URSI GASS 2023, Sapporo, Japan, 19 – 26 August 2023

SERENEt: Deep learning approach for identification of HII regions and 21-cm signal recovery from SKA-Low reionization observations

Michele Bianco*(1), Sambit K. Giri(1)(2), David Prelogović(4), Tianyue Chen(1), Emma Tolley(1), Andrei Mesinger(4), and Jean-Paul Kneib(1)

(1) Institute of Physics, Laboratory of Astrophysics, École Polytechnique Fédérale de Lausanne (EPFL), 1290 Sauverny, CH; e-mail: michele.bianco@epfl.ch

(2) Institute for Computational Science, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, CH

(3) Nordita, KTH Royal Institute of Technology & Stockholm University, Hannes Alfvens väg 12, SE-106 91 Stockholm, Sweden

(4) Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy

Abstract

The upcoming Square Kilometre Array (SKA-Low) will map the distribution of neutral hydrogen during reionization and produce a tremendous amount of 3D tomographic data. The biggest challenge for the observational analysis of these images is to separate the 21-cm signal from the undesired foreground and instrumental noise contaminations. Here we present SERENEt. A deep learning approach that works on SKA mock observation with an observation time of 1000 h and in the presence of the Galactic synchrotron foreground. We use a PCA and functional PCA (fPCA) preprocess with BlueBild code to reduce the dynamic range in foreground contaminated 21-cm image and show that our network identifies regions of neutral hydrogen (H I) and recovers the reionisation 21-cm signal from those same regions identified as neutral. We show as our approach can identify neutral regions during reionization with more than 87 per cent accuracy and recover the 21-cm 2D power spectra with an average of 95 per cent accuracy.

1 Introduction

The Epoch of reionization (EoR) is a period of great importance in studying structure formation and evolution in the Universe. During this period, the predominately cold and neutral intergalactic medium (IGM) transitioned to a hot and ionized state due to the appearance of the first luminous cosmic sources. These sources, which may have been star-forming galaxies and quasi-stellar objects (QSOs), produced the ionizing photons, which over a period of approximately 500 million years completed the reionization of the Universe [1, 2, 3]. When observed, the 21-cm signal would have redshifted to the electromagnetic spectrum radio band. The low-frequency component of the SKA will be sensitive enough to detect the 21-cm signal produced during EoR and create images of its distribution on the sky [4, 5, 6]. However, the biggest challenge for SKA observational data analysis of these astronomical images is to separate the 21-cm signal from the undesired foreground and instrumental noise contamination, as these outshine the cosmological signal by several orders of magnitude.

2 Our Network, SERENEt

In this work, we present SERENEt (**SE**gmentation and **RE**covery **NEt**work), a novel approach for the identification of the distribution of H I regions and the recovery of the 21-cm signal from SKA-Low multi-frequency tomographic images for the Cosmic Epoch of Reionization (EoR). Our code consists of a pre-processing step for foreground mitigation and two U-shaped deep convolutional neural network (CNN) [7] for segmentation and 21-cm signal recovery, SegU-Net and RecU-Net respectively. In Figure 1, we show an overview of the SERENEt pipeline.

The SKA-Low mock observation image, I_{obs}, contains foreground contamination and systematic noise that outshine the cosmological signal by several orders of magnitude. Therefore, we pre-process I_{obs} with standard PCA foreground removal technique [8] to partially subtracts the foreground contamination. The resulting image, I_{res}, will still contain some foreground residual and most systematic noise. However, this step is essential to reduce the dynamic range in the contaminated image to a reasonable level for neural network training. Therefore, we plan to use a functional principle component analysis (fPCA) decomposition with the BlueBild[1] code as the main pre-process method. BlueBild calculates the sky intensity from calibrated visibilities and separates the image into a series of energy levels. The energy decomposition is performed directly on the visibilities, allowing for more efficient separation of the foreground and the 21-cm signal before convolving the image. We can filter the high-energy components associated with the foreground contamination and employ the remaining component to build back the residual image. In Figure 2, left panel, we show an example of the residual image after PCA pre-processing. In the second step, we combine the input/output of two independently trained deep neural networks. We refer to this step as the SERENEt pipeline. The former network is SegU-Net, a stable and reliable segmentation CNN for identifying neutral hydrogen regions in SKA mock images [9]. We employ this network to iden-

[1] https://github.com/epfl-radio-astro/bipp

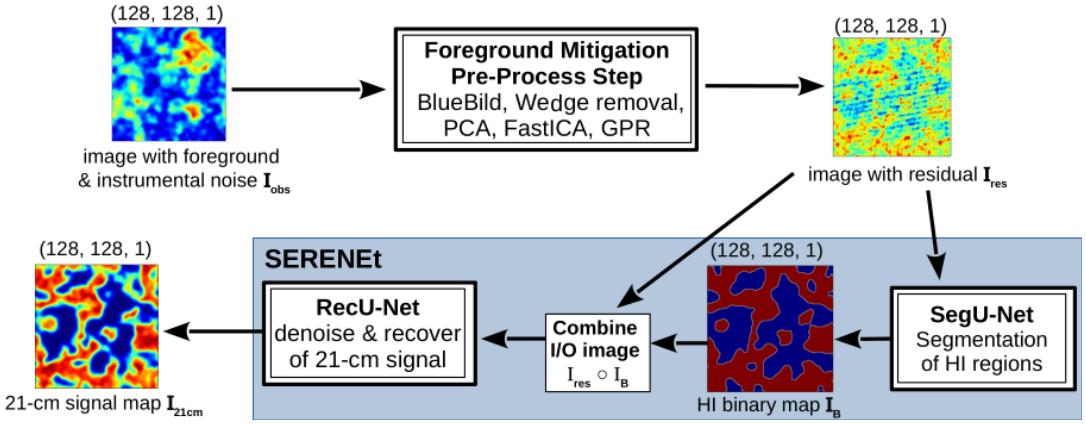

Figure 1. An overview of the `SERENEt` pipeline. The data include the mock observation with foreground and instrumental noise contamination, I_{obs}. The residual image after the pre-processing step, I_{res}. The binary prior for H I region identification, I_B and the recovered 21-cm image, I_{21cm}. Data input and output are shown with an example image. Each code step is shown in double-bordered boxes. Single-bordered boxes indicate operation on data.

tify regions of 21-cm emission from the pre-process tomographic dataset, I_{res}. The resulting binary image, I_B, will be used as a prior map for the second and final component of `SERENEt` that aims to recover the 21-cm signal, the RecU-Net network.

One of the main drawbacks of machine learning is that it often does not provide for uncertainties and confidence intervals for its predictions. Therefore, we have developed a procedure [9] that provides an error estimation with a pixel-by-pixel error in `SegU-Net` outputs image. An example of the resulting uncertainty map can be seen in Figure 2, right panel.

3 Simulation of the 21-cm & Foreground

Radio interferometry-based telescopes record the differential brightness temperature $\delta T_{\rm b}$ while observing the redshifted 21-cm signal. $\delta T_{\rm b}$ depends on position on the sky $\boldsymbol{r}$ and redshift z and can be given as [4],

$$\delta T_{\rm b}(\boldsymbol{r},z) \propto \sqrt{1+z}(\boldsymbol{r},z))\, x_{\rm HI}(\boldsymbol{x},z)\big(1+\delta_{\rm b}(\boldsymbol{r},z)\big) \quad (1)$$

where $x_{\rm HI}$ and $\delta_{\rm b}$ are neutral fractions and the baryon density contrast, respectively. Here we assumed the spin saturated approximation relevant for redshift $z < 12$ [10]. With this approximation, the differential brightness signal is always in emission ($\delta T_{\rm b} \geq 0$ mK) and locations with $\delta T_{\rm b} = 0$ mK correspond to ionised regions.

Between 250 and 30 MHz, the dominant emission comes from the Galactic synchrotron radiation. This emission alone is expected to contribute to most of the total foreground contamination of the comic 21-cm signal [11, 12]. 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 in [13]. We express the foreground radiation with a Gaussian random field with an angular power spectrum as:

$$C_l^{\rm syn}(\nu) = A_{150}\left(\frac{1000}{l}\right)^{\bar{\beta}}\left(\frac{\nu}{\nu_\star}\right)^{-2\bar{\alpha}_{syn}-2\Delta\bar{\alpha}_{syn}\log\left(\frac{\nu}{\nu_\star}\right)} \quad (2)$$

here the parameter for the Galactic synchrotron emission is the power spectra amplitude $A_{150} = 512\,{\rm mK}^2$ at the reference frequency $\nu_\star = 150$ MHz, the angular scaling $\bar{\beta} = 2.34$, the spectra index $\bar{\alpha}_{syn} = 2.8$ and the spectral running index $\Delta\bar{\alpha}_{syn} = 0.1$. These quantities are taken from [14], and [15]. We then generate the foreground temperature fluctuations map following the relation

$$\delta T_b^{\rm syn}(U,\nu) = \sqrt{\frac{\Omega_{\rm SKA}\,\mathscr{C}_l^{syn}(\nu)}{2}}\,[x_l(U)+i\cdot y_l(U)] \quad (3)$$

where $\Omega_{\rm SKA}$ is the total SKA-Low solid angle and $U = l/2\pi$. The two quantities x_l and y_l are two independent random Gaussian variables with mean zero and variance of one, $\mathscr{N} \sim (0,1)$. By performing two-dimensional inverse fast-Fourier transform of Equation 3, we get the spatial distribution of the foreground contamination $\delta T_b^{\rm syn}(\boldsymbol{r},z)$. With each lightcone simulation, we fix the random variables seed for the lowest redshift, $z = 7$, and compute Equation 2 for the corresponding frequency of the image.

4 Mock Observation for SKA-Low

To train `SERENEt`, we require a large set of simulations that represent the 21-cm radio signal for a wide range of redshift during reionization and different assumptions about the astrophysical sources of ionizing radiation. To do so, we employ `py21cmFAST` semi-numerical cosmological simulation code [16] with the approximation in Equation 3.

The simulated 21-cm signal is smoothed into a 3D lightcone of shape $(N_x, N_y, N_\nu) = (128, 128, 552)$. N_{ra} and N_{dec} correspond to the sky coordinated in comoving Mpc with a resolution of $\Delta x = 2$ cMpc. This intrinsic resolution corresponds to an angular aperture of $\Delta\theta = 0.777$ arcmin along

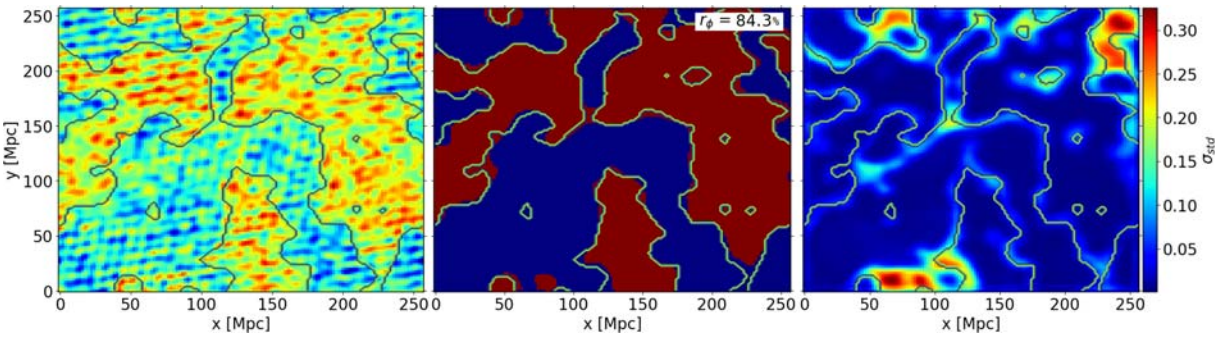

Figure 2. Slice comparison of the binary field, in blue ionized regions and in red neutral. *Left panel*: residual image after the pre-process step, I_{res}. *Middle panel*: binary field recovered by `SegU-Net`, I_B. Green lines indicate the true separation between ionized/neutral regions, derived from a smoothed version of the simulated neutral hydrogen distribution. *Right panel*: the per-pixel error as calculated by `SegU-Net`. The color bar indicates the intensity of the network uncertainty.

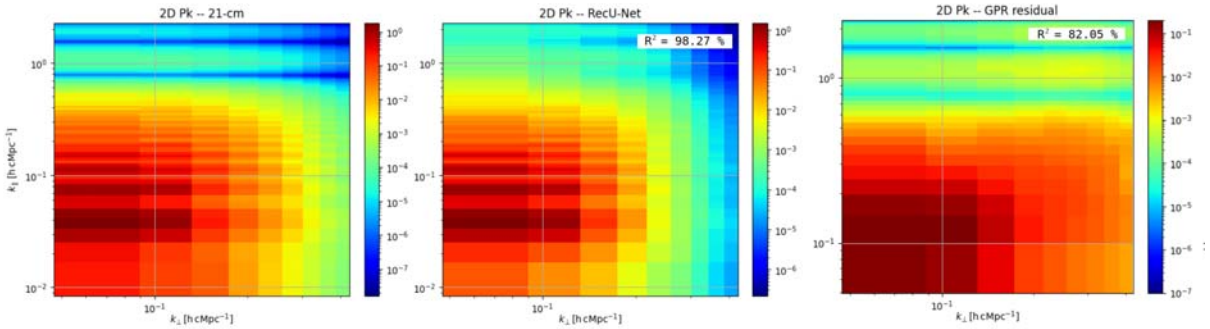

Figure 3. 2D power spectrum comparison for a tomographic data centred at 142 MHz ($z = 9$) and frequency width of 20, MHz. *Left panel*: the ground truth power spectrum from EoR 21-cm signal. *Middle panel*: 2D power spectrum from recovered 21-cm field with `RecU-Net`. *Right panel*: residual power spectrum after foreground contamination removal with GPR.

the line of sight at $z = 7$. N_ν is the number of frequency channels from 118 to 178 MHz, corresponding to redshift 11 to 7.

We simulate the instrumental noise produced from 1000 hours of observation following the methods in [17]. We then mimic the telescope limited resolution in the field-of-view direction by smoothing with a Gaussian kernel with full-width at half maximum (FWHM) of $\lambda_0(1+z)/B$, where B is the maximum baseline. For example, $B = 2$ km corresponds to a resolution of 2.905 arcmins at redshift $z \approx 7$ and 3.631 arcmins at redshift $z \approx 9$ respectively. In the frequency direction, we reduce the resolution by convolving with a top-hat bandwidth filter of a width matching the FWHM of the angular smoothing in comoving units. This width corresponds to 0.462 MHz at redshift $z \approx 7$ and 0.551 MHz at redshift $z \approx 9$, respectively.

5 Resutls

In Figure 2, we show a visual comparison of slices of the binary field predicted by `SegU-Net` (central panel) with the ground truth (green contours). Here, the ground truth is the boundary of ionized regions extracted from the simulation neutral fraction field at the same resolution by putting a threshold of 0.5. The red and blue pixels represent neutral and ionized pixels, respectively. We show the pixel error estimated from `SegU-Net` with a colour bar in the right panel. In the left panel, we show the residual image after the pre-processing step for foreground mitigation.

`SegU-Net` shows an accuracy of $r_\phi \simeq 85\%$ in recovering shapes of the H I regions. As expected, most of the network uncertainty is located at the boundaries of neutral regions or between two large ionized bubbles when these are percolating, and the gap is getting narrower. This uncertainty affects small neutral islands of a few cMpc scale residing in vast ionized regions. Moreover, larger uncertainties, $\sigma_{\rm std} \geq 0.25$, are located around narrow ionized regions protruding into large neutral regions (e.g. right-most panel, at coordinates $x \sim 75$ and $y \sim 25$). This behaviour suggests that the uncertainty mainly depends on the contrast between the local neutral and ionized regions. The network selects regions in the image based on the largest gradient in the 21-cm signal intensities to recover the binary field. Therefore, we expect larger uncertainties for reionization scenarios where the contrast in the 21-cm intensities is relatively small.

As illustrated in Figure 1, the output of `SegU-Net`, I_B, and the residual image are combined together and constitute the input of `RecU-Net`, $I_B \circ I_{res}$. The resulting output is the recovered 21-cm signal at a given frequency, I_{21cm}. Sliding trough the N_ν frequency channels, we can recover the entire tomographic data from frequency 118 to 178 MHz. In Fig-

ure 3 central panel, we show the 2D power spectrum of the recovered 21-cm signal for a sub-region of the tomographic dataset, centred at 142 MHz ($z = 9$) and a frequency depth of ± 10 MHz. A correlation of $R^2 \simeq 98\%$ between the power spectra of the 21-cm field recovered by `RecU-Net` and the ground truth (on the same figure, left panel) shows that our approach can achieve an improvement of 15% accuracy when compared to state-of-the-art foreground removal technique such as the Gaussian Regression Process (GPR) [18] (Figure 3, right panel).

References

[1] S. R. Furlanetto, S. P. Oh, and F. H. Briggs, "Cosmology at low frequencies: The 21 cm transition and the high-redshift Universe," *Physics Reports*, vol. 433, pp. 181–301, 10 2006.

[2] S. Zaroubi, "The epoch of reionization," *Astrophysics and Space Science Library*, p. 45–101, Sep 2012. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-32362-1_2

[3] A. Ferrara and S. Pandolfi, "Reionization of the Intergalactic Medium," *Proc. Int. Sch. Phys. Fermi*, vol. 186, pp. 1–57, 2014.

[4] G. Mellema, L. V. E. Koopmans, F. A. Abdalla, G. Bernardi, B. Ciardi, S. Daiboo, A. G. de Bruyn, K. K. Datta, H. Falcke, A. Ferrara, I. T. Iliev, F. Iocco, V. Jelić, H. Jensen, R. Joseph, P. Labroupoulos, A. Meiksin, A. Mesinger, A. R. Offringa, V. N. Pandey, J. R. Pritchard, M. G. Santos, D. J. Schwarz, B. Semelin, H. Vedantham, S. Yatawatta, and S. Zaroubi, "Reionization and the Cosmic Dawn with the Square Kilometre Array," *Experimental Astronomy*, vol. 36, no. 1-2, pp. 235–318, 8 2013. [Online]. Available: http://link.springer.com/10.1007/s10686-013-9334-5

[5] S. Wyithe, P. M. Geil, and H. Kim, "Imaging HII Regions from Galaxies and Quasars During Reionisation with SKA," *PoS*, vol. AASKA14, p. 015, 2015.

[6] L. V. E. Koopmans *et al.*, "The Cosmic Dawn and Epoch of Reionization with the Square Kilometre Array," *PoS*, vol. AASKA14, p. 001, 2015.

[7] O. Ronneberger, P. Fischer, and T. Brox, 2015.

[8] E. Chapman, F. B. Abdalla, G. Harker, V. Jelić , P. Labropoulos, S. Zaroubi, M. A. Brentjens, A. G. de Bruyn, and L. V. E. Koopmans, "Foreground removal usingscpfastica/scp: a showcase of LOFAR-EoR," *Monthly Notices of the Royal Astronomical Society*, vol. 423, no. 3, pp. 2518–2532, may 2012. [Online]. Available: https://doi.org/10.1111%2Fj.1365-2966.2012.21065.x

[9] M. Bianco, S. K. Giri, I. T. Iliev, and G. Mellema, "Deep learning approach for identification of H ii regions during reionization in 21-cm observations," *MNRAS*, vol. 505, no. 3, pp. 3982–3997, 2021.

[10] S. R. Furlanetto, "The global 21-centimeter background from high redshifts," *MNRAS*, vol. 371, no. 2, pp. 867–878, 08 2006. [Online]. Available: https://doi.org/10.1111/j.1365-2966.2006.10725.x

[11] T. Di Matteo, B. Ciardi, and F. Miniati, "The 21-cm emission from the reionization epoch: extended and point source foregrounds," *MNRAS*, vol. 355, pp. 1053–1065, 12 2004.

[12] V. Jelić, S. Zaroubi, P. Labropoulos, R. M. Thomas, G. Bernardi, M. A. Brentjens, A. G. de Bruyn, B. Ciardi, G. Harker, L. V. E. Koopmans, V. N. Pandey, J. Schaye, and S. Yatawatta, "Foreground simulations for the LOFAR-epoch of reionization experiment," , vol. 389, no. 3, pp. 1319–1335, Sep. 2008.

[13] S. Choudhuri, S. Bharadwaj, A. Ghosh, and S. S. Ali, "Visibility-based angular power spectrum estimation in low-frequency radio interferometric observations," *MNRAS*, vol. 445, pp. 4351–4365, 12 2014.

[14] P. Platania, M. Bensadoun, M. Bersanelli, G. De Amici, A. Kogut, S. Levin, D. Maino, and G. F. Smoot, "A Determination of the Spectral Index of Galactic Synchrotron Emission in the 1-10 GHz Range," *ApJ*, vol. 505, pp. 473–483, 10 1998.

[15] X. Wang, M. Tegmark, M. G. Santos, and L. Knox, "21 cm Tomography with Foregrounds," *ApJ*, vol. 650, pp. 529–537, 10 2006.

[16] S. G. Murray, B. Greig, A. Mesinger, J. B. Muñoz, Y. Qin, J. Park, and C. A. Watkinson, "21cmfast v3: A python-integrated c code for generating 3d realizations of the cosmic 21cm signal." *Journal of Open Source Software*, vol. 5, no. 54, p. 2582, 2020. [Online]. Available: https://doi.org/10.21105/joss.02582

[17] S. K. Giri, G. Mellema, K. L. Dixon, and I. T. Iliev, "Bubble size statistics during reionization from 21-cm tomography," *MNRAS*, vol. 473, no. 3, pp. 2949–2964, 10 2018. [Online]. Available: https://doi.org/10.1093/mnras/stx2539

[18] F. G. Mertens, M. Mevius, L. V. E. Koopmans, A. R. Offringa, G. Mellema, S. Zaroubi, M. A. Brentjens, H. Gan, B. K. Gehlot, V. N. Pandey, A. M. Sardarabadi, H. K. Vedantham, S. Yatawatta, K. M. B. Asad, B. Ciardi, E. Chapman, S. Gazagnes, R. Ghara, A. Ghosh, S. K. Giri, I. T. Iliev, V. Jelić , R. Kooistra, R. Mondal, J. Schaye, and M. B. Silva, "Improved upper limits on the 21 cm signal power spectrum of neutral hydrogen at z $\approx$ 9.1 from LOFAR," *Monthly Notices of the Royal Astronomical Society*, vol. 493, no. 2, pp. 1662–1685, feb 2020. [Online]. Available: https://doi.org/10.1093%2Fmnras%2Fstaa327