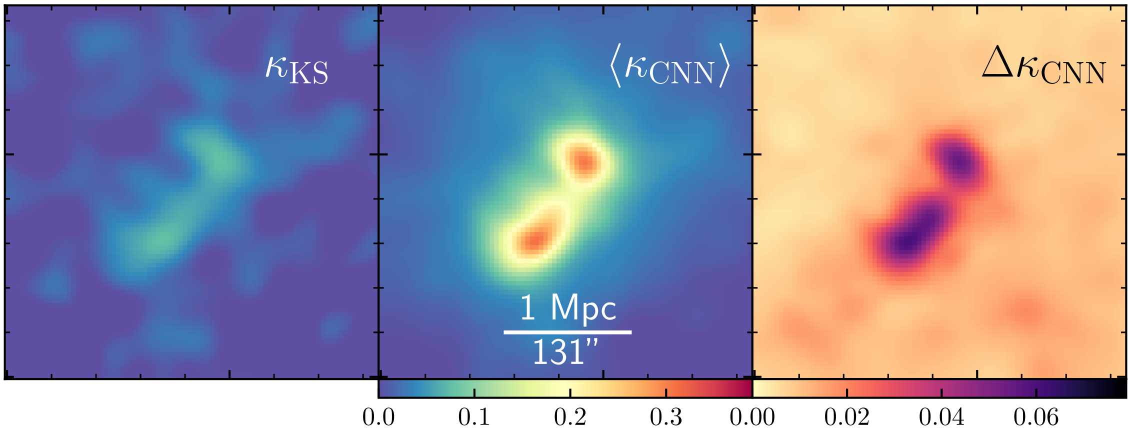

Figure 11. Application of our CNN mass reconstruction to the *El Gordo* cluster. We use the *HST* WL catalog of J. Kim *et al.* in prep. in this test. The left panel shows the mass reconstruction based on the KS93 algorithm. The CNN mass map in the middle panel is the average of the results from our 10 independent runs, which are also used to estimate the standard deviation shown in the right panel. Despite the differences between the training datasets and the El Gordo data, the CNN algorithm significantly outperforms the KS93 method in terms of the dynamical range representation, noise suppression, and substructure resolution.

programs (PROP IDs: 12477, 12755, 14096, and 14153). The entire field of view of the data with the addition of the last program (PROP ID: 14153) is ~119 sq. arcmin, which covers the cluster beyond the virial radius $r_{200} \sim 2$ Mpc (J. Kim *et al.* in prep.). With the combination of all existing programs, the resulting average source density is ~95 per sq. arcmin.

Figure 11 displays the reconstructed mass map of El Gordo cluster from our CNN model. The comparison with the KS93 version indicates that the advantages of the CNN method demonstrated in §3 with the simulated catalogs also manifest themselves here. First, the dynamical range of κ is more realistic in the CNN version. El Gordo is one of the extremely massive clusters in the universe, and the projected convergence value should be $\kappa \gtrsim 0.4$ in the central region based on the effective source redshift of ~1.2 (J. Kim *et al.* in prep.) and the redshift of the cluster 0.87. The range of the convergence value in the CNN mass map is consistent with this expectation, although the relatively small field size does not allow us to lift the mass-sheet degeneracy completely. On the other hand, the peak convergence value in the KS93 case is too small. Second, the KS93 inversion generates a number of spurious features in the outskirts whereas the CNN mass reconstruction efficiently suppresses these fluctuations. Currently, our multiwavelength data from X-ray to radio do not support the possibility that the features seen in the KS93 map might be real. Third, the two mass peaks are better resolved in the CNN mass reconstruction. In the KS93 mass map, although one can see the presence of the two mass maps, there exists a "bridge" connecting the two mass peaks. Again, the existence of such a connecting substructure is not supported by our data.

5. CONCLUSION

In this paper, we have introduced a new WL mass reconstruction method based on CNN algorithms. Our CNN architecture consists of a series of 2D convolution and transposed-convolution layers with implementation of skip-connections between the input and the output of the convolution-transposed-convolution layers via multiplication operation. We generate training data sets using the ray-tracing data from cosmological simulations while the statistical properties of the source galaxies are designed to match those in our typical WL studies with Subaru/Suprime-Cam images.

Compared with the original KS93 inversion, our CNN method produces significantly improved results. The merits include enhancement in restoration of the dynamical range, agreement with the truth in both pixel-by-pixel and cluster mass comparisons, centroid determination, and noise suppression. In particular, it is remarkable that the slope of the recovered mass to the truth becomes consistent ($0.867^{+0.327}_{-0.296}$) with unity for the test sample. The slope is much lower ($0.484^{+0.218}_{-0.149}$) when we use the KS93 results instead. Also, we find that the centroid estimation based on the CNN result is much more stable in the low-mass regime. We attribute these improvements to the efficient handling of the nonlinearity and degeneracy in our CNN algorithm, which however have been plaguing the traditional mass reconstruction methods.

The performance of our CNN algorithm somewhat degrades when a bright-star masking is fortuitously placed near the cluster center. Nevertheless, we find that the CNN reconstruction can still recover the cluster mass peak in most cases and the overall performance is still better than the KS93 results.

We tested our CNN model using the *HST* WL catalog of the El Gordo cluster. Despite the difference between our training data set and the real data in field size, source density, and redshift distribution, the CNN method clearly resolves the two mass clumps of the merging cluster in excellent agreement with the cluster member distribution while suppressing the noise fluctuation in the outskirts.

Our study is the first implementation of WL mass reconstruction with CNN methods. Although further refinements in both algorithm and simulation are needed before we use the method for quantitative characterization of galaxy clusters, the result from this pilot study looks promising. One immediate application without the further improvements is shear-based galaxy cluster detection. Among the various selection methods, the shear-based galaxy cluster selection is unique in its ability to detect galaxy clusters with their projected masses. However, one of the most outstanding obstacles is the control of false positives due to noise fluctuation. As demonstrated throughout the paper, our CNN method suppresses this noise fluctuation efficiently while preserving the resolution in the high-density region, where the shear signal is high. In addition, since the projected κ values are useful mass proxies, the CNN method can be used to provide the first classification of clusters according to their masses. Finally, the CNN-aided centroid determination and its comparison with other multiwavelength data can enhance our substructure identification and also reconstruction of merging scenarios in colliding clusters.

ACKNOWLEDGMENTS

The authors thank Inkyu Park, Cristiano Sabiu, David Parkinson, and Min-su Shin for helpful discussions. SEH, SP, and DB were (partly) supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2018R1A6A1A06024977). SEH was also partly supported by the project 우주거대구조를 이용한 암흑우주 연구 ("Understanding Dark Universe Using Large Scale Structure of the Universe"), funded by the Ministry of Science. MJJ acknowledges support from the National Research Foundation of Korea under the program nos. 2017R1A2B2004644 and 2017R1A4A1015178. DB was also supported by NRF Grant 2020R1A2B5B01001473. This work was also supported by the Korean Astronomy Machine Learning (KAML) working group.

This work is based on observations made with the NASA/ESA Hubble Space Telescope and operated by the Association of Universities for Research in Astronomy, Inc. under NASA contract NAS 5-2655. Computational data were transferred through a high-speed network provided by the Korea Research Environment Open NETwork (KREONET).

We thank the Columbia Lensing group (http://columbialensing.org) for making their simulations available. The creation of these simulations is supported through grants NSF AST-1210877, NSF AST-140041, and NASA ATP-80NSSC18K1093. We thank New Mexico State University (USA) and Instituto de Astrofisica de Andalucia CSIC (Spain) for hosting the Skies & Universes site for cosmological simulation products.

Software: Astropy (Astropy Collaboration et al. 2013), Keras (Chollet et al. 2015), Matplotlib (Hunter 2007), NumPy/SciPy (Virtanen et al. 2020), Pandas (McKinney 2010), SExtractor (Bertin & Arnouts 1996), Tensorflow (Abadi et al. 2015)

Facility: HST (ACT)

APPENDIX

A. PERFORMANCE TEST WITH OTHER CNN ARCHITECTURES

Table 3. Same as Table 2, including various CNN architectures. The CNN model used in the main text is FocalLoss, and see text for the definition of each architecture. Those with the best performance for each parameter is marked as bold characters, while those with the worst performance for each parameter is underlined. KS93 and NoSkip are not marked as underlined because they always show poorer performances than the other architectures.

Model	$\mathcal{R}(\kappa_{\rm pred}, \kappa_{\rm truth})$	$\mathcal{D}(\widetilde{\kappa}_{\rm pred}, \widetilde{\kappa}_{\rm truth})$	$M^{\rm cl}_{\rm pred}/M^{\rm cl}_{\rm truth}$	$\Delta_{\rm peak}$ [pixel]
KS93	0.253 ± 0.131	6.26 ± 4.57	$0.484^{+0.218}_{-0.149}$	$22.26^{+44.94}_{-20.55}$
Fiducial	0.363 ± 0.267	4.59 ± 4.12	$0.762^{+0.305}_{-0.266}$	$\mathbf{2.88^{+21.82}_{-1.90}}$
4Channel	0.360 ± 0.272	$\underline{4.62 \pm 4.50}$	$0.789^{+0.289}_{-0.263}$	$3.05^{+19.08}_{-1.92}$
19Filter	$\underline{0.329 \pm 0.261}$	$\mathbf{4.09 \pm 4.32}$	$\underline{0.720^{+0.314}_{-0.238}}$	$\underline{3.81^{+6.22}_{-2.07}}$
29Filter	0.359 ± 0.278	4.32 ± 4.16	$0.754^{+0.305}_{-0.245}$	$3.49^{+38.08}_{-2.31}$
FocalLoss	$\mathbf{0.407 \pm 0.288}$	4.36 ± 3.77	$\mathbf{0.867^{+0.327}_{-0.296}}$	$3.10^{+25.64}_{-2.00}$
NoSkip	-0.361 ± 0.221	5.04 ± 3.79	$-1.329^{+0.457}_{-0.842}$	$3.81^{+6.22}_{-2.07}$

We present performance tests of various CNN architectures using the diagnostics given in §3. Hereafter the architecture presented in the main text of the current paper is referred to FocalLoss. It uses the modified mean-square-error loss function inspired by the focal loss (Lin et al. 2017, see equation (11)).

In addition to FocalLoss, we test the following five variations:

Fiducial: same as FocalLoss except that the loss function is given by the standard mean square error (MSE):

$$\mathcal{L} = \sum_{\mathbf{x}} \left[\kappa_{\rm pred}(\mathbf{x}) - \kappa_{\rm truth}(\mathbf{x})\right]^2 \; . \qquad \text{(A1)}$$

This MSE is also used for the rest of the variations.

4Channel: same as Fiducial except that the smoothed number distribution of background galaxies is used as an additional channel of the input layer.

19Filter: same as Fiducial except that the employed filter size is 19 × 19 (instead of 49 × 49) during the convolution and transposed convolution operations.

29Filter: same as Fiducial except that the employed filter size is 29 × 29.

NoSkip: same as Fiducial except that it uses no skip-connection.

Figure 12 compares the mass reconstructions from these various CNN architectures. Judged by visual inspection, most CNN variations produce similar results. The exception is NoSkip, whose resolution is substantially compromised compared to the others. When we examined the results from the individual runs, the NoSkip runs frequently produce null results, where the convergence map is flat ($\kappa(\mathbf{x}) \approx$ constant). Also, the non-null results lack small-scale structures. This comparison illustrates that the ResNet-like skip-connection between convolution and transposed-convolution layers plays a crucial role for recovering the details. The 4Channel result does not show any significant merit over Fiducial. This indicates that the additional information on the source number distribution does not meaningfully contribute to the mass reconstruction quality. In this example and also others, we find that the 19Filter results tend to slightly overestimate the densities near the field edges compared to those produced with the larger (29×29 or 49×49) filters, although the difference becomes insignificant when we compare the 29×29 vs 49×49 cases. This implies that there may exist a lower threshold in filter size in order to properly restore the dynamic range. Tables 3 and 4 summarize the comparisons among different CNN branches.

Table 4. Same to Table 3, but for bright-star masking (see §3.5). NoSkip is not shown because of its poor performance.

Model	$\mathcal{R}(\kappa_{\rm pred}, \kappa_{\rm truth})$	$\mathcal{D}(\widetilde{\kappa}_{\rm pred}, \widetilde{\kappa}_{\rm truth})$	$M^{\rm cl}_{\rm pred}/M^{\rm cl}_{\rm truth}$	$\Delta_{\rm peak}$ [pixel]
Fiducial-BS	0.216 ± 0.140	4.22 ± 3.50	$0.448^{+0.217}_{-0.182}$	$13.56^{+21.28}_{-11.05}$
4Channel-BS	$\underline{0.180 \pm 0.117}$	$\underline{4.93 \pm 3.93}$	$0.400^{+0.178}_{-0.135}$	$21.85^{+23.34}_{-19.12}$
19Filter-BS	0.193 ± 0.142	4.91 ± 3.90	$\underline{0.386^{+0.191}_{-0.149}}$	$\underline{31.51^{+26.00}_{-28.56}}$
29Filter-BS	0.209 ± 0.144	4.29 ± 3.55	$0.421^{+0.188}_{-0.193}$	$13.30^{+31.72}_{-10.97}$
FocalLoss-BS	$\mathbf{0.276 \pm 0.179}$	$\mathbf{3.66 \pm 3.22}$	$\mathbf{0.554^{+0.257}_{-0.211}}$	$\mathbf{7.95^{+21.94}_{-5.94}}$

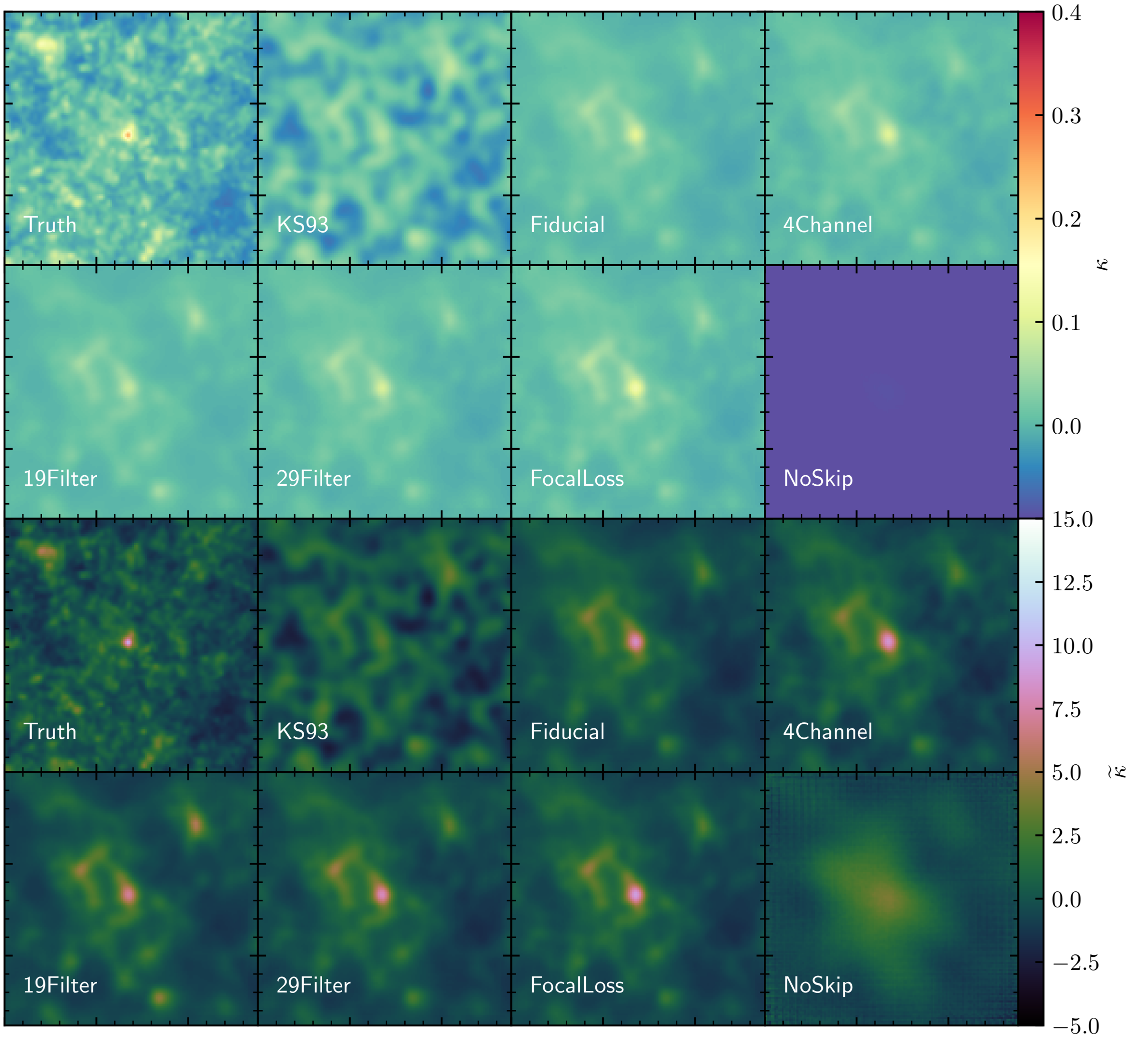

Figure 12. Comparison of mass reconstruction results from different CNN architectures. See text for details of each variation. Most CNN variations show similar performances except for NoSkip, which suffers from significant resolution loss. According to our quantitative comparison based on the entire test sample, FocalLoss (used in the main text) produces the best results.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, software available from tensorflow.org

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33

Bartelmann, M. 1995, A&A, 303, 643

Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393

Bradač, M., Lombardi, M., & Schneider, P. 2004, A&A, 424, 13

Chollet, F., et al. 2015, Keras, https://keras.io

Clerkin, L., Kirk, D., Manera, M., et al. 2017, MNRAS, 466, 1444

Finner, K., Jee, M. J., Golovich, N., et al. 2017, ApJ, 851, 46

Fischer, P., & Tyson, J. A. 1997, AJ, 114, 14

Flamary, R. 2017, in 2017 25th European Signal Processing Conference (EUSIPCO), IEEE, 2468–2472

Gorenstein, M. V., Falco, E. E., & Shapiro, I. I. 1988, ApJ, 327, 693

He, K., Zhang, X., Ren, S., & Sun, J. 2015, arXiv e-prints, arXiv:1512.03385

High, F. W., Rhodes, J., Massey, R., & Ellis, R. 2007, PASP, 119, 1295

Hikage, C., Oguri, M., Hamana, T., et al. 2019, PASJ, 71, 43

Hilbert, S., Hartlap, J., & Schneider, P. 2011, A&A, 536, A85

Hunter, J. D. 2007, Computing in Science & Engineering, 9, 90

Ioffe, S., & Szegedy, C. 2015, arXiv e-prints, arXiv:1502.03167

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111

Jain, B., Seljak, U., & White, S. 2000, ApJ, 530, 547

Jarvis, M., Schechter, P., & Jain, B. 2008, arXiv e-prints, arXiv:0810.0027

Jee, M. J., Hughes, J. P., Menanteau, F., et al. 2014, ApJ, 785, 20

Jee, M. J., Ford, H. C., Illingworth, G. D., et al. 2007, ApJ, 661, 728

Kaiser, N., & Squires, G. 1993, ApJ, 404, 441

Kim, E. J., & Brunner, R. J. 2017, MNRAS, 464, 4463

Kim, M., Jee, M. J., Finner, K., et al. 2019, ApJ, 874, 143

Kingma, D. P., & Ba, J. 2014, arXiv e-prints, arXiv:1412.6980

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv e-prints, arXiv:1110.3193

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. 2017, arXiv e-prints, arXiv:1708.02002

Liu, J., Bird, S., Zorrilla Matilla, J. M., et al. 2018, JCAP, 2018, 049

Lucas, A., Iliadis, M., Molina, R., & Katsaggelos, A. K. 2018, IEEE Signal Processing Magazine, 35, 20

Mandelbaum, R., Rowe, B., Armstrong, R., et al. 2015, MNRAS, 450, 2963

McCann, M. T., Jin, K. H., & Unser, M. 2017, IEEE Signal Processing Magazine, 34, 85

McKinney, W. 2010, in Proceedings of the 9th Python in Science Conference, ed. S. van der Walt & J. Millman, 51 – 56

Menanteau, F., Hughes, J. P., Barrientos, F., & Infante, L. 2014, Is “El Gordo” the fattest cluster in the Universe?, NOAO Proposal

Meyers, J. E., & Burchat, P. R. 2015, ApJ, 807, 182

Mittal, A., Soorya, A., Nagrath, P., & Hemanth, D. J. 2019, Earth Science Informatics, 1

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563

Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, A&A, 621, A26

Pires, S., Starck, J. L., Amara, A., et al. 2009, MNRAS, 395, 1265

Randall, S. W., Markevitch, M., Clowe, D., Gonzalez, A. H., & Bradač, M. 2008, ApJ, 679, 1173

Rivenson, Y., Zhang, Y., Günaydın, H., Teng, D., & Ozcan, A. 2017, Light: Science & Applications, 7, 17141

Schaefer, C., Geiger, M., Kuntzer, T., & Kneib, J. P. 2018, A&A, 611, A2

Seitz, S., & Schneider, P. 1996, A&A, 305, 383

Seitz, S., Schneider, P., & Bartelmann, M. 1998, A&A, 337, 325

Spergel, D., Gehrels, N., Baltay, C., et al. 2015, arXiv e-prints, arXiv:1503.03757

Squires, G., & Kaiser, N. 1996, ApJ, 473, 65

Starck, J. L., Donoho, D. L., & Candès, E. J. 2003, A&A, 398, 785

Troxel, M. A., MacCrann, N., Zuntz, J., et al. 2018, PhRvD, 98, 043528

van Waerbeke, L. 2000, MNRAS, 313, 524

Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, Nature Methods, 17, 261

von der Linden, A., Allen, M. T., Applegate, D. E., et al. 2014, MNRAS, 439, 2

Yoon, M., Lee, W., Jee, M. J., et al. 2020, ApJ, 903, 151