Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **528,** 5212–5230 (2024) https://doi.org/10.1093/mnras/stae257
Advance Access publication 2024 January 27

Deep learning approach for identification of H II regions during reionization in 21-cm observations – II. Foreground contamination

Michele Bianco [1,2★] Sambit. K. Giri [3,4] David Prelogović,[5] Tianyue Chen [1] Florent G. Mertens [6] Emma Tolley,[1] Andrei Mesinger [5] and Jean-Paul Kneib[1]

[1]*Laboratoire d'Astrophysique, Ecole Polytechnique Federale de Lausanne (EPFL), Observatoire de Sauverny, Versoix CH-1290, Switzerland*
[2]*Astronomy Centre, Department of Physics & Astronomy, Pevensey III Building, University of Sussex, Falmer, Brighton BN1 9QH, UK*
[3]*Nordita, KTH Royal Institute of Technology and Stockholm University, Hannes Alfv´ens väg 12, SE-106 91 Stockholm, Sweden*
[4]*Institute for Computational Science, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland*
[5]*Scuola Normale Superiore, Piazza dei Cavalieri 7, I-56126 Pisa, Italy*
[6]*LERMA, Observatoire de Paris, PSL Research University, CNRS, Sorbonne Université, F-75014 Paris, France*

Accepted 2024 January 18. Received 2023 December 11; in original form 2023 April 6

ABSTRACT

The upcoming Square Kilometre Array Observatory will produce images of neutral hydrogen distribution during the epoch of reionization by observing the corresponding 21-cm signal. However, the 21-cm signal will be subject to instrumental limitations such as noise and galactic foreground contamination that pose a challenge for accurate detection. In this study, we present the `SegU-Net v2` framework, an enhanced version of our convolutional neural network, built to identify neutral and ionized regions in the 21-cm signal contaminated with foreground emission. We trained our neural network on 21-cm image data processed by a foreground removal method based on Principal Component Analysis achieving an average classification accuracy of 71 per cent between redshift $z = 7$ and 11. We tested `SegU-Net v2` against various foreground removal methods, including Gaussian Process Regression, Polynomial Fitting, and Foreground-Wedge Removal. Results show comparable performance, highlighting `SegU-Net v2`'s independence on these pre-processing methods. Statistical analysis shows that a perfect classification score with AUC = 95 per cent is possible for $8 < z < 10$. While the network prediction lacks the ability to correctly identify ionized regions at higher redshift and differentiate well the few remaining neutral regions at lower redshift due to low contrast between 21-cm signal, noise, and foreground residual in images. Moreover, as the photon sources driving reionization are expected to be located inside ionized regions, we show that `SegU-Net v2` can be used to correctly identify and measure the volume of isolated bubbles with $V_{\rm ion} > (10\,{\rm cMpc})^3$ at $z > 9$, for follow-up studies with infrared/optical telescopes to detect these sources.

Key words: techniques: image processing – techniques: interferometric – dark ages, reionization, first stars – early Universe.

1 INTRODUCTION

Radiation emitted by the first luminous sources drastically influenced the chemical composition and thermal history of the intergalactic medium (IGM), transitioning the Universe from an initial cold and neutral state to a final hot and ionized state (e.g. Furlanetto, Oh & Briggs 2006; Ferrara & Pandolfi 2014; Choudhury 2022). These sources most likely formed at locations where dark matter structures collapsed into gravitational bound structures during redshift $z \gtrsim 10$ (Abel, Bryan & Norman 2001; Bromm et al. 2009; Pawlik, Milosavljević & Bromm 2011). The newly launched *JWST*[1] is already providing preliminary results by detecting possible ionizing source candidates at these high redshifts (Bakx et al. 2023; Castellano et al. 2022; Naidu et al. 2022), which will help us understand the conditions for early galaxy formation (e.g. Boylan-Kolchin 2023; Dayal & Giri 2023; Hütsi et al. 2023).

Another way to probe the appearance of these first luminous sources is to observe the evolution of neutral hydrogen (H I) in the IGM. The ground state spin-flip transition of neutral hydrogen produces a signal with a wavelength of 21 cm in the rest frame, known as the 21-cm signal. The presence of this signal is directly correlated with the number density of neutral hydrogen present in the early Universe, and with the Universe expansion, the 21-cm signal wavelength redshifts into the radio frequency. As the first stars and galaxies formed and began emitting ultraviolet radiation, they started to ionize neutral gas in their surrounding. These primordial sources produce enough photons to escape their hosting environment and propagate into the IGM. As the hydrogen in the IGM becomes ionized, the intensity of the 21-cm signal decreases. Therefore, by observing the 21-cm signal from the early Universe with radio telescopes, we can study the reionization process and learn about the properties of the first luminous sources (e.g. Madau, Meiksin & Rees 1997; Furlanetto et al. 2006). Several radio experiments, such as the

★ E-mail: mbianco@protonmail.com
[1] http://jwst.nasa.gov

Low-frequency Array[2] (LOFAR; e.g. Ghara et al. 2020; Mertens et al. 2020), Murchison Wide-field Array[3] (MWA; e.g. Trott et al. 2020; Ghara et al. 2021), and Hydrogen Epoch of Reionization Array[4] (HERA; e.g. The HERA Collaboration 2022a, b), have been trying to detect this signal during the epoch of reionization (EoR).

Currently, the low-frequency band component of the Square Kilometre Array[5] (SKA-Low; e.g. Mellema et al. 2013), which will observe the sky at a frequency range between 50 and 350 MHz, is under construction. SKA-Low will have a field of view covering $\sim(10\ \mathrm{deg})^2$ on the sky (Koopmans et al. 2015). This radio interferometer will be sensitive enough to capture the evolution of the IGM during EoR with images of the 21-cm signal from redshift $z = 30$ to 5. This sequence of 21-cm maps observed at different frequencies will be stuck together to constitute a three-dimensional (3D) set of data, known as the multifrequency tomographic data set (e.g. Mellema et al. 2015; Wyithe, Geil & Kim 2015; Giri et al. 2018a). The 21-cm signal image data produced by the SKA-Low will contain imprints of the ionized regions (or bubbles) caused by the luminous sources (Giri et al. 2018a; Giri, Mellema & Ghara 2018b) and neutral regions (or islands) tracing the cosmic voids (Giri et al. 2019). By detecting these bubbles, we can learn about the locations of the first luminous sources (Zackrisson et al. 2020). We can also understand the nature and distribution of the photon sources driving the reionization process by studying the evolution of their sizes and morphology (e.g. Giri et al. 2018a, 2019; Kapahtia, Chingangbam & Appleby 2019; Gazagnes, Koopmans & Wilkinson 2021; Giri & Mellema 2021; Kapahtia et al. 2021; Elbers & van de Weygaert 2023). However, detecting these ionized bubbles in radio telescope observations is not trivial due to several limitations of the telescope, such as the limited resolution and instrument noise.

To detect these bubbles, previous authors have developed methods using visibilities data smoothed with appropriated filters to represent the sizes and shapes of the bubbles, then a likelihood for Bayesian approach estimates the parameters of the ionized regions filtered (e.g. Datta, Bharadwaj & Choudhury 2007; Ghara & Choudhury 2020). Other authors employ the image data of radio telescopes. This approach can be intensity-based, where the method filters the image based on a threshold value or region-based, by agglomerate clustering correlated pixels into groups with common traits within the image (e.g. Achanta et al. 2012; Mehra & Neeru 2016; Giri et al. 2018b). This task is a well-known assignment in artificial intelligence (AI) called segmentation. Therefore, another approach would be to consider a deep learning application. Recent work by Gagnon-Hartman et al. (2021) demonstrated that a combination of foreground avoidance and machine learning techniques enable 21-cm segmentation and bubble detection for experiments that are not necessarily optimized for imaging. Moreover, recently, we presented our first work (see Bianco et al. 2021, hereafter Paper I), where we introduced a deep learning approach to identify the distribution of H I regions in SKA 21-cm tomographic image using a U-shaped convolutional neural network (U-Net) (Ronneberger, Fischer & Brox 2015). We named our framework `SegU-Net` and we assessed how this network could process 21-cm images during the EoR contaminated by systematic noise simulated for SKA-Low and segment the images into ionized and neutral regions with an average of 87 per cent accuracy for redshift between 7 and 9. Moreover, we assessed that our network outperforms the `Super-Pixel` method (Giri et al. 2018a), considered the state-of-the-art algorithm for EoR segmentation, with, on average, 10–20 per cent more accuracy. We also demonstrated that `SegU-Net` could be used to recover the bubble size distributions with a relative difference within the 5 per cent and other summary statistics with the same level of accuracy. Moreover, we provided our method with a per-pixel uncertainty map that provides a confidence interval for its prediction and the derived statistics. We have tested the response of our framework to different noise levels based on a shorter (250 h) and more extended (1500 h) observing time, corresponding to an under- and overestimation of the noise level, respectively. We demonstrated that `SegU-Net` tolerates noise up to $\sqrt{2}$ times larger than the one employed in the training process, obtaining the same level of accuracy. By studying the uncertainty map and the response to the noise level, we realized that machine learning models are sensitive to the dynamic range and the intrinsic resolution of the simulated images.

While our previous work demonstrated excellent performance in detecting H I regions from EoR images, it should be considered a proof-of-concept as we consider EoR images with only telescope systematic noise, and we did not include any foreground contamination. The biggest challenge for the SKA-Low observation, just like other radio telescopes, is to separate the 21-cm signal from the undesired extra-galactic and galactic foreground contamination, which outshine the cosmological signal by several orders of magnitude (Jelić et al. 2008; Bowman, Morales & Hewitt 2009). The key goal of this work is to develop tools which remove these foregrounds while recovering the regions of H I during EoR from the 21-cm signal image data.

In this work, we will further develop our deep learning-based method to determine the ionized bubbles in image data with the presence of realistic galactic and extra-galactic foregrounds expected from the SKA-Low. Therefore, here we present `SegU-Net v2`, which extends the previous work by including foreground emissions of galactic origin and a complete study of its dependency on the foreground mitigation pre-processing step that partially subtracts the foreground signal, thus reducing the dynamic range in the 21-cm images before starting the network training. In the last three decades, several foreground removal methods with different approaches have been developed. Some of the early attempts take advantage of the spectral smoothness of the galactic and extra-galactic contaminants to fit along the line of sight and remove the foreground in either real or *uv* space (e.g.: Wang et al. 2006; Morales et al. 2006a; Morales, Bowman & Hewitt 2006b; Gleser, Nusser & Benson 2008; Liu et al. 2009b; Wang et al. 2013). However, more recent approaches suggest a non-parametric subtraction (e.g. Harker et al. 2009; Chapman et al. 2012, 2013; Gu et al. 2013; Bonaldi & Brown 2015; Mertens, Ghosh & Koopmans 2018) as the frequency smoothness of the foreground spectrum can be corrupted by beam effect and incomplete *uv* coverage (Liu, Tegmark & Zaldarriaga 2009a). Therefore, we perform a complete study of different available approaches for foreground subtraction in the case of the SKA-Low 21-cm tomographic data set applied to `SegU-Net v2`. We analyse the effect of the subtraction process on the predicted binary maps so that we can establish if a particular foreground removal method provides a concrete advantage for our task.

This paper is organized as follows. In Section 2, we present how we generate the simulated data sets used for this work, including details of our foreground model in Section 2.3 and a description of the mock 21-cm observation in Section 2.4. In Section 4, we describe the design and the training of our neural network. In Section 5, we discuss its application to our simulated SKA-Low data sets contaminated by the foreground signal, and we analyse summary

[2]https://www.astron.nl/telescopes/lofar
[3]https://www.mwatelescope.org
[4]https://reionization.org/
[5]https://skatelescope.org

statistics such as the mean ionization fraction, power spectra and topological quantities. In Section 5.2, we test our framework on a different foreground removal method. We discuss and summarize our conclusions in Section 6. Throughout this work, we assume a flat ΛCDM cosmology with the following parameters: $\Omega_\Lambda = 0.73$, $\Omega_m = 0.27$, $\Omega_b = 0.046$, $H_0 = 70\,\mathrm{km\,s^{-1}Mpc^{-1}}$, $\sigma_8 = 0.82$, $n_s = 0.96$. These values are based on the Wilkinson Microwave Anisotropy Probe (WMAP) 5 yr observation (Komatsu et al. 2009) and consistent with *Planck 2018* (Planck Collaboration VI 2020) results.

2 21-CM SIGNAL

This section illustrates the process we follow to create 21-cm mock observations of the EoR. Development of the network requires mock 21-cm observations of the EoR for network training, validation and testing, which will be described in Section 4.

2.1 Simulating the cosmological 21-cm signal during EoR

The intensity of the redshifted 21-cm signal emerging from a neutral cloud of hydrogen can be observed by a radio interferometric telescope as the difference against the CMB temperature $T_{\rm CMB}$, i.e. $\delta T_b \equiv T_b - T_{\rm CMB}$. For a given sky angular position $\hat{\boldsymbol{n}}$ and redshift z, we can define it to be (e.g. Zaroubi 2012; Mellema et al. 2013)

$$\delta T_{\rm b}(\boldsymbol{r}, z) = T_0(z)\left(1 - \frac{T_{\rm CMB}(z)}{T_{\rm S}(\boldsymbol{r}, z)}\right)[1 + \delta_b(\boldsymbol{r}, z)]x_{\rm HI}(\boldsymbol{r}, z), \quad (1)$$

$$T_0(z) \approx 27\ \mathrm{mK}\left(\frac{\Omega_{\rm b}}{0.044}\right)\left(\frac{\rm h}{0.7}\right)\sqrt{\left(\frac{1+\rm z}{10}\right)\left(\frac{0.27}{\Omega_{\rm m}}\right)}. \quad (2)$$

where $x_{\rm H\,I}$ is the neutral hydrogen fraction, δ_b is the baryonic overdensity, and T_S is the spin temperature. We assume that the IGM is heated well above the CMB temperature ($T_{\rm S} \gg T_{\rm CMB}$) at $z \lesssim 12$, which is consistent with theoretical predictions (e.g. Pritchard & Furlanetto 2007; Ross et al. 2017, 2019, 2021).[6] In this context, equation (1) is always positive and can be approximated as $\delta T_b \propto (1 + \delta_b)\, x_{\rm HI}$, while the presence of ionized regions is characterized by a lack of signal, $\delta T_b = 0\,\mathrm{mK}$. The radio interferometer cannot observe the absolute δT_b. Therefore, the ionized regions cannot be identified by finding pixels with zero signal in the 21-cm image data. To model the large-scale cosmological 21-cm signal expected during reionization, we employ the PYTHON wrapper of the `21cmFAST` seminumerical code (Mesinger, Furlanetto & Cen 2011; Murray et al. 2020). The code models the dark matter density evolution and gravitational collapse using the second-order Lagrangian perturbation theory. From the generated large-scale density field, a region is considered collapsed when it exceeds a defined mass threshold, which can be related to a minimum virial temperature $T_{\rm vir}^{\rm min}$. The excursion set formalism is then employed to calculate ionized regions (Furlanetto, Zaldarriaga & Hernquist 2004). The code outputs a coeval cube at different redshifts that are then used for constructing 21-cm lightcones. We refer the readers to e.g. Giri et al. (2018a) for more general details on the construction of lightcone from coeval cube simulations. In this work, we simulate the signal in coeval cubes for a total of ∼20 snapshot for redshift $z = [7, 11]$ with a mesh grid of 128^3 that is 256 Mpc along each direction.

[6]Note that the current 21-cm signal measurements have not completely ruled out the possibility of cold reionization (see e.g. Ghara et al. 2020, 2021; The HERA Collaboration 2022a). The signal becomes very complicated if $T_{\rm S} \sim T_{\rm CMB}$ when reionization begins (Ross et al. 2021; Schneider, Schaeffer & Giri 2023). Therefore, we defer a detailed exploration to the future.

Table 1. The telescope parameters used in this work. For the frequency channel width, we indicate the quantity at $z = 7$ and 11.

Parameters	Values	
System temperature	$T_{\rm sys}$	$60(\frac{\nu}{300\mathrm{MHz}})^{-2.55}$ K
Effective collecting area	$A_{\rm eff}$	962 m^2
Declination	θ_c	-30°
Frequency channel width	$\Delta\nu$	118 − 96 kHz
Observation hour per day	$t_{\rm daily}$	6 h
Signal integration time	$t_{\rm int}$	10 s

2.2 Systematic noise

We model the SKA-Low antenna receiver noise by a random Gaussian distribution with mean value zero and variance (Ghara et al. 2017; Giri et al. 2018b)

$$\sigma_{\rm uv} = \frac{k_{\rm B}\, T_{\rm sys}}{A_{\rm eff}}\sqrt{\frac{2\, t_{\rm daily}}{\Delta\nu\, N_{\rm uv}\, t_{\rm obs}\, t_{\rm int}}}. \quad (3)$$

Here, $t_{\rm int}$ is the integration time, $t_{\rm daily}$ is the window of observation per day, $T_{\rm sys}$ is the system temperature, $A_{\rm eff}$ is the effective collecting area, $\Delta\nu$ is the bandwidth, $N_{\rm uv}$ is the number of measurements that are detected in each cell of the uv-coverage grid. We assume an observation length of $t_{\rm obs} = 1000\,\mathrm{h}$. We list the SKA-Low telescope parameters in Table 1. The uv-coverage grid is simulated assuming the current plan for antennae distribution of SKA-Low.[7] In the top right-hand panel of Fig. 1, we show an example slice of the 21-cm signal and a noise realization at $z = 8.24$. As the map is degraded to a resolution corresponding to a maximum baseline of $B = 2\,\mathrm{km}$, we can see the large-scale distribution of the neutral and ionized regions.

2.3 Foreground contamination

Between 250 and 30 MHz, the dominant emission comes from the Galactic synchrotron radiation. This emission alone is expected to contribute to the majority of the total foreground contamination of the comic 21-cm signal (Di Matteo et al. 2002; Di Matteo, Ciardi & Miniati 2004; Santos, Cooray & Knox 2005; Datta et al. 2007; Jelić et al. 2008; Kerrigan et al. 2018). Other contributors can include emissions from unresolved extra-galactic point sources, Galactic free–free emissions, supernova remnants and extra-galactic radio clusters, which, for simplicity, have been neglected in this study. We based our Galactic synchrotron emission model on the Choudhuri et al. (2014) study. We express the foreground radiation with a Gaussian random field with an angular power spectrum as

$$\mathcal{C}_l^{\rm syn}(\nu) = A_{150}\left(\frac{1000}{l}\right)^{\overline{\beta}}\left(\frac{\nu}{\nu_\star}\right)^{-2\overline{\alpha}_{\rm syn} - 2\Delta\overline{\alpha}_{\rm syn}\log\left(\frac{\nu}{\nu_\star}\right)}. \quad (4)$$

Here, the parameter for the Galactic synchrotron emission is the power spectra amplitude $A_{150} = 512\,\mathrm{mK}^2$ at the reference frequency $\nu_\star = 150\,\mathrm{MHz}$, the angular scaling $\overline{\beta} = 2.34$, the spectra index $\overline{\alpha}_{\rm syn} = 2.8$ and the running spectral index $\Delta\overline{\alpha}_{syn} = 0.1$. These quantities are taken from Platania et al. (1998) and Wang et al. (2006). We then generate the foreground temperature fluctuations

[7]The SKA-Low design is given at https://www.skao.int/sites/default/files/documents/d18-SKA-TEL-SKO-0000422_02_SKA1_LowConfigurationCoordinates-1.pdf.

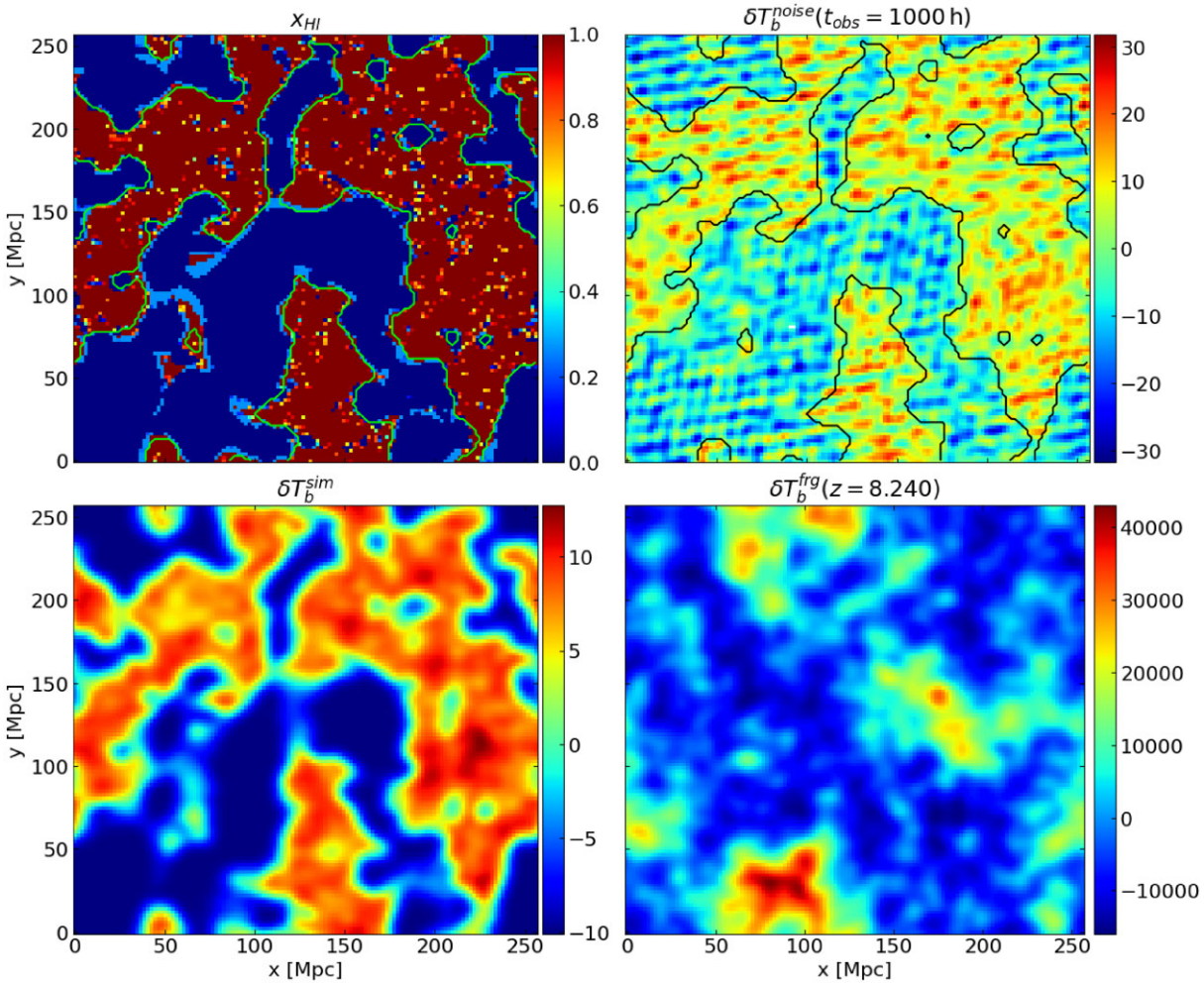

Figure 1. An example of a slice through the sky-plane used during the network training. *Top Left*: The neutral hydrogen fraction at simulation resolution when the reionization process is halfway complete. *Bottom Left*: The simulated 21-cm signal after the interferometric smoothing with a maximum baseline of $B = 2$ km and matching frequency resolution. We then subtract the frequency mean signal to mimic the effect of the lack of a zero baseline. *Top Right*: Systematic noise added to the 21-cm signal for an observing time of 1000 h. A solid black line indicates the neutral field after the same interferometric smoothing scale. *Bottom right*: The Galactic synchrotron emission added to the 21-cm signal with the systematic. We can notice how the dynamic range is a few orders of magnitude larger and completely outshines the 21-cm signal. For all the differential brightness images, the units are in mK.

map following the relation:

$$\delta T_b^{\rm frg}(U,\,\nu) = \sqrt{\frac{\Omega_{\rm SKA}\,C_l^{\rm syn}(\nu)}{2}}\,[x_l(U) + i\cdot y_l(U)]\,. \tag{5}$$

$\Omega_{\rm SKA}$ is the total SKA-Low solid angle and $U = l/2\pi$. The two quantities x_l and y_l are independent random Gaussian variables with mean zero and variance of one, $\mathcal{N} \sim (0, 1)$. By performing two-dimensional inverse fast-Fourier transform of equation (5), we get the spatial distribution of the foreground contamination $\delta T_b^{\rm frg}(\hat{\boldsymbol{n}},\,z)$. With each lightcone simulation, we fix the random variables seed for the lowest redshift, $z = 7$, and compute equation (4) for the corresponding frequency of the image.

2.4 Mock 21-cm observation

From the simulated coeval cubes described in Section 2.1, we create 3D lightcones with differential brightness $\delta T_b^{\rm sim}(\hat{\boldsymbol{n}}, z) \equiv \delta T_b^{\rm sim}(x, y, z)$ at x, y coordinates for a total box size of 256 cMpc and spatial resolution of $\Delta x = 2$ cMpc, both in comoving units, corresponding to an angular mesh-size of 128^2. This scale corresponds to an angular resolution of $\Delta\theta = 0.77$ arcmin at redshift $z = 7$. The redshift coordinate is divided into 552 bins at equal comoving distance Δx from $z = 11$ to 7, corresponding of frequencies from $\nu_{\rm obs} = 118$ MHz to 178 MHz and a frequency resolution of approximately $\Delta\nu \simeq 0.11$ MHz.

We select one tomographic simulation from the prediction data set as our *fiducial* simulation. In Fig. 1, left column, we show a slice of this *fiducial* lightcone at redshift $z = 8.24$, corresponding to $\nu_{\rm obs} = 152.90$ MHz. At this stage, the simulated lightcone is 50 per cent ionized. The top panel shows the neutral fraction $x_{\rm H\,I}$, with blue and red regions being the neutral and ionized regions, respectively. At the same time, the green colour indicated regions of transitions with $x \simeq 0.5$. The differential brightness is calculated with equation (1) with the approximation discussed in Section 2.1.

From radio interferometry telescope, we can obtain images by gridding the uv-plane and inverse Fourier transform the gridded visibility (Smirnov 2011; Offringa et al. 2014). Image weighting can be applied to the visibilities before the gridding, and in the case of large-scales 21-cm EoR experiment with SKA-Low, the

so-called natural weighting is preferable as the more redundant, short baselines ensures the highest signal-to-noise ratio in the image at the expense of a limited image resolution and large side lobes effect (Briggs 1995). In our case, we do not simulate the 21-cm signal from the visibility space but instead work on images already in the real space. Therefore, to mimic the effect of the limited resolution due to the visibility weighting, in the angular direction, we apply a Gaussian kernel, $G(\hat{\boldsymbol{n}}, z)$, with full width at half-maximum (FWHM) of 21 cm$(1+z)/B$, where $B = 2$ km that corresponds to the maximum baseline of SKA-Low. According to the planned SKA-Low design,[8] it will be densely filled within this 2 km providing enough sensitivity to construct images. The bottom panel in Fig. 1 shows the differential brightness after smoothing the field with $G(\hat{\boldsymbol{n}}, z)$. For reference, this interferometric smoothing corresponds to an angular resolution of ~ 2.9 arcmins at $z \approx 7$ and ~ 4.3 arcmins at $z \approx 11$. In the frequency direction, we apply a top-hat bandwidth filter with the same width as the FWHM in the angular direction. We implement the method explained in Section 2.2 and the parameters listed in Table 1 to simulate the effect of the systematic noise, $\delta T_b^{\rm noise}(\hat{\boldsymbol{n}}, z)$. We create a random field with the same mesh size as the lightcone and add the simulated differential brightness. We then apply the same interferometric smoothing mentioned above, and the result is shown in Fig. 1 (top right-hand panel). As a reference for the reader, this was the network input in our previous work (Paper I).

In this paper, we want to extend our previous effort as we want to recover the neutral binary map in the presence of contamination due to the synchrotron Galactic foreground, $\delta T_b^{\rm frg}(\hat{\boldsymbol{n}}, z)$. The result of the model described in Section 2.3 is shown in Fig. 1 (bottom right-hand panel). As we can see, the dynamic range of the observed changes drastically. Our previous work showed that our method is sensitive to the SNR level between the noise and the 21-cm signal. Therefore, we need to introduce an additional pre-processing step in our framework to mitigate foreground contamination and decrease the dynamic range of the contaminated images before providing them for network training. We will discuss this method in more detail in Section 3.

We can describe our mock observation pipeline by combining the components and operations described here above as (e.g. Liu & Shaw 2020)

$$\delta T_{\rm obs}(\hat{\boldsymbol{n}}, z) = \delta T_b^{\rm sim}(\hat{\boldsymbol{n}}, z) + \delta T_b^{\rm frg}(\hat{\boldsymbol{n}}, z) + \delta T_b^{\rm noise}(z). \quad (6)$$

For each realization of the lightcone $\delta T_{\rm obs}(\hat{\boldsymbol{n}}, z)$, illustrated with Fig. 1, we calculate the mean along the frequency channels,

$$\delta \overline{T}_{\rm obs}(z) = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \delta T_{\rm obs}(x_i, y_j, z), \quad (7)$$

where N_x and N_y are the dimension in the angular-direction of the 128^2 mesh. We subtract this quantity from $\delta T_{\rm obs}$ to account for the effect of the null baseline in interferometry telescopes. For this reason, the colour bar in the figure shows a negative value. W convolve the subtracted term with the Gaussian kernel G mentioned above

$$\delta \widetilde{T}_{\rm obs}(\hat{\boldsymbol{n}}, z) = \int_{\Omega_{\rm SKA}} \left[\delta T_{\rm obs}(\hat{\boldsymbol{n}}', z) - \delta \overline{T}_{\rm obs}(z)\right] \cdot G(\hat{\boldsymbol{n}} - \hat{\boldsymbol{n}}', z)\, {\rm d}\hat{\boldsymbol{n}}'. \quad (8)$$

This result constitutes a realistic mock observation of the SKA-low interferometric telescope, including systematic noise, Galactic foreground contamination, and telescope limited resolution effect. We employ this pipeline to create the training, validation and *random testing set*. In Section 3, we explain how we pre-process this type of data before inputting it into our neural network.

[8]The construction document can be found at https://www.skao.int/en/resources/402/key-documents.

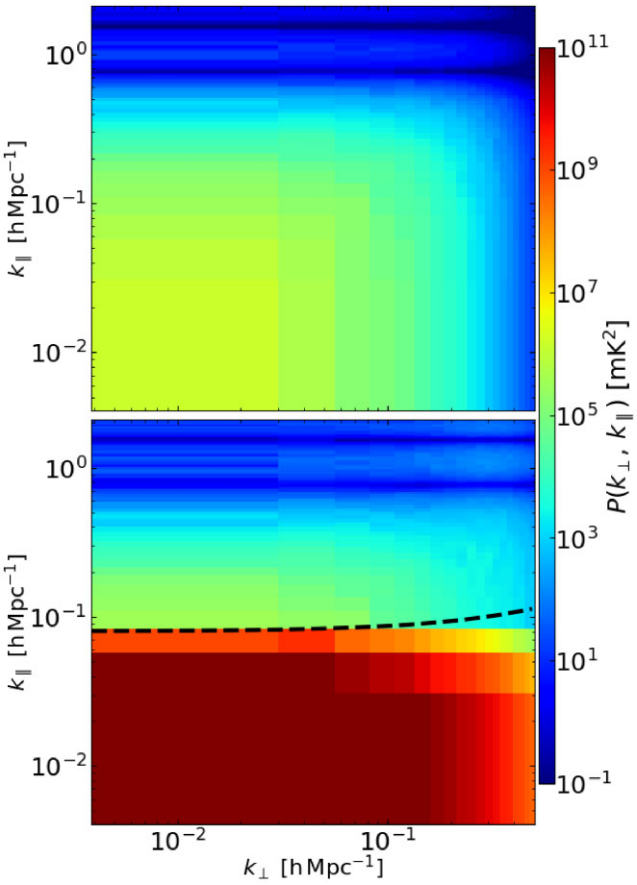

Figure 2. Cylindrical power spectra for a lightcone subvolume centred at redshift $z_c = 8.24$ and frequency depth of ± 10 MHz. *Top Panel*: 2D Power spectra from the simulated 21-cm signal only. *Bottom Panel*: Same quantity but with the galactic foreground contribution. The black dashed line indicates the wedge slope with $\theta = 2.25^\circ$ and $b = 8 \times 10^{-2}$ h Mpc^{-1}.

Finally, we create an additional field that serves as the target of the network training. We apply the interferometric smoothing explained above to the simulated neutral fraction field $x_{\rm H\,I}$ (the top left-hand panel of Fig. 1). We then choose a threshold of $x_{\rm th} = 0.5$ to discern the ionized and neutral regions. The result is a binary lightcone, $x_{\rm HI}^B(\hat{\boldsymbol{n}}, z)$, where neutral and ionized regions are classified by 1 and 0, respectively. For a visual comparison, we overplot the contour of this binary field as a black line in the top right-hand panel of Fig. 1.

3 FOREGROUND MITIGATION

As we outlined in Section 2.4, foreground contamination poses a huge problem in detecting the 21-cm signal, as this signal is several orders of magnitude fainter in comparison. In Fig. 2, we illustrate the effect of the foreground contamination on the two-dimensional (2D) cylindrical power spectrum for a lightcone subvolume centred at redshift $z_c = 8.24$ and frequency width of $\Delta\nu \pm 10$ MHz. This quantity of the 21-cm signal (top panel) is compared with the same signal contaminated by the Galactic foreground signal (bottom panel). We observe that the contamination is visible at $k_\parallel \leq 10^{-1}$ Mpc/h

with signal intensity of $\geq 10^9\,\mathrm{mK}^2$. The black dashed line in the figure indicates the foreground wedge. We will discuss this line later in Section 3.2. To reduce the dynamic range of the foreground contaminated images to a level that is manageable for the neural network, we include a pre-processing step on the observed data, $\delta\widetilde{T}_{\rm obs}(\hat{\boldsymbol{n}}, z)$. Hereafter, we refer to the resulting images of this pre-process as *residual lightcone* or *images*, $\delta T_{\rm res}(\hat{\boldsymbol{n}}, z)$.

In foreground mitigation, we can consider two methods: foreground subtraction or avoidance (Chapman & Jelić 2019). Here, we consider three of the former cases, namely principal component analysis (PCA), Gaussian regression processes (GPR), and Polynomial fitting, and one of the latter techniques, Wedge removal. In this section, we briefly describe four different pre-processing methods that we test and we provide the residual image in Fig. 3 for each method. The top panels show the residual image of the example illustrated in Fig. 1, while black contours indicate the ground truth. The bottom panel shows the 2D cylindrical power spectrum for the fiducial lightcone subvolume centred at $z_c = 8.24$ and frequency depth of ± 10 MHz.

3.1 Principal component analysis

PCA is a commonly used method to remove foregrounds in 21-cm experiments (e.g. Alonso et al. 2015; Cunnington et al. 2023; Chen et al. 2023a). The method exploits the fact that foregrounds have large amplitude and smooth frequency coherence. PCA simultaneously identifies the largest foreground components and an optimal set of basis functions that describe the frequency structure of the foregrounds. As the foregrounds are highly correlated in frequency, the frequency–frequency co-variance matrix of the foregrounds will have a particular eigensystem where most of the information can be sufficiently described by a small set of very large eigenvalues, the other ones being negligibly small. Thus, we can attempt to subtract the foregrounds by eliminating the components corresponding to the eigenvectors of the frequency co-variance matrix with the largest associated eigenvalues. In practice, we remove four components, which captured most of the variance of the foreground modes. PCA is a relatively fast and computationally efficient method that requires no prior assumptions about the foregrounds or the 21-cm signal. However, PCA is not well-suited to handle non-linear relationships between the foregrounds and the 21-cm signal, and it can struggle to remove residual foregrounds not well-described by the largest components.

In Fig. 3, left column, we show the residual image at $z_c = 8.24$, on top. After removing the first four components with PCA decomposition on the 20 MHz subvolume of the fiducial lightcone, we obtain this image. On the bottom panel, we show the corresponding 2D power spectra.

3.2 Wedge remove

We consider another pre-process that focuses on discarding the Fourier modes dominated by foreground contamination. This method assumes that the contaminated modes are contained in specific regions in the $k_\perp - k_\parallel$ space, named the *foreground wedge*. These contaminated k-modes can be defined by (e.g. Liu, Parsons & Trott 2014; Murray & Trott 2018)

$$k_\parallel \leq |\boldsymbol{k}_\perp| \, \frac{H(z)}{1+z} \int_0^z \frac{\mathrm{d}z'}{H(z')} \cdot \sin\theta + b, \tag{9}$$

where $H(z)$ is the Hubble parameter and $k_\perp$ is the Fourier component perpendicular to the line of sight. θ is the angular size of the field of view, which can be interpreted as the horizon limit angle. b is the bias that accounts for the presence of an intrinsic foreground limit at low $k_\parallel$ values. Pessimistic and arguably more realistic assumptions consider the horizon limit to $\theta = 90^\circ$ justified by antenna side-lobes effect (Dillon et al. 2014; Pober et al. 2014). In our case, we select $\theta = 2.25^\circ$, corresponding to the field of view (FoV), at redshift $z = 7$ and comoving size of 256 cMpc, of our data set. We then select $b = 8 \times 10^{-2}\,\mathrm{h\,Mpc^{-1}}$ based on the 2D cylindrical power spectrum shown in the right-hand panel of Fig. 2. The dashed black line indicates equation (9) for the θ and b mentioned before.

In this work, we employ a simplified version of the code developed by Prelogović et al. (2021). Here, we give a brief description, referring the reader to the original paper for more details. First, we perform a 2D Fourier transform in the angular direction of a lightcone subvolume, equation (8), centred at redshift z_c and with a given frequency depth, $\pm\Delta\nu$. Subsequently, an iterating procedure along the line-of-sight axis calculates equation (9) and sets the k-modes that obey the condition to zero. A Blackmann–Harris taper function of the same angular and redshift size is multiplied by the lightcone to avoid artificial ringing in the Fourier space. However, this taper has the limitation that at low $k_\parallel$, it reduces the Fourier-space side lobes, while the opposite effect occurs at high $k_\parallel$. Finally, we do an inverse Fourier transform to regain the real-space lightcone subvolume.

An example of data with the foreground contamination removed by this algorithm can be seen in the second column of Fig. 3. From the residual image (top panel), we see a large portion of the foreground residual is still present. The bottom panel shows the 2D cylindrical power. The dark blue colour indicates the $k_\perp - k_\parallel$ modes where the wedge removes method is applied.

3.3 Gaussian process regression

The GPR method was developed in Mertens et al. (2018) to separate foregrounds from 21-cm signal by modelling the two components as a stochastic process and separating them using a Bayesian approach. The method involves constructing a prior statistical model of the foregrounds and the 21-cm signal and then using the model to estimate the posterior distribution of the 21-cm signal given the observed data. This is done by assuming that the foregrounds and 21-cm signals are realizations of Gaussian processes, fully defined by their covariance. The selection of the prior covariance model in GPR is made under a Bayesian framework by maximizing the marginal likelihood. The Matérn class of covariance functions is commonly used as prior covariance for the different data components. Following Mertens et al. (2018), a radial basis function kernel is used as the prior covariance model for the foreground component, while an exponential kernel is used for the 21-cm signal. This method can effectively remove foreground contamination from the 21-cm signal and has the advantage of being able to incorporate prior knowledge about the signal and foregrounds. However, it requires accurate modelling of the foregrounds and assumptions about the statistical properties of the signal and foregrounds.

In Fig. 3, third column, we show the result obtained by the GPR presented here above. Similar to PCA, see Section 3.1, GPR removes a good portion of the foreground contamination providing a better contrast between the 21-cm emitting regions and the ionized one. For instance, the regions around $(x, y) = (225, 100)$Mpc and $(x, y) = (150, 200)$Mpc. From the 2D power spectra at $k_\parallel > 3 \times 10^{-2}$, we see more signal when compared to PCA pre-process data.

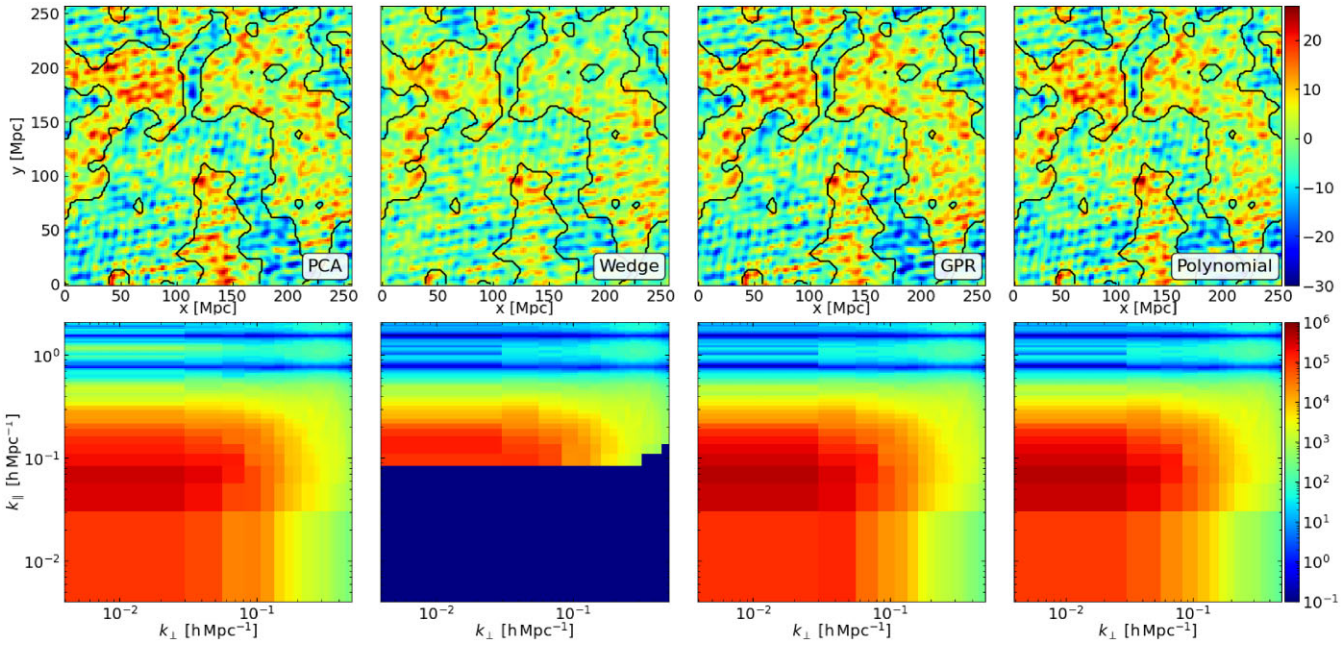

Figure 3. Comparison between different foreground mitigation methods. From left to right, we have PCA, wedge removal, GPR, and polynomial fitting. First row, a visual example at redshift $z = 8.24$ of the residual image after the corresponding method. Second row, the cylindrical power spectrum for a lightcone subvolume centred at $z_c = 8.24$ and frequency depth ± 10 MHz.

3.4 Polynomial fitting

We can also use polynomial fitting to remove foreground contamination from the 21-cm signal (Wang et al. 2006; Alonso et al. 2015). The method involves modelling the foregrounds as a smooth polynomial function in log space and fitting this function to the observed data, $\delta\widetilde{T}_{\rm obs}$.

$$\log\left(T(\hat{\boldsymbol{n}}, z)\right) = \sum_{k=1}^{N_{fg}} \alpha_k(\hat{\boldsymbol{n}}) \left[\log\left(\frac{\nu_0}{1+z}\right)\right]^{k-1}. \tag{10}$$

Here, ν_0 is the 21-cm frequency and $N_{\rm fg}$ indicates the polynomial degree. In our study, we consider a fourth-degree polynomial. The resulting fit is then subtracted from the data to remove the foreground contamination $\delta T_{\rm res} = \delta\widetilde{T}_{\rm obs} - T(\hat{\boldsymbol{n}}, z)$.

This approach has the advantage of being simple and computationally efficient but may not be as effective at removing foregrounds as other, more sophisticated methods. One limitation of the polynomial fitting is that it assumes the foregrounds can be well-described by a smooth polynomial, which may not always be the case (e.g. Thyagarajan et al. 2015). Additionally, if the polynomial fit is not in high enough order, it may leave some foregrounds in the data, while an overly high-order polynomial may also remove the signal. The polynomial fitting has been combined with other foreground removal methods in some studies to improve the overall performance of the foreground removal process.

In Fig. 3, fourth column, we show the result obtained by the polynomial fitting. In both cases, from the residual image and the 2D power spectra, visually, we see similar results to GPR, see Section 3.3, with a more considerable difference between the positive (neutral) and negative (ionized) regions in the residual image, although presenting the same level of residual foreground located at $(x, y) \sim (80, 125)$ Mpc as in the other methods.

4 U-NET FOR 21-CM IMAGE SEGMENTATION

The network architecture of `SegU-Net v2` is the same as in Paper I. The only implementation consists of a simplistic hyperparameter optimization analysis on seven network hyperparameters. In Appendix A, we give a brief overview of the hyperparameter space exploration method we employed and in Table A1, we list the six best-performing set-ups we found. Moreover, in Appendix B, we present a first attempt to open the *black box* and performed a Gradient-weighted Class Activation Mapping (Grad-CAM; Selvaraju et al. 2019) importance analysis to highlight the features in the input image that the network employs to identify and predict the neutral regions from residual images. In Fig. B1, we give a visual representation of the Grad-CAM importance analysis we performed.

4.1 Network architecture

Here, we give a brief description of our network architecture. We refer the reader to our previous work for more details. `SegU-Net` is a U-shaped deep convolutional neural network composed of a contracting (encoder) and an expanding path (decoder). The former has two convolutional blocks, followed by the 2D averaging pooling operation of size 2^2 and a dropout layer with a 5 per cent rate, `Encoder-Level = 2*ConvBlock+AvrgPool + Drop`. A convolutional block consists of a 2D convolutional layer with kernel size 7^2, followed by batch normalization and Rectified Linear Unit (`ReLU`) activation function, `ConvBlock = Conv2D+BN + ReLU`. The latter path consists of transposed 2D convolution followed by the concatenation with the corresponding output of the convolutional encoder block, dropout layer and two convolutional blocks, `Decoder-Level = TConv2D+CC+Drop + 2*ConvBlock`. This structure is repeated four times for both the encoder and decoder. At each level, the pooling operation halves the angular dimension of the input and doubles the number of channels. The network takes as

input a redshift slice from the residual lightcone, $\delta T_{\rm res}$, and outputs the corresponding 2D binary image, $x_{\rm HI}^{B}$.

4.2 Data set

We generated a large set of realizations of the SKA multifrequency tomographic data set by changing the initial conditions and the following three astrophysical parameters. We sample the high-redshift galaxy efficiency ζ and the MFP of ionizing photons $R_{\rm mfp}$ with a normal distribution with mean and variance $\mathcal{N}(82,\ 18)$ and $\mathcal{N}(17.5\,{\rm Mpc},\ 4.5\,{\rm Mpc})$, respectively. At the same time, the minimum virial temperature for star-forming haloes $T_{\rm vir}^{\rm min}$ is sampled in logarithmic space with distribution $\mathcal{N}(4.7,\ 0.2)$. We chose this sampling of parameters because we want the global volume-averaged neutral fraction $\overline{x}_{\rm HI}$ of all data to be at least greater than 90 per cent at redshift $z = 11$ and less than 10 per cent at redshift 7. Moreover, with this parameter sampling, we can postulate the spin-saturation assumption, $T_{\rm S} \gg T_{\rm CMB}$, which assures that the differential brightness is strictly positive and that neutral hydrogen is correlated with a positive signal in each image.

In this work, we updated the data set from Paper I for a total of 10 000 samples for the network training and 1500 for validation. Once the network is trained, we will test its accuracy and generalization ability on an additional 300 mock observations during the prediction step. We will refer to this data set as the *random testing set*. The training data set is employed during the forward- and back-propagation (Rumelhart & Zipser 1985), while the validation data set is used to validate the accuracy of network results during training. We want to clarify that we trained `SegU-Net v2` on $\delta T_{\rm res}$ data pre-processed only with the PCA eigen-decomposition on the full redshift range, $z = 7$ to 11, which is explained in Section 3.1. The testing data set is an independent set of simulations on which we will validate the final results of the trained network.

4.3 Metrics

We consider a true positive (TP) detection to be the number of pixels correctly identified as neutral, while a true negative (TN) is the opposite. False positives (FP) and false negatives (FN) represent the number of pixels wrongly classified as neutral or ionized. Therefore, we can define the Matthews correlation coefficient (MCC) for quantifying the accuracy of our network predictions as

$$r_{\phi} = \frac{\rm TP \cdot TN - FP \cdot FN}{\sqrt{\rm (TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{11}$$

This metric can have values between $-1 \leq r_{\phi} \leq 1$, quantifying the quality of binary field (two-class) classifications. A negative value indicates anticorrelation, zero represents a completely random classification, and positive values indicate a positive correlation. For a direct comparison with previous studies on segmentation of 21-cm image data (e.g. Gagnon-Hartman et al. 2021), we define three additional statistical metrics as follows:

$$\text{Accuracy} = \frac{\rm TP + TN}{\rm TP + FP + FN + TN}. \tag{12}$$

Here, this metric indicates how well a model is able to predict the target variable correctly:

$$\text{Precision} = \frac{\rm TP}{\rm TP + FP}. \tag{13}$$

This second metric refers to the level of consistency or repeatability of a predicted value. While accuracy and precision are important metrics in evaluating the performance of a network, they may not be sufficient in certain scenarios. For instance, in our binary classification problem, there can be scenarios when neutral regions can be much rarer than ionized regions and vice versa. In this case, accuracy can be misleading as the model may achieve high accuracy by simply predicting the majority class for all instances. Precision and recall are more informative metrics in such cases as they consider the class imbalance:

$$\text{IoU} = \frac{\rm TP}{\rm TP + FP + FN}. \tag{14}$$

However, here, we include the third additional metric, the Intersection over Union (IoU), that quantifies how well the predicted neutral region of interest overlaps with the true one. We will use these metrics later in Section 5.2.

In our case, we are targeting binary maps that indicate the location in the sky at a given redshift as either neutral or ionized. Therefore, an easy way for the reader to interpret the results is in the number of pixels guessed correctly or wrongly. For this reason, we introduce the false positive rate (FPR), also referred to as non-specificity, and the true positive rate (TPR), also known as sensitivity:

$$\text{TPR} = \frac{\rm TP}{\rm TP + FN} \quad , \quad \text{FPR} = \frac{\rm FP}{\rm FP + TN}. \tag{15}$$

The former quantity gives the percentage of neutral pixels (positive case in our context) correctly identified as neutral. A value of TPR = 1 will indicate that the network identified all the neutral pixels correctly. Otherwise, 1 − TPR indicates the percentage of pixels falsely classified as ionized. Similarly, the FPR gives the percentage of pixels falsely detected as neutral.

4.4 Per-pixel error estimation

The error calculation uses the same method as in Paper I. In the prediction step, we employ temporal time augmentation (TTA) operations (Perez & Wang 2017; Wang et al. 2020) on the network input data to create several copies of the same realization but that we modify by rotating and vertical/horizontal flip operation. In this work, we fix the axis of symmetry and rotation to the frequency direction. Thus, the number of manipulations was reduced to a sample of 16 copies. This number corresponds to the maximum independent operations we can apply to an image. `SegU-Net v2` then gives a prediction for each modified copy that is then rotated or flipped back to obtain a different prediction of the same input image. We calculated the standard deviation, $\sigma_{\rm std}$, on the 16 copies and obtained a per-pixel uncertainty map as shown in Fig. 4, bottom panel. The method is simple but efficient, showing how difficult it was for the network to give the predicted binary field for each pixel in the image.

5 RESULTS

This section discusses the result obtained with `SegU-Net v2` acting on data pre-processed with the PCA foreground removal method as explained in Section 3.1. Here, we evaluate the result on the predicted binary maps and the network performance on the different methods (illustrated in Section 3) in Sections 5.1 and 5.2, respectively. Finally, in Section 5.3, we demonstrate a possible astrophysical application of `SegU-Net v2`.

5.1 Identifying H II regions with `SegU-Net v2`

In Fig. 4, we visually evaluate one realization of the network predicted neutral (red) and ionized (blue) regions. We refer to this

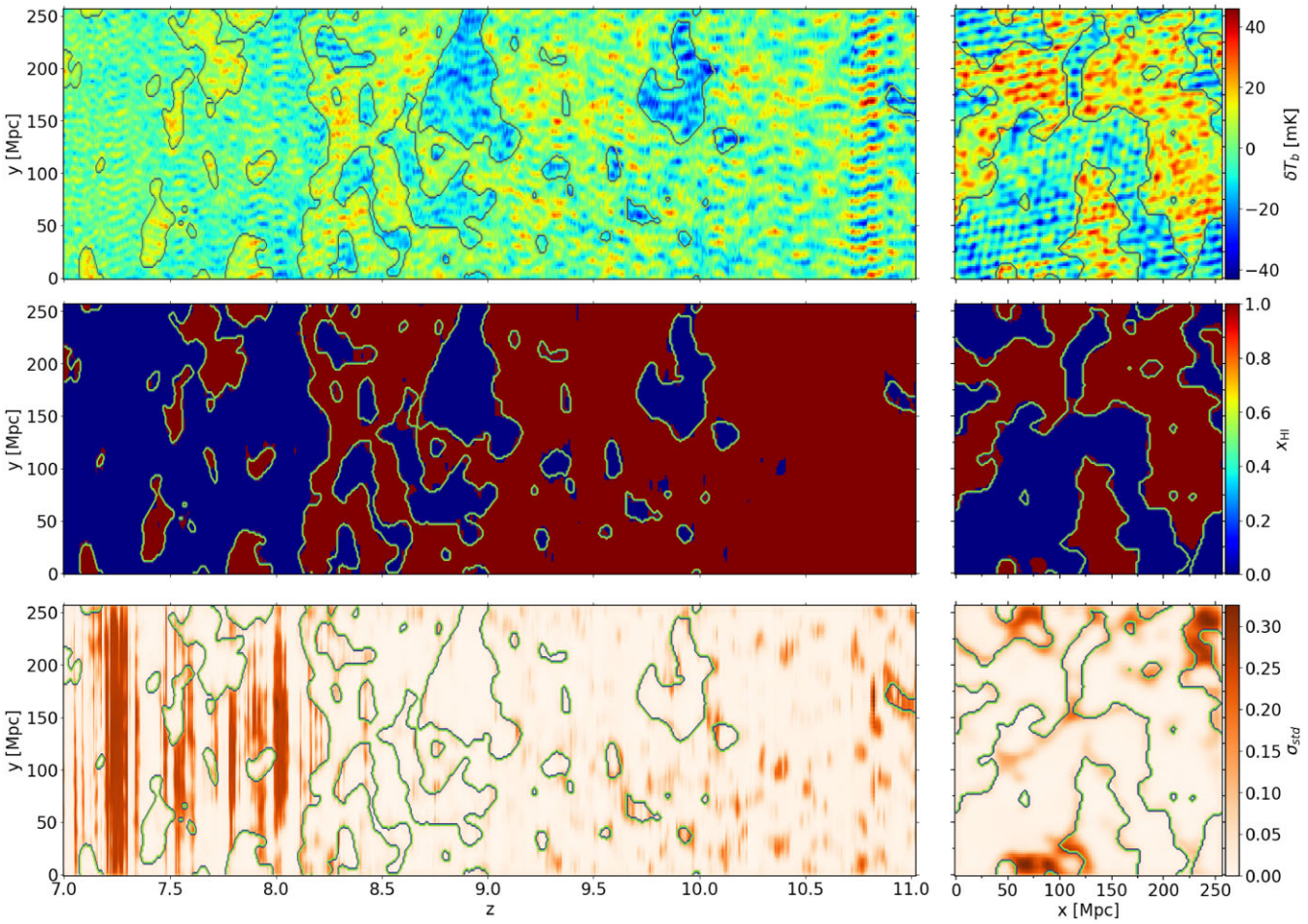

Figure 4. Visualization of the different fields for our fiducial lightcone. *Top left*: For a given position on the x-direction, the redshift evolution of the residual lightcone after the PCA pre-processing step. *Top Right*: Residual image at redshift $z = 8.24$ ($\overline{x}_{\rm HI} = 0.5$). Same image as in Fig. 1. *Middle left*: Redshift evolution of the predicted neutral (red) and ionized (blue) lightcones. *Middle right*: Predicted map at the corresponding redshift. *Bottom left*: The corresponding per-pixel error lightcone, orange colour indicates the intensity of the uncertainty. *Bottom right*: The corresponding per-pixel error map. For all panels, we overplot contours that represent the ground truth.

simulated lightcone as the *fiducial* simulation. In the right column, we show a slice at redshift $z = 8.24$ ($\nu_{\rm obs} = 152.90$ MHz), corresponding when the global volume average neutral fraction is $\overline{x}_{\rm HI} = 0.5$. From top to bottom, we show the residual image after the PCA pre-processing employed as the input of the neural network, the binary map predicted with `SegU-Net v2` from the PCA pre-processed data and the derived per-pixel uncertainty, respectively. In the left column, we show the redshift evolution of the same fields along one given direction of the corresponding fields.

First, when we compare the bottom right-hand panel in Fig. 1 with the top right-hand panel in Fig. 4, we can notice that the pre-processing step drastically reduces the signal from $\delta T_b \sim \pm 10^5$ mK to just an observed differential brightness of few tens $\delta T_b \sim \pm 40$ mK. Nevertheless, some of the foreground contamination is still visible. For instance, in Fig. 4 top left-hand panel, we can see that across a few frequency bands at $z \approx 10.8$ presents an anomalous feature. Moreover, we can see that foreground residual is still present between $7 \leq z \leq 8.2$. This signal excess is self-evident in the per-pixel uncertainty for the same redshift range. Some frequency bands are saturated with considerable uncertainty $\sigma_{\rm std} \sim 0.3$. This is because the foreground component is correlated along the frequency direction and is primarily diffused over large angular scales. The foreground residuals thus observe extended features along the z direction over multiple adjacent frequency channels. From the redshift evolution of the predicted binary field (left middle panel), we notice that the network can either falsely detect bubbles when most of the lightcone is still highly neutral, $z \geq 9.5$, or completely miss ionized bubbles that are entirely surrounded by neutral hydrogen. In both cases, the mislabelling is limited to bubbles with sizes close to or smaller than the interferometric smoothing scale, $\Delta x \sim 9$ Mpc, as the network confuses structures with small-scale noise fluctuations. Thus posing a hard limit on the possibility of measuring and detecting the smallest H II bubble close to the instrument resolution. We discuss this further in Section 5.3. This limitation is visible from the recovered binary field at redshift $z = 8.24$ (middle right-hand panel). Here, the detection of the bubbles at 180 Mpc $\leq$ x $\leq$ 210 Mpc is entirely missed. We observe the same outcome for the island of neutral hydrogen at coordinates $(x, y) \approx (75, 75)$ Mpc. These erroneous findings are associated with a moderate to high uncertainty $\sigma_{\rm str} \geq 0.2$. As we mentioned above, the per-pixel uncertainty shows that at the early stage of reionization, $z > 9$, most of the uncertainty is either situated around small H II volumes, $V \leq (10^3\,{\rm Mpc})^3$, or at the border between neutral and ionized regions. On the other hand, at the late stages, $z < 8.2$,

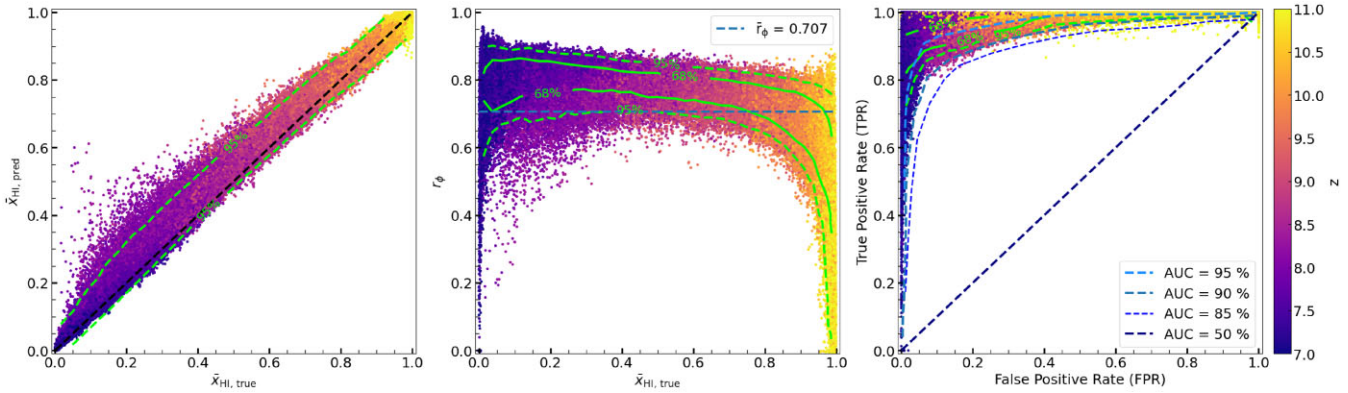

Figure 5. Statistical analysis of the predicted binary maps for the testing data set. Each point indicates an image at a given redshift in the colour bar. *Left-hand panel*: Correlation plot between the ground truth volume average neutral fraction, $\overline{x}_{\rm HI,\,true}$, against the predicted, $\overline{x}_{\rm HI,\,pred}$. *Right-hand panel*: Matthew correlation coefficient r_ϕ against global volume-averaged neutral fraction. The dashed blue line indicates the redshift averaged r_ϕ. Here, solid green lines indicate the 68 per cent (1σ) and dashed green lines the 95 per cent (2σ) data contour. *Right-hand panel*: Receiver operating characteristic curve for the same data set. The dashed line of different blue shades indicates the percentage of reliability of the prediction.

high uncertainty is mostly located in the vast, interconnected ionized IGM.

In Fig. 5, we show three statistical analyses for the entire *random testing set*. In the left-hand panel, we show the correlation plot between the true global averaged neutral fraction $\bar{x}_{\rm HI,true}$ against the predicted $\bar{x}_{\rm HI,pred}$. The dashed green line indicates the 95 per cent data contour, corresponding to a 2σ difference from the ground truth. The 2σ contour clearly shows a deviation on the left-hand side of the black dashed line (perfect correlation), indicating that the predicted images tend to be considered more neutral than they should be. This trend is more visible at lower redshift $z < 8.5$ ($\bar{x}_{\rm HI,true} < 0.4$) as more points reside outside of the 95 per cent percentile. This behaviour can be motivated by the presence of residuals from the foreground that the PCA process could not remove. As we mention in Section 3.1, we consider the first four components to contain most foreground information. These components are most representative at higher frequency as the foreground amplitude increases inversely proportional to redshift, equation (4). Therefore, for tomographic data with a wide redshift range, the decomposition can under-represent foreground contamination at lower redshift, resulting in more residuals when we reconstruct the image from the remaining components at the corresponding redshift slices. This effect is visible in the uncertainty map in Fig. 4.

In Fig. 5, middle panel, we show the correlation coefficient against the same quantity as before, $x_{\rm H\,I,\,true}$. Each point corresponds to an image at a redshift indicated by the colour bar. We add the 68 per cent data contour (solid line) on this panel, corresponding to a 1σ difference from the ground truth. We first noticed that we obtain a global accuracy that is approximately 15 per cent lower, $\bar{r}_\phi = 0.71$, compared to our previous work in Paper I. This lower score with the same network structure and architecture is justified because any signal extrapolation in foreground contamination is extremely arduous compared to forecasting in the presence of just telescope systematic noise. Moreover, as we stated before, we notice that at lower redshift $z < 8.5$ ($\bar{x}_{\rm HI,true} < 0.4$), a sizable portion of the redshift slices have a difference larger than 2σ. This behaviour is also evident from the increase of the uncertainty map in Fig. 4 for images at $z < 8.5$.

Lastly, in Fig. 5, right-hand panel, we show the correlation between the true positive rate (TPR), also known as sensitivity, and the FPR, also known as non-specificity, on the *random testing set*. In our case, these quantities indicate the percentage of pixels correctly labelled as neutral and the fraction of pixels mislabelled as ionized, respectively. This plot is known as the receiver operating characteristic (ROC) curve, and it is a standard analysis in classification problems as it gives an intuitive overall performance of the method. The results from our network show that most of the realizations with redshift range $z \in [7.5,\ 10]$ are located in the top-left corner, representing the ideal performance or perfect classification. This indicates that most binary maps have high sensitivity and specificity, i.e. neutral and ionized regions are correctly identified. Data points close to the diagonal line indicate that the method performance is not much better than a random classifier. In our case, this is true for the values at the extreme of the redshift range. The data points on the top-right corner have high sensitivity but low specificity, meaning that the network labels correctly neutral regions, from equation (15), left metric, FN $\ll$ TP, while misclassifying most of the ionized pixels as neutral, TN $\ll$ FP. This is the case for images with $z > 10$; however, at this redshift, the images are mostly neutral; thus, the incorrect detection is limited to a few pixels of the image. The data point in the bottom-left represents the opposite situation where the network has high specificity but low sensitivity. This scenario indicates that the model is not able to differentiate well between neutral and ionized instances, from equation (15), right function, TP $\ll$ FN and FP $\ll$ TN. We see the opposite trend as in the previous case, where images with $z \sim 7$ occupy this instance. Another important quantity derivable from the ROC curve is the area under the curve (AUC). This quantity gives an overall evaluation of the classification method. In Fig. 5, right-hand panel, we overplot four curves that represent different AUC scores. In our case, we can see that the network performs well as the *random testing set* points are mostly located above the 85 per cent line and are well centred around the AUC = 95 per cent.

5.2 Sensitivity to the choice of pre-processing method

We trained `SegU-Net v2` on the signal that is pre-processed using the PCA method. Therefore, it is vital to investigate how sensitive the trained model is to the pre-processing method used to mitigate foreground. Here, we test `SegU-Net v2` on the foreground mitigation processes we presented in Section 3. We cannot use the entire lightcone as the GPR module currently available has been validated only for a bandwidth of 20 MHz. From the entire

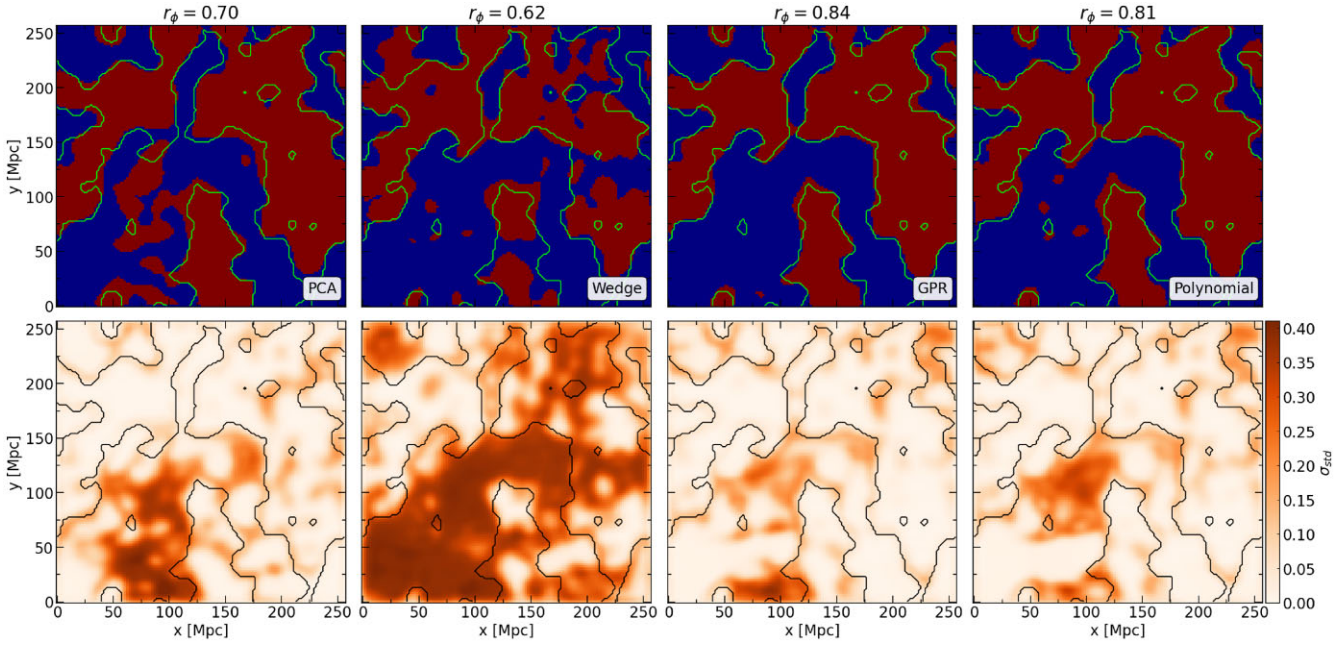

Figure 6. Comparison of the recovered binary field from different foreground mitigation pre-processes. We have PCA, wedge removal, GPR, and polynomial fitting from left to right. *Top panels*: A visual example of the recovered binary map at redshift $z = 8.24$ after the mentioned pre-processing step. The red/blue indicates the predicted neutral/ionized regions, while the green contour indicates the ground truth. *Bottom panels*: The corresponding per-pixel uncertainty map derived by `SegU-Net v2`. The orange indicates the intensity of the uncertainty, defined as a general standard deviation. The title includes the resulting r_ϕ at this redshift.

lightcone, we use three subvolume centred at redshift $z_c = 7.68$, 8.24, and 8.97 with frequency size of 20 MHz, corresponding to 172, 181, and 186 redshifts bins from $z \in [7.19,\ 8.24]$, $[7.68,\ 8.88]$, and $[8.31,\ 9.72]$, respectively. The volume average neutral fraction of these subvolumes is $\overline{x}_{\rm HI} \simeq 0.25$, 0.50, and 0.75, corresponding to the late, middle, and early stages of reionization, respectively.

We then apply four different foreground mitigation pre-processing steps to each subvolume: PCA, wedge remove, GPR, and Polynomial fitting. From the residual volumes, we predict the neutral/ionized regions from the trained `SegU-Net v2`, with PCA, pre-processing step as presented in Section 5.1. By applying different foreground mitigation processes, we can quantify the robustness and adaptability of our trained network.

5.2.1 *Visual evaluation*

We visually compare the middle stage of reionization subvolume for the four cases in Fig. 6. From the left to right column, we have PCA, wedge remove, GPR, and polynomial fitting, respectively. The top panels visually compare an image at the subvolume central redshift $z_c = 8.24$ for the different pre-processes. In the bottom panels, we show the corresponding uncertainty map from the `SegU-Net v2`. We notice that for the case of the fiducial simulation, the polynomial fitting and GPR pre-processing obtain similar results with correlation $r_\phi(z_c) = 0.81$ and $r(z_c)_\phi = 0.84$, respectively. The former case appears to overestimate the extent of the neutral regions (see at position $(x, y) \simeq (75, 125)$ Mpc) as well as falsely detecting the presence of isolated neutral island in the vast ionized region, for instance, see around $(x, y) \sim (75, 100)$ Mpc. The PCA obtains approximately 10 per cent less accuracy, $r_\phi(z_c) = 0.70$, its limitation comes forth when predicting the vast ionized region (see at position 50 Mpc $\leq x \leq$ 125 Mpc and 75 Mpc $\leq y \leq$ 125 Mpc) as the network is over-predicting the presence of an interconnected neutral hydrogen region. Wedge Remove method has the lowest performance, with $r_\phi(z_c) = 0.62$. In this example, the pre-process forecasts an excess of neutral hydrogen outside the ground truth. On the other hand, this method underestimates its presence within the extensive neutral cloud. In Table 2 (third column), we show the resulting $r_\phi(z_c)$ for each pre-process.

Among the methods presented, the Wedge Remove method appears to be the least efficient for `SegU-Net v2`. The uncertainty map in Fig. 6 shows that the wedge remove method has high incertitude in the vast interconnected H II regions, for $x \in [0,\ 125]$ Mpc and $y \in [0,\ 150]$ Mpc, as well as between nearby H I regions, for instance at $(x,\ y) \simeq (120, 160)$ Mpc. The presence of a higher foreground residual compared to the other methods (visible in the same region in Fig. 3) indicates that lower performance is attributed to a harsh and perhaps undisclosed subtraction that does not aim at portraying the foreground contamination but rather removes its contribution. Overall, the GPR method, followed by PCA decomposition, appears to give an advantage compared to the other pre-processing. At the same time, all the cases fail to detect ionized or neutral regions of sizes close to the interferometric smoothing scale, $\Delta x \simeq 9$ Mpc.

5.2.2 *Redshift evolution*

In Fig. 7, we show the redshift evolution of the Matthew correlation coefficient r_ϕ for the four different methods. On each panel, we show the results from the early ($z_c = 8.97$, in red), middle ($z_c = 8.24$, in green) and late ($z_c = 7.68$, in blue) stage of reionization subvolumes with the corresponding error bar represented by the shadow area. The horizontal dashed line denotes the redshift averaged correlation coefficient, $\overline{r}_\phi$. In Table 2 (fourth column), we show the resulting $\overline{r}_\phi$ for each subvolume and subvolume. Based on this quantity, we notice that the ranking goes by the GPR method with $\overline{r}_\phi = 0.71$ at $z_C = 7.68$, 0.67 at $z_C = 8.24$ and 0.63 at $z_C = 8.97$, followed by the

Table 2. Result summary of the predicted binary field for the tested pre-processing step on the three lightcone subvolume at representative stages of reionization.

z_c	Pre-process	$r_\phi(z_c)$	Accuracy	Precision	IoU	TPR (per cent)	FPR (per cent)	$\overline{r}_\phi$	$\overline{x}_{\rm HI}$	$\overline{R}_C$ [cMpc]
7.68	Ground truth	–	–	–	–	–	–	–	**0.24**	**19.89**
	All z PCA	0.78	0.94	0.81	0.67	83.12	5.98	0.82	0.26 ± 0.12	$21.62^{+4.34}_{-3.90}$
	PCA	0.75	0.89	0.81	0.70	84.08	8.32	0.73	0.26 ± 0.15	$17.96^{+8.66}_{-4.66}$
	Wedge	0.55	0.80	0.65	0.20	52.56	49.82	0.28	0.07 ± 0.12	$11.96^{+9.46}_{-2.54}$
	GPR	0.77	0.90	0.82	0.73	86.22	7.97	0.77	0.28 ± 0.14	$19.75^{+6.93}_{-5.03}$
	Polynomial	0.75	0.89	0.82	0.70	83.81	8.09	0.76	0.27 ± 0.15	$19.17^{+7.84}_{-5.18}$
8.24	Ground truth	–	–	–	–	–	–	–	**0.45**	**29.54**
	All z PCA	0.84	0.91	0.86	0.72	90.60	5.32	0.80	0.48 ± 0.07	$31.37^{+3.09}_{-3.93}$
	PCA	0.70	0.85	0.81	0.75	91.12	21.48	0.69	0.49 ± 0.11	$27.65^{+9.13}_{-6.12}$
	Wedge	0.62	0.64	0.65	0.22	74.95	45.43	0.22	0.16 ± 0.13	$15.20^{+24.13}_{-6.18}$
	GPR	0.84	0.92	0.91	0.85	93.02	9.44	0.75	0.48 ± 0.09	$29.14^{+5.26}_{-4.89}$
	Polynomial	0.81	0.91	0.89	0.83	92.18	11.01	0.74	0.49 ± 0.10	$29.21^{+5.83}_{-5.21}$
8.97	Ground truth	–	–	–	–	–	–	–	**0.72**	**49.09**
	All z PCA	0.78	0.92	0.93	0.85	93.43	15.52	0.76	0.74 ± 0.29	$48.57^{+5.93}_{-6.36}$
	PCA	0.72	0.88	0.90	0.85	93.80	23.75	0.68	0.75 ± 0.33	$46,06^{+9.47}_{-8.74}$
	Wedge	0.53	0.51	0.76	0.37	70.96	77.96	0.19	0.38 ± 0.11	$28.57^{+11.46}_{-8.54}$
	GPR	0.75	0.90	0.91	0.86	94.53	22.10	0.72	0.74 ± 0.28	$46.64^{+7.21}_{-7.52}$
	Polynomial	0.74	0.89	0.90	0.86	94.53	22.78	0.72	0.74 ± 0.29	$47.24^{+7.07}_{-7.81}$

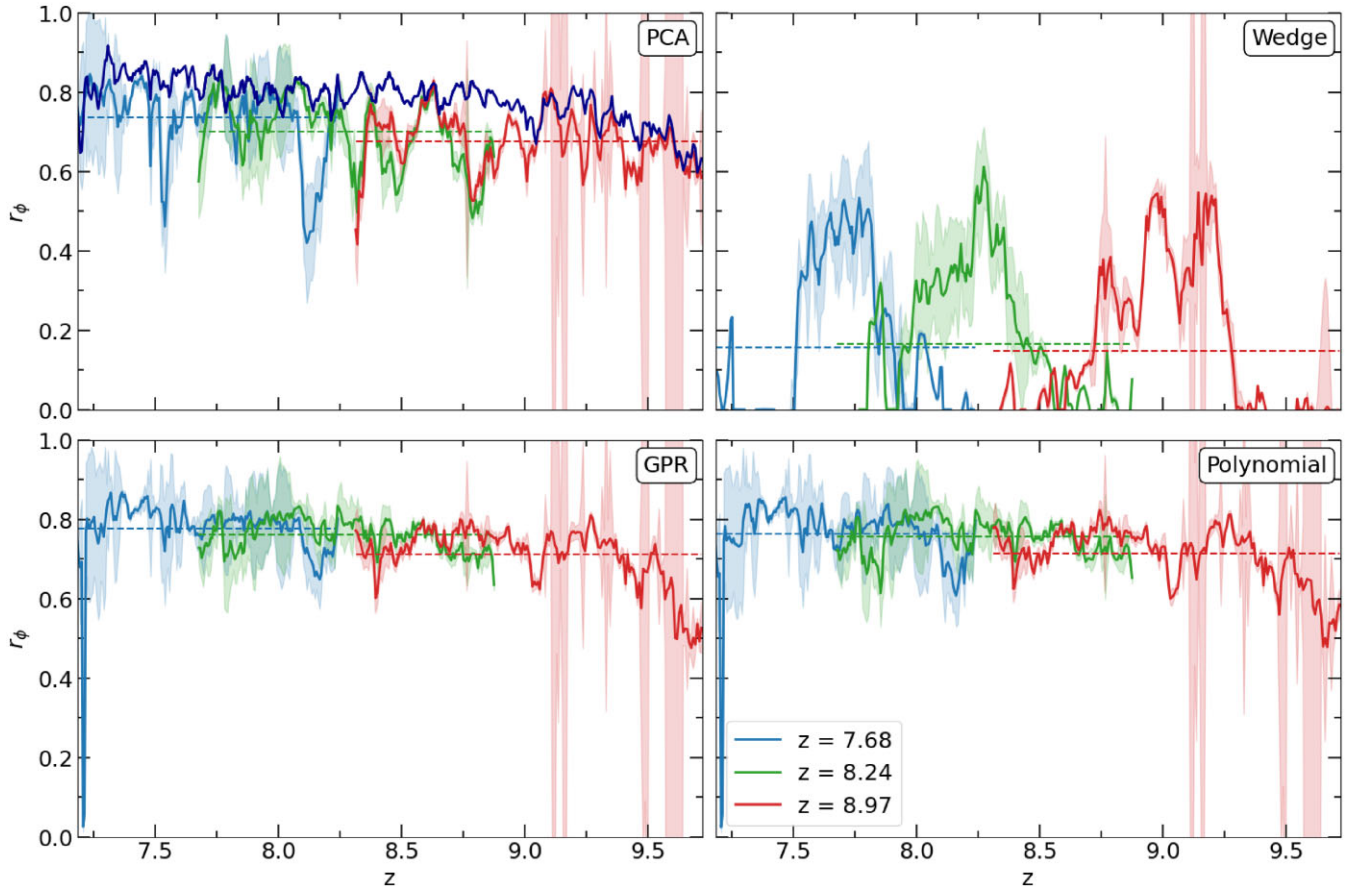

Figure 7. Redshift evolution of the r_ϕ correlation coefficient for the different tested pre-processing step. Each panel shows the result on three lightcone subvolumes centred at $z_c = 7.68$ (blue), 8.24 (green) and 8.97 (red) with a ± 10 MHz frequency depth. These redshifts correspond to the late, middle and early stages of reionization, respectively. Solid lines indicate the r_ϕ coefficient for the predicted binary maps. Shadow areas indicate the error due to the uncertainty map. Horizontal dashed lines indicate the redshift averaged $\overline{r}_\phi$ coefficient. For the case of PCA, we plot the decomposition executed on the full redshift range (dark blue) as a reference.

PCA with $\overline{r}_\phi = 0.68, 0.67$, and 0.62, respectively. Polynomial fitting follows with $\overline{r}_\phi = 0.65, 0.62$, and 0.60, while wedge remove follows with $\overline{r}_\phi = 0.18, 0.19$, and 0.15, respectively. An important remark: in this comparison, we limit the PCA decomposition to the subvolumes redshift bins (172, 181, and 186), and it is performing slightly worse when compared to the same results in the previous section on the 552 redshift bins. Therefore, we attribute the performance decrease to the reduced number of redshift bins that directly lower the number of orthogonal components with which the data are represented. For the case of PCA in Fig. 7, we plot on the same panel the performance of the PCA decomposition on the 552 redshifts (dark blue line). Here, we can notice how the redshift averaged correlation coefficient is substantially higher, $\overline{r}_\phi = 0.82$ at $z_C = 7.68$, 0.80 at $z_C = 8.24$, and 0.76 at $z_C = 8.97$, hence indicating that the PCA pre-process is preferred if we have at our disposal a tomographic data set with an extended redshift range. The sharp increase at $z \simeq 8.76$, the sudden increase at $z \geq 9$ and the constant broadening for $z \leq 8.1$ of the uncertainty error in Fig. 7 indicates that the PCA, GPR and Polynomial fitting are sensible to the evolution and distinctiveness of the same structures in the data.

Moreover, all processes, except for PCA, show a slight decrease in accuracy close to the redshift extremities values of the subvolume. The wedge removal efficiently helps recover the binary maps only for the selected subvolume central part, close to the central redshift. While the accuracy decreases rapidly toward the edges as the foreground removal becomes inefficient, in our simplified version of the wedge removal code, we do not include the sliding trough process (see Section 3.2). Therefore, a comparison between the wedge removal and the other pre-processing should be strictly limited to the subvolume central part.

5.2.3 *Recovered neutral island size distribution*

In Fig. 8, we compare the neutral island size distribution (ISD) derived from the H I binary field predicted with the different pre-processing methods presented in Section 3. We employ the mean-free path (MFP; Mesinger & Furlanetto 2007) method to derive the probability density distribution ($R\mathrm{d}P/\mathrm{d}R$) of the neutral region sizes or radius R. This size distribution measures the topological evolution of the reionization process (Friedrich et al. 2011; Giri et al. 2018a). See Giri et al. (2019) for a detailed study of ISDs during reionization.

In Fig. 8, each panel shows the predicted ISD (solid line) for three subvolumes centred at redshift $z_c = 7.68$ (blue), 8.24 (green) and 8.97 (red) against the ground truth ISD (dashed line). In the bottom part of each panel, we show the difference with the ground truth. Similarly to before, in the case of PCA, the estimated distribution with PCA decomposition on the full redshift range, from 7 to 11, is shown with a darker colour. We show the uncertainty error on the predicted ISD with a shadow area of the same colour. The GPR method and the polynomial fitting from neutral island distribution analysis appear to be the best fit. Differences are visible only at a large scale, $R \geq 100$ Mpc, with a factor $\sim$3 larger for the early and middle reionization subvolume stage. The only noticeable difference for the early stage subvolume is for the extremely large sizes, $R \approx 300$ Mpc. The results from the training pre-processing (darker colour) predict an ISD consistently shifted toward a larger scale for the case of $z_c = 7.68$ and 8.24. Deviations from the ground truth start to be visible for scale $R \geq 40$ Mpc and $R \geq 80$ Mpc with differences from up to a factor of $\sim$2 and a maximum of 5 at $R \approx 200$ Mpc. On the other hand, for the case of the subvolume centred at $z_c = 8.97$, the predicted ISD shows no virtual difference. These results confirm what we concluded in Section 5.1, with the analysis from Fig. 5 (left-hand panel). The PCA performed on the subvolume redshift range shows the same factorial difference but with an opposite behaviour. Differences are more prominent for the late stage of reionization subvolume and get gradually better at the early stage. In this analysis, the Wedge method fails to depict the HI distribution for all the subvolumes. For small neutral regions, $R \leq 20$ Mpc, the predicted distribution is a factor of 2 larger, while for larger sizes, the distribution can be severely underestimated, with $R\mathrm{d}P/\mathrm{d}R$ two orders of magnitude smaller than the ground truth distribution. This performance is an indication that with the Wedge pre-processing, `SegU-Net v2` is struggling to connect large neutral regions due to the missing 21-cm signal lying in the *foreground wedge* region that has been removed along with the foreground.

From the probability density distribution $R\mathrm{d}P/\mathrm{d}R$, we can estimate the mean radius of the neutral islands at a given redshift, defined as

$$\overline{R}_C(z) = \int_{R_{\rm min}}^{\infty} R\,\frac{\mathrm{d}P}{\mathrm{d}R}(z)\,\mathrm{d}R. \tag{16}$$

In our case, we set the lower limit to the intrinsic resolution of our simulation $R_{\rm min} = 2$ cMpc. In Table 2, rightmost column, we list this quantity derived from the predicted binary field with the different pre-process. The ground truth average radius is $\overline{R}_C = 19.89$ cMpc for the subvolume centred at $z_c = 7.68$, $\overline{R}_C = 29.54$ cMpc for $z_c = 8.24$ and $\overline{R}_C = 49.09$ cMpc for $z_c = 8.97$. Based on this quantity, we notice that the GPR method and Polynomial fitting produce a better prediction for the late and middle EoR subvolumes, with a difference to the ground truth below the cMpc, while for the early stage scenario, they tend to underestimate of a few cMpc. In the case of both PCA decompositions, the predicted quantity differs by a few cMpc in excess and deficit, respectively. This trend is also visible from the predicted ISD, as PCA shows a systematic underestimation, while the same decomposition on the entire redshift range shows an overestimation for the same scale, $R \geq 30$ cMpc. Considering the uncertainty, the wedge method seems to work reasonably well only for the late stage of reionization. However, for this scenario, the predicted ISD does not match. At late stages, the Wedge Removal prediction of $\overline{R}_C$ cannot be trusted, as this quantity differs substantially.

5.3 Relation between ionized volume and total ionizing photons

Zackrisson et al. (2020) illustrated the possibility of employing SKA-Low tomographic data as a foreshadowing method to identify the region of interest for future and ongoing experiments that aim to observe galaxy formation in the early Universe, such as the *JWST*, Euclid, and Nancy Grace Roman Space Telescope (e.g. Beardsley et al. 2015; Geil et al. 2017). This work demonstrated that there is a simple relation between the volume of isolated H II bubbles, $\mathrm{V_{ion}}$, and the grand total of ionizing photons, $\mathrm{N_{\gamma,\,tot}}$, produced by the primordial sources within the same ionized region. Although we are overlooking relevant instrumental effects (e.g. incomplete uv-coverage, absence of gain error, beam effect and more), we assume that our framework, described in Section 2.4, produces realistic enough mock observation to demonstrate the challenge of identifying and measuring the sizes of such bubbles and its derived relation.

For this analysis, we require the mass and the position of the sources within the ionized bubbles. Therefore, we decided to use a simulation run with the `C`^{2}`Ray` radiative transfer code (Mellema et al. 2006). In Paper I, we demonstrated how `SegU-Net` works reasonably well on simulations other than those employed for the training and validation. Moreover, recent works demonstrated the

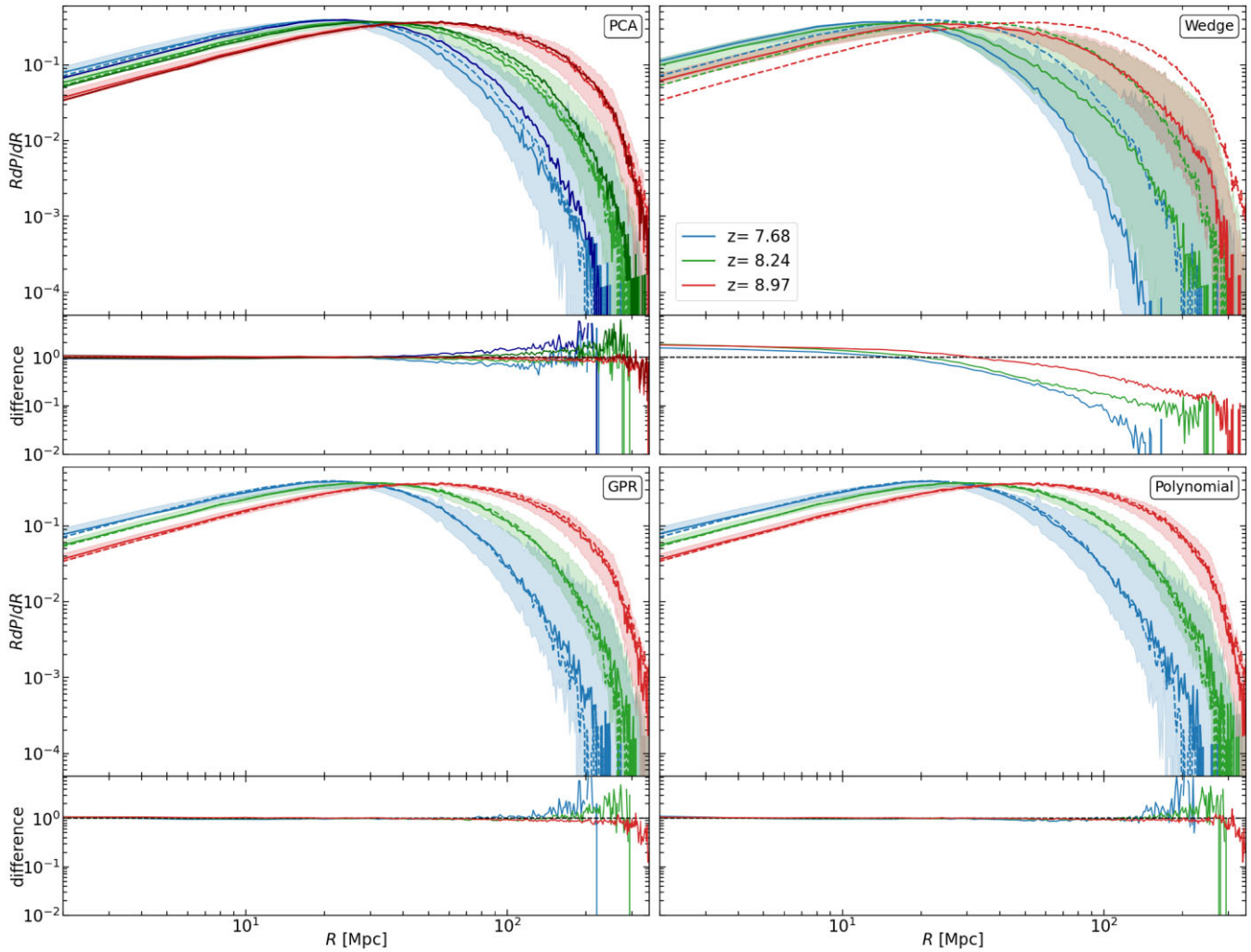

Figure 8. Island size distribution for the different pre-processing steps. Each panel shows the predicted size distribution $R\,dP/dR$ (top section) and the difference to the ground truth (bottom section). The colours indicate the lightcone subvolume at the late ($z_c = 7.68$, blue), middle ($z_c = 8.24$, green), and early ($z_c = 8.97$, red) stage of reionization. The results from the neutral regions in the predicted fields are shown with solid lines and the ground truth with dashed lines. For the case of PCA, we plot as a reference the predicted size distribution with a dot–dashed line.

limitations of U-Net when cross-validating different cosmological models (Chen et al. 2023b). Here, we employ the obtained ionized hydrogen and density coeval cubes to calculate the 21-cm differential brightness with equation (1) and follow the mock observation procedure explained in Section 2.4. We consider the third axis the frequency direction to create the corresponding network input and target. We use one realization of the simulated coeval cube at redshift z = 8.89 with box and mesh size of 348 cMpc and 250, respectively. We interpolate the 250 mesh grid into a 166 grid per side to a corresponding intrinsic resolution similar to our $\Delta x = 2.09$ Mpc data set. One of the inputs of the C^{2}Ray code is the cumulative halo mass smoothed into the mesh grid. In this way, we can associate an ionized bubble to the sources within the same region by converting the total halo distribution mass $M_{\rm h,\,tot}$ to the total ionizing photon produced $\mathrm{N}_{\gamma,\mathrm{tot}} = \mathrm{f}_\gamma\,\Omega_\mathrm{m}/\,\Omega_\mathrm{b}\,\mathrm{M}_{\mathrm{h,tot}}$. We refer the reader to Iliev et al. (2006, 2012) and Dixon et al. (2016) for further reading on the halo source model.

Though SegU-Net v2 is not trained on simulations produced with C^{2}Ray, we still find that the ionized regions are accurately identified. This analysis shows that the trained model is quite general[9] and, therefore, capable of finding physical features in real observations. In Fig. 9, we show the relation between $\mathrm{V}_{\rm ion}$ and $\mathrm{N}_{\gamma,\,\mathrm{tot}}$ derived from the simulation data (blue crosses) and the predicted binary maps (orange points). We notice that SegU-Net v2 is failing to correctly quantify the number of ionizing photons for volumes $\mathrm{V}_{\rm ion} \lesssim (10\,\mathrm{cMpc})^3$, vertical black dash line. This limitation corresponds to the 2 km interferometric smoothing scale we apply in our mock observation pipeline. At $z = 8.89$, the Gaussian kernel has an angular scale of $\Delta\theta \approx 3.57$ arcmin, corresponding to a comoving size of 9.9 cMpc. This limitation is also consistent with the results in Fig. 5, where the correlation between prediction and ground truth slowly decreases, $r_\phi \leq 80$ per cent, for higher redshift, z $\geq$ 9.

[9]We should note that we have not tested the framework on radiative transfer hydrodynamical simulations due to the unavailability of models with box lengths exceeding 200 Mpc, which is essential for studying the 21-cm signal (e.g. Giri et al. 2023).

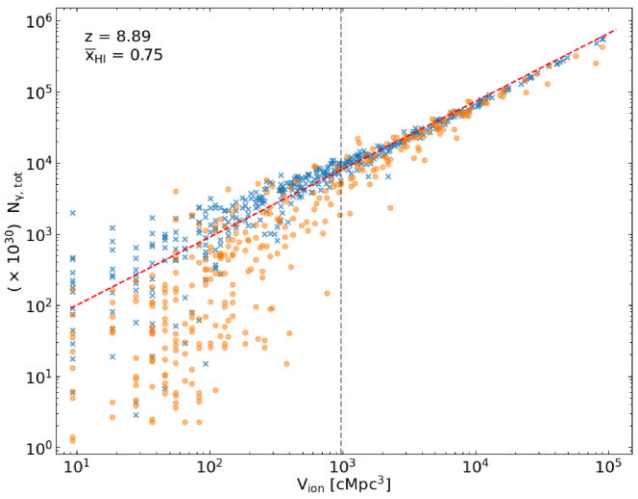

Figure 9. Relation between the volume of ionized region versus the grand total of ionizing photons within the same region. For a coeval cube at redshift $z = 9$ ($\bar{x}_{HI} = 0.75$) and box size of $L_{box} \approx 348$ cMpc. Relation derived from the ground truth is represented with blue cross data, while orange circle points are derived from `SegU-Net` prediction. The dashed red line corresponds to the linear fit of the ground truth data points. The vertical line indicates the 2 km baseline smoothed resolution.

6 DISCUSSION AND CONCLUSIONS

With this work, we improved our previous effort in Paper I and updated our deep learning framework, `SegU-Net v2`, for the identification of neutral and ionized regions in realistic 21-cm mock observation expected from SKA-Low. One of the advantages of our network is the possibility to provide per-pixel uncertainty maps on its predictions. In Section 2.4, we introduced our extended mock observation pipeline by including synchrotron Galactic foreground contamination, presented in Section 2.3. Additionally, we performed machine learning hyperparameter optimization. We show the best-performing hyperparameters set-up we analysed in Appendix A.

In this work, we combine our network with a foreground mitigation method that pre-processes the input data and reduces, in part, the foreground contribution. We trained `SegU-Net v2` on 10,000 lightcones with 552 redshift slices from $z = 7$ to 11 pre-processed with PCA on 4 components for the full redshift range. We chose this pre-processing method as it is the most commonly used method for foreground contamination and provides fast and efficient mitigation. In Section 5.1, the analysis on a random sample data set, composed of 300 lightcone with the same redshift extent and bins, shows that the updated version of our network works well, with an average correlation of 71 per cent, on 21-cm images contaminated and pre-processed by a foreground contamination method. This level of accuracy is almost ~ 20 per cent less than our previous results and is attributed to the added complexity due to the presence of the Galactic foreground. We show that `SegU-Net v2` recovered binary fields that tend to be considered more neutral at $z \leq 8.5$. We attribute this to the under-subtraction of the PCA pre-processing method employed during training. This trend is confirmed by the increase of the uncertainty map for the same redshift extent that saturates entire frequency channels (see the bottom panel in Fig. 4).

In Section 5.2, we compared the binary maps predicted with `SegU-Net v2` on different pre-processing foreground mitigation and one avoidance method. We consider three subvolume of the fiducial simulation with frequency width $\Delta\nu = \pm 10$ MHz centred at redshift $z_c = 7.68$, 8.24 and 8.97, representing a late, middle and early stage of reionization. In this work, we consider PCA decomposition (Section 3.1), Wedge removal (Section 3.2), Gaussian Process Regression (Section 3.3) and Polynomial fitting (Section 3.4). We demonstrated that `SegU-Net v2` is able to recover H I regions with varying accuracy for all the pre-processing methods we tested. In our case, the network is able to generalize enough and work with the same level of accuracy as the training case on pre-processing methods that were not employed during its training (see summary statistics in Table 2). Moreover, in Section 5.2.3, we study the ISD of the predicted binary maps. GPR and Polynomial fitting work better in recovering the ISDs, as well as the average distribution size R_C of neutral regions, than the two cases of the PCA pre-processing (applied on the full redshift range and the subvolume redshift range).

Therefore, we can conclude that `SegU-Net v2` is the pre-processing method agnostic, providing accurate predictions independent of the pre-processing method, as long as the foreground mitigation provides reasonable residual images of the original 21-cm signal. Another conclusion is that PCA decomposition on lightcone data with a wide redshift range, e.g. frequency depth of the order of 60 MHz or larger, is to be preferred. In the case of smaller available subvolumes, with frequency depth between 20 MHz and 30 MHz, other methods such as GPR or Polynomial fitting are to be preferred as they provide better prediction when compared to PCA on the same redshift range.

Finally, we provided a concrete use case of `SegU-Net v2` in the context of 21-cm SKA-Low tomographic observation. Previous work demonstrated that a linear relation could be derived between the size of the ionized volume and the grand total number of ionizing photons produced by the hosted source. In Section 5.3, we demonstrated that our network could recover with precision the linear relation for ionized volumes that are resolved. Here, we stipulate the limited resolution of the SKA-Low layout by the interferometric smoothing scale for the maximum baseline of $B = 2$ km, which corresponds to an angular scale of approximately 3.57 arcmin at redshift $z = 8.89$, corresponding to an early stage of reionization scenario, $\bar{x}_{HI} = 0.75$.

The current version of `SegU-Net v2` is trained using semi-numerical simulations, known for their non-conservation of photons (e.g. Choudhury & Paranjape 2018; Hutter 2018). This discrepancy arises when the number of photons emitted by the sources does not match the number of IGM ionizations. However, it is important to highlight that `SegU-Net v2` does not exhibit sensitivity to the model linking the sources and sinks in the simulations. Instead, it learns the ionization patterns present in the 21-cm signal distribution. Consequently, the framework successfully predicts the accurate volume of ionized regions in simulations generated by `C2Ray`, a numerical simulation code that conserves photons (Section 5.3). In future work, we plan to retrain the network on models from photon-conserving frameworks, such as `Beorn` (Schaeffer, Giri & Schneider 2023) and `pyC2Ray` (Hirling et al. 2023).

In this paper, `SegU-Net` was trained on one NVIDIA® Tesla® P100 with 16GB for a total computational cost of approximately 12 GPU hours. When comparing the pre-processing method, we also consider the computational time required to compute the foreground mitigation/avoidance method. In our set-up, one lightcone subvolume of frequency depth 20 MHz with 200 redshift bins takes about 7 s CPU time to compute with PCA and 2 s with Polynomial fitting. Wedge remover provides faster pre-processing with 230 ms but inefficient foreground mitigation. On the other hand, GPR provides slow but reliable mitigation with a computing time of $\sim$1.2 CPU hours.

The Grad-CAM importance score analysis conducted in Appendix B shows that the network decoder convolutional layer starts by identifying and grouping the region with the strongest positive emission. In the bottleneck of the U-Net model, the low-dimensional latent space then uses the encoded information to identify the threshold that defines the boundary of the neutral regions. The decoder layers use the compressed information and the U-Net skip connection with the encoder layer to define the location of the borders. Finally, the last convolutional layer further refines the decoder output. However, the analysis showed that the network struggled to correctly identify the residual foreground when this signal is similar to the 21-cm intensity. This explains why the final predictions include a positive detection of 21-cm signal regions and a false negative due to the noise or foreground residuals.

In our case, `SegU-Net` is a deterministic deep learning model. Recently, a series of works have imported probabilistic models in radio astronomy and astrophysics (Friedman & Hassan 2022; Sortino et al. 2023; Wang et al. 2023). This approach inherently handles noise and variability in the data compared to the deterministic case. At the same time, they can learn the underlying probability distribution of the data, which can help for a better interpretation. On the other hand, deterministic models like U-Net often have the advantage of being computationally efficient and easier to train. In future work with `SegU-Net`, we consider converting the model to be probabilistic.

Our analysis shows that using image data from SKA-Low, `SegU-Net v2` accurately determines the ionization fraction at different stages of reionization. Additionally, we have identified how the ionized regions detected by `SegU-Net v2` can be used as markers for locating the galaxies responsible for driving the reionization process. These findings demonstrate the potential of our framework for synergy studies with other telescopes, such as the *JWST*, Euclid, and Nancy Grace Roman Space Telescope.

ACKNOWLEDGEMENTS

The authors would like to thank Bharat Kumar Geholt for his useful discussions and comments. MB acknowledges the financial support from the Swiss National Science Foundation (SNSF) under the Sinergia Astrosignals grant (CRSII5_193826). We acknowledge access to Piz Daint at the Swiss National Supercomputing Centre, Switzerland, under the SKA's share with the project ID sk09. This work has been done in partnership with the SKACH consortium through funding by SERI. Nordita is supported in part by NordForsk.

The deep learning implementation was possible thanks to the application programming interface of `Tensorflow` (Abadi et al. 2015) and `Keras` (Chollet, Allaire et al. 2017). The algorithms and image processing tools operated on our data were performed with the help of NUMPY (Harris et al. 2020), SCIPY (Virtanen et al. 2020), SCIKIT-LEARN (Pedregosa et al. 2011), and SCIKIT-IMAGE (van der Walt et al. 2014) packages. All figures were created with MATHPLOTLIB (Hunter 2007).

DATA AVAILABILITY

The data underlying this article is available upon request and can also be re-generated from scratch using the publicly available `21cmFAST` (Mesinger et al. 2011), `CUBEP`^{3}`M` (Harnois-Déraps et al. 2013), `C`^{2}`RAY` (Mellema et al. 2006), and `Tools21cm` (Giri, Mellema & Jensen 2020) code. The `SegU-Net` code and its trained network weights are available on the author's `GitHub` page: https://github.com/micbia/SegU-Net.

REFERENCES

Abadi M. et al., 2015, TensorFlow. Available at: http://tensorflow.org/
Abel T., Bryan G. L., Norman M. L., 2001, Science, 295, 93,
Achanta R., Shaji A., Smith K., Lucchi A., Fua P., Süsstrunk S., 2012, IEEE Trans. Pattern Anal. Machine Intell., 34, 2274
Akiba T., Sano S., Yanase T., Ohta T., Koyama M., 2019, KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Association for Computing Machinery, New York, NY, p. 2623
Alonso D., Bull P., Ferreira P. G., Santos M. G., 2015, MNRAS, 447, 400
Bakx T. J. L. C. et al., 2023, MNRAS, 519, 5076
Beardsley A. P., Morales M. F., Lidz A., Malloy M., Sutter P. M., 2015, ApJ, 800, 128
Bianco M., Giri S. K., Iliev I. T., Mellema G., 2021, MNRAS, 505, 3982 (Paper I)
Bonaldi A., Brown M. L., 2015, MNRAS, 447, 1973
Bowman J. D., Morales M. F., Hewitt J. N., 2009, ApJ, 695, 183
Boylan-Kolchin M., 2023, Nat. Astron., 7, 731
Briggs D. S., 1995, PhD thesis, New Mexico Institute of Mining and Technology
Bromm V., Yoshida N., Hernquist L., McKee C. F., 2009, Nature, 459, 49
Castellano M. et al., 2022, ApJL, 938, L15
Chapman E., Jelić V., 2019, preprint (arXiv:1909.12369)
Chapman E. et al., 2012, MNRAS, 423, 2518
Chapman E. et al., 2013, MNRAS, 429, 165
Chen Z., Chapman E., Wolz L., Mazumder A., 2023a, MNRAS, 524, 3724
Chen T., Bianco M., Tolley E., Spinelli M., Forero-Sanchez D., Kneib J. P., 2023b, preprint (arXiv:2311.00493)
Chollet F. et al., 2017, Keras. Available at: https://github.com/rstudio/keras
Choudhuri S., Bharadwaj S., Ghosh A., Ali S. S., 2014, MNRAS, 445, 4351
Choudhury T. R., 2022, Gen. Rel. Grav., 54, 102
Choudhury T. R., Paranjape A., 2018, MNRAS, 481, 3821
Cunnington S. et al., 2023, MNRAS, 518, 6262
Datta K. K., Bharadwaj S., Choudhury T. R., 2007, MNRAS, 382, 809
Dayal P., Giri S. K., 2023, preprint (arXiv:2303.14239)
Di Matteo T., Perna R., Abel T., Rees M. J., 2002, ApJ, 564, 576
Di Matteo T., Ciardi B., Miniati F., 2004, MNRAS, 355, 1053
Dillon J. S. et al., 2014, PRD, 89, 23002
Dixon K. L., Iliev I. T., Mellema G., Ahn K., Shapiro P. R., 2016, MNRAS, 456, 3011
Elbers W., van de Weygaert R., 2023, MNRAS, 520, 2709
Ferrara A., Pandolfi S., 2014, Proc. Int. Sch. Phys. Fermi, 186, 1
Friedman R., Hassan S., 2022, preprint (arXiv:2211.12724)
Friedrich M. M., Mellema G., Alvarez M. A., Shapiro P. R., Iliev I. T., 2011, MNRAS, 413, 1353
Furlanetto S. R., Zaldarriaga M., Hernquist L., 2004, ApJ, 613, 1
Furlanetto S. R., Oh S. P., Briggs F. H., 2006, Phys. Rep., 433, 181
Gagnon-Hartman S., Cui Y., Liu A., Ravanbakhsh S., 2021, MNRAS, 504, 4716
Gazagnes S., Koopmans L. V. E., Wilkinson M. H. F., 2021, MNRAS, 502, 1816
Geil P. M., Mutch S. J., Poole G. B., Duffy A. R., Mesinger A., Wyithe J. S. B., 2017, MNRAS, 472, 1324
Ghara R., Choudhury T. R., 2020, MNRAS, 496, 739
Ghara R., Choudhury T. R., Datta K. K., Choudhuri S., 2017, MNRAS, 464, 2234
Ghara R. et al., 2020, MNRAS, 493, 4728
Ghara R., Giri S. K., Ciardi B., Mellema G., Zaroubi S., 2021, MNRAS, 503, 4551
Giri S. K., Mellema G., 2021, MNRAS, 505, 1863
Giri S. K., Mellema G., Dixon K. L., Iliev I. T., 2018a, MNRAS, 473, 2949
Giri S. K., Mellema G., Ghara R., 2018b, MNRAS, 479, 5596
Giri S. K., Mellema G., Aldheimer T., Dixon K. L., Iliev I. T., 2019, MNRAS, 489, 1590
Giri S. K., Mellema G., Jensen H., 2020, J. Open Source Softw., 5, 2363
Giri S. K., Schneider A., Maion F., Angulo R. E., 2023, A&A, 669, A6
Gleser L., Nusser A., Benson A. J., 2008, MNRAS, 391, 383

Gu J., Xu H., Wang J., An T., Chen W., 2013, ApJ, 773, 38
Harker G. et al., 2009, MNRAS, 397, 1138
Harnois-Déraps J., Pen U.-L., Iliev I. T., Merz H., Emberson J. D., Desjacques V., 2013, MNRAS, 436, 540
Harris C. R. et al., 2020, Nature, 585, 357
Hirling P., Bianco M., Giri S. K., Iliev I. T., Mellema G., Kneib J.-P., 2023, preprint (arXiv:2311.01492)
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Hütsi G., Raidal M., Urrutia J., Vaskonen V., Veermäe H., 2023, Phys. Rev. D, 107, 043502
Hutter A., 2018, MNRAS, 477, 1549
Iliev I. T., Mellema G., Pen U. L., Merz H., Shapiro P. R., Alvarez M. A., 2006, MNRAS, 369, 1625
Iliev I. T., Mellema G., Shapiro P. R., Pen U.-L., Mao Y., Koda J., Ahn K., 2012, MNRAS, 423, 2222
Jelić V. et al., 2008, MNRAS, 389, 1319
Kapahtia A., Chingangbam P., Appleby S., 2019, J. Cosmol. Astropart. Phys., 2019, 053
Kapahtia A., Chingangbam P., Ghara R., Appleby S., Choudhury T. R., 2021, J. Cosmol. Astropart. Phys., 2021, 026
Kerrigan J. R. et al., 2018, ApJ, 864, 131
Kingma D. P., Ba J., 2014, preprint (arXiv:1412.6980)
Komatsu E. et al., 2009, ApJ, 180, 330
Koopmans L. V. E. et al., 2015, Proc. Sci. The Cosmic Dawn and Epoch of Reionization withthe Square Kilometre Array. SISSA, Trieste, PoS(AASKA14)001
Li X., Chen S., Hu X., Yang J., 2018, preprint (arXiv:1801.05134)
Liu A., Shaw J. R., 2020, Publ. Astron. Soc. Pac., 132, 062001
Liu A., Tegmark M., Zaldarriaga M., 2009a, MNRAS, 394, 1575
Liu A., Tegmark M., Bowman J., Hewitt J., Zaldarriaga M., 2009b, MNRAS, 398, 401
Liu A., Parsons A. R., Trott C. M., 2014, Phys. Rev. D, 90, 023019
Madau P., Meiksin A., Rees M. J., 1997, ApJ, 475, 429
Mehra J., Neeru N., 2016, Imperial J. Int. Res., 03, 8
Mellema G., Iliev I. T., Alvarez M. A., Shapiro P. R., 2006, New Astron., 11, 374
Mellema G. et al., 2013, Exp. Astron., 36, 235
Mellema G., Koopmans L., Shukla H., Datta K. K., Mesinger A., Majumdar S., 2015, Proc. Sci., HI tomographic imaging of the Cosmic Dawn andE-poch of Reionization with SKA. SISSA, Trieste, PoS(AASKA14)010
Mertens F. G., Ghosh A., Koopmans L. V. E., 2018, MNRAS, 478, 3640
Mertens F. G. et al., 2020, MNRAS, 493, 1662
Mesinger A., Furlanetto S., 2007, ApJ, 669, 663
Mesinger A., Furlanetto S., Cen R., 2011, MNRAS, 411, 955
Morales M. F., Bowman J. D., Cappallo R., Hewitt J. N., Lonsdale C. J., 2006a, New Astron. Rev., 50, 173
Morales M. F., Bowman J. D., Hewitt J. N., 2006b, ApJ, 648, 767
Murray S. G., Trott C. M., 2018, ApJ, 869, 25
Murray S. G., Greig B., Mesinger A., Muñoz J. B., Qin Y., Park J., Watkinson C. A., 2020, J. Open Source Soft., 5, 2582
Naidu R. P. et al., 2022, ApJ, 940, L14
Offringa A. R. et al., 2014, MNRAS, 444, 606
Pawlik A. H., Milosavljević M., Bromm V., 2011, ApJ, 731, 54
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Perez L., Wang J., 2017, preprint (arXiv:1712.04621)
Planck Collaboration VI, 2020, A&A, 641, A6
Platania P., Bensadoun M., Bersanelli M., De Amici G., Kogut A., Levin S., Maino D., Smoot G. F., 1998, ApJ, 505, 473
Pober J. C. et al., 2014, ApJ, 782, 66
Prelogović D., Mesinger A., Murray S., Fiameni G., Gillet N., 2021, MNRAS, 509, 3852
Pritchard J. R., Furlanetto S. R., 2007, MNRAS, 376, 1680
Ronneberger O., Fischer P., Brox T., 2015, preprint (arXiv:1505.04597)
Ross H. E., Dixon K. L., Ilie'v I. T., Mellema G., 2017, MNRAS, 468, 3785
Ross H. E., Dixon K. L., Ghara R., Iliev I. T., Mellema G., 2019, MNRAS, 487, 1101
Ross H. E., Giri S. K., Mellema G., Dixon K. L., Ghara R., Iliev I. T., 2021, MNRAS, 506, 3717
Rumelhart D. E., Zipser D., 1985, Cognitive Sci., 9, 75
Salehi S. S. M., Erdogmus D., Gholipour A., 2017, preprint (arXiv:1706.05721)
Santos M. G., Cooray A., Knox L., 2005, ApJ, 625, 575
Schaeffer T., Giri S. K., Schneider A., 2023, MNRAS, 526, 2942
Schneider A., Schaeffer T., Giri S. K., 2023, Phys. Rev. D, 108, 043030
Selvaraju R. R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D., 2019, Int. J. Comp. Vision, 128, 336
Smirnov O. M., 2011, A&A, 527, A106
Sortino R. et al., 2023, preprint (arXiv:2307.02392)
The HERA Collaboration, 2022a, ApJ, 924, 51
The HERA Collaboration, 2022b, ApJ, 925, 221
Thyagarajan N. et al., 2015, ApJ, 804, 14
Trott C. M. et al., 2020, MNRAS, 493, 4711
Virtanen P. et al., 2020, Nature Methods, 17, 261
van der Walt S. et al., 2014, PeerJ, 2, e453
Wang X., Tegmark M., Santos M. G., Knox L., 2006, ApJ, 650, 529
Wang J. et al., 2013, ApJ, 763, 90
Wang Y., Huang G., Song S., Pan X., Xia Y., Wu C., 2020, preprint (arXiv:2007.10538)
Wang R., Chen Z., Luo Q., Wang F., 2023, preprint (arXiv:2305.09121)
Wyithe S., Geil P. M., Kim H., 2015, Proc. Sci., Imaging HII Regions from Galaxies and QuasarsDuring Reionisation with SKA. SISSA, Trieste, PoS(AASKA14)015
Zackrisson E. et al., 2020, MNRAS, 493, 855
Zaroubi S., 2012, in Wiklind T., Mobasher B., Bromm V., eds, Astrophysics and Space Science Library, Vol 396, The Epoch of Reionization. Springer, Berlin, Heidelberg

APPENDIX A: HYPERPARAMETER EXPLORATION

As we mentioned in Section 4, we perform an optimization analysis of the `SegU-Net` hyperparameters. We are aware of tools that automatize the exploratory analysis of the network hyperparameter space, such as `Optuna` (Akiba et al. 2019). However, constrained by time and computational resources, we manually searched the best-performing parameters through a trial-and-error approach. First, we selected a few combinations of the network parameters and performed a short training of no more than five to ten epochs. Based on the result obtained in this short training, we selected six of the best-performing results with the lowest validation loss and performed a full training to identify the ideal hyperparameters set-up. In future work, we intend to undertake a more comprehensive study.

We list the six best-performing set-ups we tested in Table A1. We include an analysis of seven model parameters, from left to right: the activation function of the convolutional layers, the number of channels for the bottom layer, the number of pooling operations, the dropout rate, the final activation function before the binary operation, the size of the kernel filters and the type of pooling operation. As a loss function, we employed the balanced cross-entropy (BCE) function (Salehi, Erdogmus & Gholipour 2017) and the Adaptive Moment Estimator (Adam) (Kingma & Ba 2014) as the stochastic gradient descent algorithm to minimize the loss. We employed the in bold text for the results presented in this paper. Although we monitored the validation loss to select the best set-up, we noticed that the first and second models gave the worst prediction for images at the edges of the redshift range $z \sim 7$ and 11. The fifth model provided the most balanced result, with an overall $r_\phi \approx 0.7$ score, as shown in Fig. 5 central panel. Moreover, in contrast to the findings by Li et al. (2018), we observe that setting the dropout rate to zero enhances accuracy only for the third-ranked set-up. Meanwhile, other configurations exhibit improved performance when both batch normalization and dropout are included.

Table A1. `SegU-Net` hyperparameter optimization analysis for the best-performing set-ups of seven parameters with optimization on the validation loss.

Ranking	Activation	Channels latent space	Depth	Dropout	Final activation	Kernel	Pooling type	r_ϕ (per cent)	Validation loss ($\times 10^{-2}$)
1	`LeakyReLu`	256	3	0.42	$\sigma(x)$	6	max	89.08	6.59
2	`LeakyReLu`	128	4	0.00	$\sigma(x)$	5	max	89.02	6.62
3	`Elu`	128	3	0.34	x	11	average	87.76	7.27
4	`LeakyReLu`	128	4	0.50	$\sigma(x)$	5	max	88.72	6.85
5	**`ReLu`**	**256**	**4**	**0.05**	$\boldsymbol{x}$	**7**	**average**	**88.50**	**7.48**
6	`ReLu`	256	5	0.14	x	7	max	86.53	9.15

APPENDIX B: INSIDE THE BLACK BOX

The trained model we presented in this paper is able to recover the ionized field from noisy images with residual foreground contamination. This is an indication that the network learns to identify the regions of interest from important hidden features that maximize the recovery. However, the machine learning model's complexity, high dimensionality and non-linearity make them difficult to interpret and regulate, so these applications are often referred to as a *black box*. Here, we present a first attempt to open and look inside `SegU-Net` black box. A standard tool to visualize and understand the decisions made by a general convolutional neural network is the Gradient-weighted Class Activation Mapping (Grad-CAM) technique (Selvaraju et al. 2019). This method applied in segmentation and object classification highlights the features of an input image employed in predicting a particular class. Grad-CAM achieves this by computing the gradients of the predicted class $y_c \equiv x^B_{\rm HI}$ score with respect to the feature maps F_k of all the $k > 0$ convolutional layers up to the layer in the network under study. A weighted combination of these feature maps gives the *importance score* $M_c^{(i,j)} \in [0, 1]$ that indicates the importance of the feature in the image at location (i, j) for the class cth. This score is given as

$$M_c^{(i,j)} = \frac{1}{Z} \sum_k w_k \frac{\partial y_c}{\partial F_k^{(i,j)}}, \tag{B1}$$

where $Z = \sum_k w_k$ is the normalization factor, corresponding to the sum of the kth weights, while w_k is the weight corresponding to the feature map F_k. In our case, we focus on the neutral regions, categorized with a value of $c = 1$ in our maps. Values close to one indicate high importance, while in the opposite case, it indicates irrelevance.

In Fig. B1, we show the result for three hidden layers. The first column shows the input image and the M_c score represented by dark shadows, indicating the location in the image employed in the classification of the neutral regions. Solid line contours indicate the ground truth. The central column shows the Grad-CAM filtered region obtained by element-wise multiplication of the input image with the importance score. This plot shows us what features the network emphasizes in the image for identifying the neutral region. The right column visualizes the hidden layer output, with the number of subpanels corresponding to the number of channels. From top to bottom, we have the output of the convolution block at the second level of the encoder after two convolutional layers and a pooling operation. The hidden state has angular and channel dimensions (64, 64, 32). We can see that in the encoder, the network focuses on the regions with the highest intensity, which, thanks to the pre-process presented in Section 3, are mostly located within the neutral region. The different channels in the hidden layer show a similar conclusion, with the convolutional operation capturing the large-scale region that produces 21-cm signals. In the second row, the bottom of the U-Net, known as the low-dimensional latent space, with dimension (16,16,128), gives a compressed representation of the input image, and it appears to focus on location in the image with the highest and lowest values. Our interpretation is that the network focuses on these extreme values to quantify the 'threshold' value that sets the boundary between neutral and ionized regions. This interpretation is also supported by the 128 hidden layer plots in the right-hand panel, as the compressed data shows different constant values across the channels. In the last row, we show the importance score from the final convolution before the binarization of the output. Here, it appears that the network uses the threshold value defined in the bottom layer of the U-Net to locate the delimitation that defines the neutral regions, as the shadow is located along the contour of the ground truth (black solid line). We notice that some locations in the image with substantial foreground residuals are wrongly included. The hidden layer plot shows that the network struggles to remove the foreground residual completely.

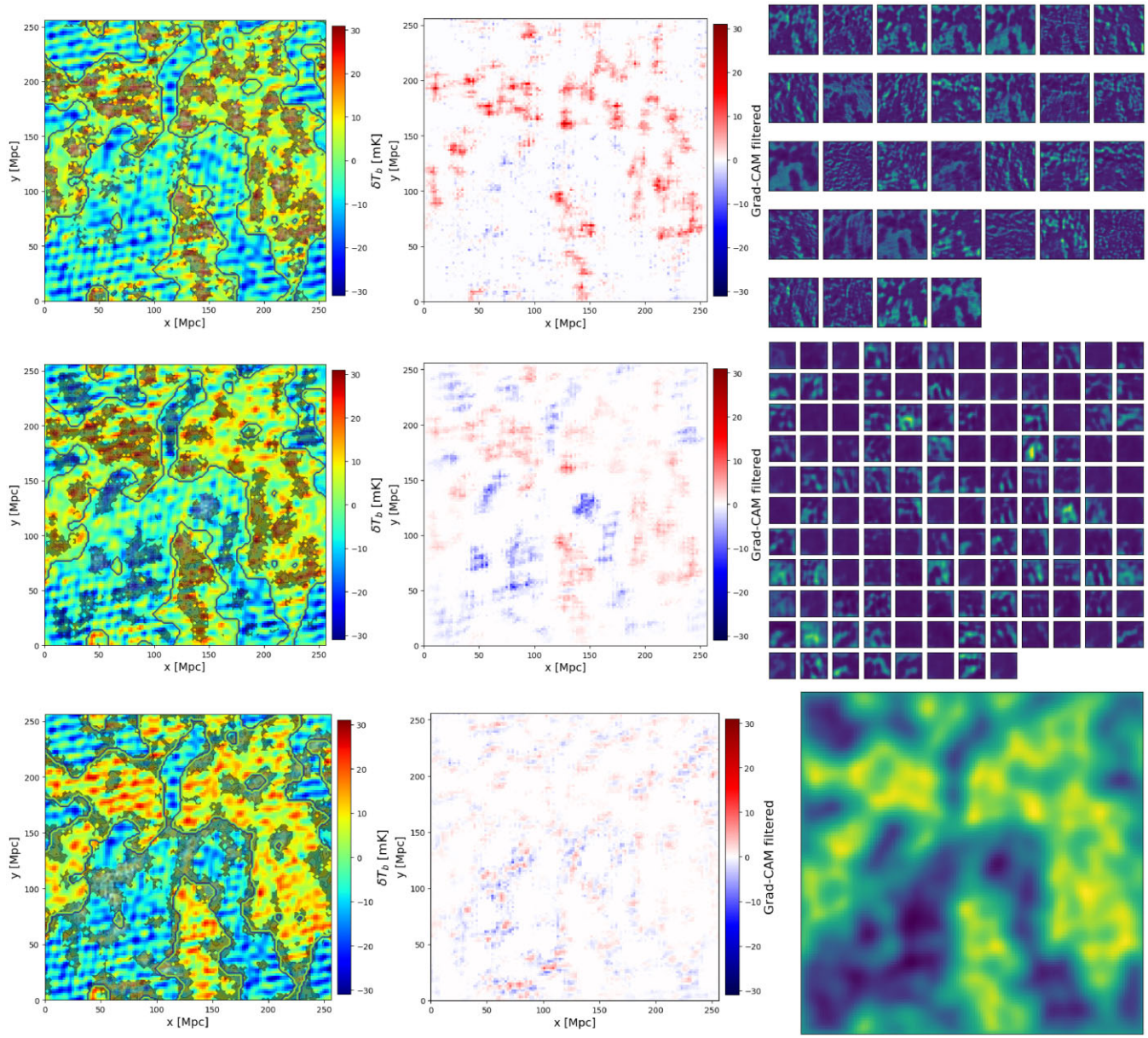

Figure B1. Region of interest detected by Grad-CAM for three hidden layers. *Left-hand panels*: Input image with shadow areas that indicate the region of attention detected by the Grad-CAM method. Black solid contours indicate the ground truth for comparison. *Central panel*: The filtered Grad-CAM image element-wise multiplication between the input and the M_c filter. *Right-hand panel*: A visualization of the hidden layer output. The number of subpanels indicates the channel size of the hidden layer.

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