



Machine Learning Frameworks for Large-Scale Radio Surveys: A Summary of Recent Studies

Nikhel Gupta⁽¹⁾

(1) CSIRO Space & Astronomy, PO Box 1130, Bentley WA 6102, Australia

Abstract

The rapid growth of large-scale radio surveys, generating over 100 petabytes of data annually, has created a pressing need for automated data analysis methods. Recent research has explored the application of machine learning techniques to address the challenges associated with detecting and classifying radio galaxies, as well as discovering peculiar radio sources. This paper provides an overview of our investigations with the Evolutionary Map of the Universe (EMU) survey, detailing the methodologies employed—including supervised, unsupervised, self-supervised, and weakly supervised learning approaches—and their implications for ongoing and future radio astronomical surveys.

1 Introduction

The exploration of the cosmos has been revolutionized by large-scale radio surveys, enabling deeper insights into the Universe. Advanced radio interferometers, such as the Australian Square Kilometre Array Pathfinder (ASKAP; [1]–[3]), the Murchison Widefield Array (MWA; [4], [5]), MeerKAT [6], the Low Frequency Array (LOFAR; [7]), and the Karl G. Jansky Very Large Array (JVLA; [8]), play a crucial role in conducting these surveys. Among them, the Evolutionary Map of the Universe (EMU; [9]) survey, carried out with ASKAP, exemplifies the transformative impact of modern radio interferometers. Over five years, EMU aims to detect about 20 million compact and extended radio galaxies, producing an extensive dataset that will refine our understanding of galaxy evolution and the Universe’s history. Furthermore, this wealth of data is expected to reveal novel astrophysical phenomena and provide deeper insights into the origins of radio emissions. However, to fully harness its potential, traditional data analysis methods must be supplemented with advanced techniques.

In response to the challenges posed by the vast data volumes generated by next-generation radio telescopes, machine learning has emerged as a key analytical tool (e.g., [10]–[27]). These methods have significantly accelerated the identification of new radio morphologies while enhancing the efficiency of source detection, classification, and cataloguing. This paper explores various machine learning techniques applied to radio astronomical data from the EMU survey, emphasizing the role of labeled datasets. We

discuss different approaches, including supervised, unsupervised, self-supervised, and weakly-supervised learning, demonstrating their impact on large-scale radio surveys. This paper is structured as follows: Section 2 describes the EMU survey. Section 3 outlines the machine learning techniques employed, along with a summary of the investigations conducted using the EMU survey data. Section 6 gives a brief conclusion.

2 ASKAP’s EMU Survey

The EMU survey aims to develop a comprehensive radio atlas of the southern sky using the advanced ASKAP radio telescope. The survey will result in 853 tile footprints from 1,014 individual tile observations. Of these, 692 tiles will have a 10-hour integration time, while 161 tiles, observed twice, will have a 5-hour integration time. The 10-hour integration observations will achieve a median RMS noise of 30 $\mu\text{Jy}/\text{beam}$ and typically detect around 750 sources per square degree, while the 5-hour integrations will have a median RMS noise of 46 $\mu\text{Jy}/\text{beam}$, detecting about 460 sources per square degree. The five-year survey will cover the 800–1088 MHz frequency range centered at 944 MHz. EMU will map declinations from -11.4° to $+7.0^\circ$ and is expected to catalogue approximately 20 million radio sources across 2π steradians, with completion anticipated by 2028. In the first year of EMU observations, a total of 160 tiles were observed, covering about 4,500 square degrees of the sky. Data collection for the survey commenced in late 2022, with validated data being delivered between February 2023 and March 2024. These datasets include tiles with Scheduling Block unique ID (SBID) numbers ranging from 45638 to 59612, and are accessible via the CSIRO Data Access Portal (CASDA¹). Prior to the main EMU survey, the first EMU Pilot Survey (EMU-PS1; [28]) was conducted in late 2019, covering an area of 270 deg^2 in the region of $301^\circ < \text{RA} < 336^\circ$ and $-63^\circ < \text{Dec} < -48^\circ$, composed of 10 tiles with a total integration time of about 10 hours per tile.

3 Machine Learning Methods

In this section, we discuss different machine learning techniques employed to analyse the data from the EMU-PS1 and the first year of the main EMU survey.

¹<https://research.csiro.au/casda/>

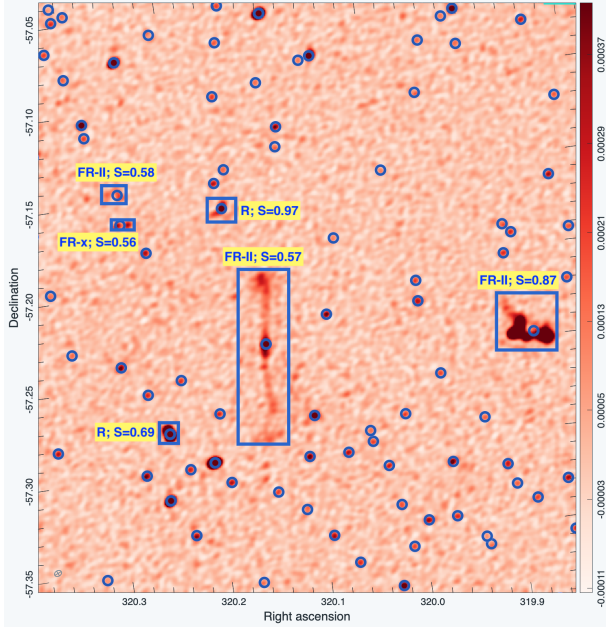


Figure 1. An example of radio galaxies from the RG-CAT catalogue, superimposed on the EMU-PS1 radio image, highlighting their host positions, bounding boxes, and classifications generated by the Gal-DINO network. See more examples in [21].

3.1 Supervised Learning

Supervised learning is a machine learning paradigm where algorithms learn from labeled datasets to predict outcomes or make decisions based on input data. To develop these algorithms, appropriate labels are required for detection and classification tasks. We introduced RadioGalaxyNET [14], which consisted of a labeled dataset and a suite of computer vision models designed for detecting extended radio sources and their corresponding infrared host galaxies. The dataset included 2,800 images, comprising both radio and infrared sky channels, and contained 4,155 instances of labeled sources. Radio images were sourced from the EMU-PS1. The identification of radio galaxies and the curation of annotations, such as class labels, bounding boxes, and segmentation masks, were performed through visual inspections [29]. Additionally, infrared images from the Wide-field Infrared Survey Explorer (WISE; [30], [31]) were aligned with the sky positions of the radio images, and the infrared host galaxies were visually identified in the infrared images. We employed object detection methods for the detection and classification of radio galaxies. Our computer vision methods included multimodal models, such as Gal-DETR², Gal-Deformable DETR³, and Gal-DINO⁴, where we introduced a mechanism for the simultaneous detection of radio galaxies and infrared hosts through bounding box and keypoint detections, respectively. Our findings revealed that the Gal-DINO model exhibited su-

perior performance in detecting both radio galaxies and infrared hosts. In addition to these models, we conducted a comparative analysis of our radio galaxy detection results with those produced by the Gal-SIOD⁵, Gal-Faster RCNN⁶, and Gal-YOLOv8⁷ models. Our results indicated that Gal-DINO and Gal-YOLOv8 yielded comparable outcomes in radio galaxy detection.

In another work, we developed the RG-CAT pipeline [21] to build radio galaxy catalogues using a two-step process: first detecting compact and extended radio galaxies and their host galaxies in radio and infrared images with computer vision networks, then using these predictions to construct a catalogue. This pipeline led to the creation of the radio galaxy catalogue for the EMU-PS1. We built the detection pipeline based on the GalDINO network, expanding RadioGalaxyNET’s 2,800 radio galaxies. After visual inspections, we labeled 2,090 compact and 99 rare and peculiar radio sources, resulting in a dataset of $\sim 5,000$ radio galaxies for training and evaluation. In evaluations, 99% of radio galaxy predictions in the centre of the images had an Intersection over Union (IoU) greater than 0.5, and 98% of keypoint detections had less than $3''$ separation from the ground truth. The final catalogue included 211,625 radio galaxies: 201,211 compact and 10,414 extended, with sub-categories. Cross-matching with infrared (CatWISE; [32]) and optical catalogues (DES DR2; [33], DESI Legacy Surveys DR8 redshifts; [34], SuperCosmos; [35]) revealed that 75% of radio galaxies had counterparts in CatWISE. Additionally, 64%, 37%, and 11% were matched with DES, DESI, and SuperCosmos, respectively. In ongoing work [36], we apply the RG-CAT pipeline to construct catalogues from the EMU main survey, further improving the Gal-DINO network with rotated bounding boxes and incorporating online learning.

As Gal-DINO is trained to detect both rare and peculiar radio sources, along with other radio galaxies, we further applied the trained model to the first year of the EMU main survey to identify these sources in a supervised manner [26]. This allowed for the efficient processing of large radio datasets, enabling the rapid discovery of rare and peculiar radio sources such as Odd Radio Circles (ORCs; [37]) with edge-brightened rings, Galaxies with Large-scale Ambient Radio Emissions (GLAREs), and Starburst Radio Ring Galaxies (SRRGs). We identified 5 new ORCs, 2 candidate ORCs, 56 GLAREs, and 18 SRRGs, offering new insights into their potential formation mechanisms. While such tasks are typically performed using unsupervised machine learning methods—since the target objects are unknown—here we demonstrate an efficient supervised learning approach.

²<https://github.com/Nikheli1/Gal-DETR>

³<https://github.com/Nikheli1/Gal-Deformable-DETR>

⁴<https://github.com/Nikheli1/Gal-DINO>

⁵<https://github.com/Nikheli1/Gal-SIOD>

⁶<https://github.com/Nikheli1/Gal-Faster-RCNN>

⁷<https://github.com/Nikheli1/Gal-YOLOv8>

3.2 Unsupervised Learning

Unsupervised learning is a machine learning approach where algorithms identify patterns or structures in unlabeled data without explicit guidance on the desired output. These methods are often applied to anomaly detection tasks. We presented an unsupervised machine learning approach to search for the rarest and most intriguing sources in EMU-PS1 [11]. We utilized self-organizing maps (SOMs; [38]) and accounted for affine transformations of astronomical images to mitigate the effects of rotational and flipping variances. SOMs were applied to approximately 42,000 cutouts at the positions of extended radio sources in EMU-PS1. The trained model was then used to map these sources into a 10×10 lattice of neurons based on their relative similarity. The Euclidean distance metric was employed to identify the rarest and most interesting sources in the survey, among which two new ORC candidates and several other complex radio morphologies were identified.

4 Self-supervised Learning

Self-supervised learning is a technique where models learn representations from unlabeled data by generating their own supervision signals. Unlike traditional unsupervised learning, which focuses on finding patterns without explicit objectives, self-supervised learning uses pretext tasks to generate meaningful supervision signals. Multimodal foundation models are pretrained using self-supervised techniques to build versatile representations across text, images, and other modalities. These models can then be fine-tuned on domain-specific datasets to adapt their general knowledge for specialized tasks. In recent work, we investigated the use of a multimodal foundation model in radio astronomy, focusing on OpenCLIP [39], an open-source pretrained model, to classify and retrieve radio sources from the EMU survey. This study aims to improve the identification and retrieval of various radio galaxies by integrating visual and textual information into a shared embedding space. We fine-tune OpenCLIP on the RadioGalaxyNET dataset, which includes extended radio galaxies and other rare and peculiar radio sources [27]. The model maps radio and infrared images to a shared latent space alongside their associated textual descriptions, enabling zero-shot classification and retrieval tasks without task-specific training. This latent representation significantly reduces storage requirements by preserving only the embeddings containing relevant information, while also accelerating data retrieval from hours to sub-seconds. Such methods have the potential to aid in task-specific data compression. This approach led to the development of a search tool for the EMU main survey, EMUSE (Evolutionary Map of the Universe Search Engine), accessible through a web-based application⁸, and it can also be used locally⁹.

⁸<https://askap-emuse.streamlit.app/>

⁹follow steps at <https://github.com/Nikhel1/EMUSE>

5 Weakly-supervised Learning

Weakly supervised learning involves training models using incomplete, imprecise, or indirect supervision, such as noisy labels or approximate annotations, offering more flexibility than fully supervised learning while relying less on explicit labeling. Unlike unsupervised or self-supervised learning, which do not use labeled data, weakly supervised learning bridges this gap by leveraging limited or imperfect labels. In a recent study, we applied a weakly-supervised deep learning method, Gal-CAM¹⁰, to detect and classify extended radio galaxies [14]. The algorithm was trained on weak class-level labels to generate class activation maps (CAMs; [40]), which were further refined using an inter-pixel relations network (IRNet; [41]) to produce instance segmentation masks for the radio galaxies and their infrared host positions. By using only class-level information for radio galaxies, we successfully detected them in the EMU-PS1 survey images. Our results demonstrated that weakly-supervised deep learning could accurately predict pixel-level information, such as masks or bounding boxes for extended radio emissions and the positions of infrared hosts, even without direct pixel-level labels during training. We evaluated our method using mean Average Precision (mAP) at a standard intersection over union (IoU) threshold of 0.5, achieving mAP₅₀ scores of 67.5% for radio masks and 76.8% for infrared host positions.

6 Conclusions

The integration of machine learning techniques has significantly advanced the analysis of data from the ongoing Evolutionary Map of the Universe (EMU) radio survey of the southern sky, allowing for more efficient detection, classification, and cataloging of radio sources. By leveraging supervised, unsupervised, self-supervised, and weakly-supervised learning approaches, we demonstrate that how innovative methods can enhance our understanding of complex radio galaxies and uncover new astrophysical phenomena. Through the development of the curated dataset RadioGalaxyNET and the application of cutting-edge object detection models such as Gal-DINO, substantial progress has been made in detecting and classifying radio galaxies. The advancements, enabled by the catalogue construction pipeline RG-CAT and the search engine EMUSE, have further enhanced the processing of vast radio datasets from surveys like EMU, leading to the discovery of rare and peculiar radio sources. Machine learning’s ability to efficiently analyse large-scale data has bridged the gap between complex datasets and meaningful astronomical insights, offering a promising future for generating expedited scientific results from next-generation radio surveys. As these surveys grow, and with the advent of the Square Kilometre Array (SKA¹¹) era, the continuous application and refinement of machine learning models will be essential in uncovering

¹⁰<https://github.com/Nikhel1/Gal-CAM>

¹¹<https://www.skao.int/en>

deeper insights into the radio universe and further enhancing the exploration of cosmic phenomena.

7 Acknowledgements

NG acknowledges support from CSIRO's Machine Learning and Artificial Intelligence Future Science Impossible Without You (MLAI FSP IWY) Platform. The scientific work discussed here uses data obtained from Inyarrimanha Ilgari Bundara / the Murchison Radio-astronomy Observatory. We acknowledge the Wajarri Yamaji People as the Traditional Owners and native title holders of the Observatory site. The Australian SKA Pathfinder is part of the Australia Telescope National Facility (<https://ror.org/05qajvd42>) which is managed by CSIRO. Operation of ASKAP is funded by the Australian Government with support from the National Collaborative Research Infrastructure Strategy. ASKAP uses the resources of the Pawsey Supercomputing Centre. The establishment of ASKAP, the Murchison Radio-astronomy Observatory and the Pawsey Supercomputing Centre are initiatives of the Australian Government, with support from the Government of Western Australia and the Science and Industry Endowment Fund. This paper includes archived data obtained through the CSIRO ASKAP Science Data Archive, CASDA (<http://data.csiro.au>).

References

- [1] S. Johnston *et al.*, "Science with the Australian Square Kilometre Array Pathfinder," vol. 24, no. 4, pp. 174–188, Dec. 2007. DOI: 10.1071/AS07033. arXiv: 0711.2103 [astro-ph].
- [2] D. R. DeBoer *et al.*, "Australian SKA Pathfinder: A High-Dynamic Range Wide-Field of View Survey Telescope," *IEEE Proceedings*, vol. 97, no. 8, pp. 1507–1521, Aug. 2009. DOI: 10.1109/JPR0C.2009.2016516.
- [3] A. W. Hotan *et al.*, "Australian square kilometre array pathfinder: I. system description," *PASA*, vol. 38, e009, e009, Mar. 2021. DOI: 10.1017/pasa.2021.1. arXiv: 2102.01870 [astro-ph.IM].
- [4] S. J. Tingay *et al.*, "The Murchison Widefield Array: The Square Kilometre Array Precursor at Low Radio Frequencies," vol. 30, e007, e007, Jan. 2013. DOI: 10.1017/pasa.2012.007. arXiv: 1206.6945 [astro-ph.IM].
- [5] R. B. Wayth *et al.*, "The Phase II Murchison Widefield Array: Design overview," *PASA*, vol. 35, e033, e033, Nov. 2018. DOI: 10.1017/pasa.2018.37. arXiv: 1809.06466 [astro-ph.IM].
- [6] J. Jonas and MeerKAT Team, "The MeerKAT Radio Telescope," in *MeerKAT Science: On the Pathway to the SKA*, Jan. 2016, 1, p. 1.
- [7] M. P. van Haarlem *et al.*, "LOFAR: The LOW-Frequency ARray," vol. 556, A2, A2, Aug. 2013. DOI: 10.1051/0004-6361/201220873. arXiv: 1305.3550 [astro-ph.IM].
- [8] R. A. Perley *et al.*, "The Expanded Very Large Array: A New Telescope for New Science," *APJ*, vol. 739, no. 1, L1, p. L1, Sep. 2011. DOI: 10.1088/2041-8205/739/1/L1. arXiv: 1106.0532 [astro-ph.IM].
- [9] A. M. Hopkins *et al.*, "The evolutionary map of the universe: A new radio atlas for the southern hemisphere sky," submitted to *PASA* journal.
- [10] R. I. J. Mostert *et al.*, "Unveiling the rarest morphologies of the LOFAR Two-metre Sky Survey radio source population with self-organised maps," vol. 645, A89, A89, Jan. 2021. DOI: 10.1051/0004-6361/202038500. arXiv: 2011.06001 [astro-ph.IM].
- [11] N. Gupta *et al.*, "Discovery of peculiar radio morphologies with ASKAP using unsupervised machine learning," vol. 39, e051, e051, Oct. 2022. DOI: 10.1017/pasa.2022.44. arXiv: 2208.13997 [astro-ph.GA].
- [12] M. Walmsley *et al.*, "Practical galaxy morphology tools from deep supervised representation learning," vol. 513, no. 2, pp. 1581–1599, Jun. 2022. DOI: 10.1093/mnras/stac525. arXiv: 2110.12735 [astro-ph.GA].
- [13] G. Segal *et al.*, "Identifying anomalous radio sources in the EMU Pilot Survey using a complexity-based approach," *arXiv e-prints*, arXiv:2206.14677, arXiv:2206.14677, Jun. 2022. arXiv: 2206.14677 [astro-ph.IM].
- [14] N. Gupta *et al.*, "Deep learning for morphological identification of extended radio galaxies using weak labels," *Publications of the Astronomical Society of Australia*, vol. 40, e044, 2023. DOI: 10.1017/pasa.2023.46.
- [15] M. Lochner *et al.*, "A unique, ring-like radio source with quadrilateral structure detected with machine learning," vol. 520, no. 1, pp. 1439–1446, Mar. 2023. DOI: 10.1093/mnras/stad074. arXiv: 2211.02062 [astro-ph.GA].
- [16] N. Gupta *et al.*, "A Multimodal Dataset and Benchmark for Radio Galaxy and Infrared Host Detection," *NeurIPS ML4PS 2023*, arXiv:2312.06728, arXiv:2312.06728, Dec. 2023. DOI: 10.48550/arXiv.2312.06728. arXiv: 2312.06728 [cs.CV].
- [17] I. V. Slijepcevic *et al.*, "Radio galaxy zoo: towards building the first multipurpose foundation model for radio astronomy with self-supervised learning," *RAS Techniques and Instruments*, vol. 3, no. 1, pp. 19–32, Jan. 2024. DOI: 10.1093/rasti/rzad055. arXiv: 2305.16127 [astro-ph.IM].

- [18] K. Mohale and M. Lochner, “Enabling unsupervised discovery in astronomical images through self-supervised representations,” vol. 530, no. 1, pp. 1274–1295, May 2024. DOI: 10.1093/mnras/stae926. arXiv: 2311.14157 [astro-ph.IM].
- [19] N. Gupta *et al.*, “RadioGalaxyNET: Dataset and novel computer vision algorithms for the detection of extended radio galaxies and infrared hosts,” vol. 41, e001, e001, Jan. 2024. DOI: 10.1017/pasa.2023.64. arXiv: 2312.00306 [astro-ph.IM].
- [20] E. Lastufka *et al.*, “Self-supervised learning on MeerKAT wide-field continuum images,” vol. 690, A310, A310, Oct. 2024. DOI: 10.1051/0004-6361/202449964. arXiv: 2408.06147 [astro-ph.IM].
- [21] N. Gupta *et al.*, “RG-CAT: Detection pipeline and catalogue of radio galaxies in the EMU pilot survey,” vol. 41, e027, e027, Apr. 2024. DOI: 10.1017/pasa.2024.25. arXiv: 2403.14235 [astro-ph.GA].
- [22] S. Riggi *et al.*, “Self-supervised contrastive learning of radio data for source detection, classification and peculiar object discovery,” vol. 41, e085, e085, Nov. 2024. DOI: 10.1017/pasa.2024.84. arXiv: 2404.18462 [astro-ph.IM].
- [23] M. Lochner and L. Rudnick, “Astronomy Protege: Discovery Through Human-Machine Collaboration,” *arXiv e-prints*, arXiv:2411.04188, arXiv:2411.04188, Nov. 2024. DOI: 10.48550/arXiv.2411.04188. arXiv: 2411.04188 [astro-ph.IM].
- [24] R. I. J. Mostert *et al.*, “Constraining the giant radio galaxy population with machine learning and Bayesian inference,” vol. 691, A185, A185, Nov. 2024. DOI: 10.1051/0004-6361/202348897. arXiv: 2405.00232 [astro-ph.GA].
- [25] B. Lao *et al.*, “Identification of 4876 Bent-Tail Radio Galaxies in the FIRST Survey Using Deep Learning Combined with Visual Inspection,” *arXiv e-prints*, arXiv:2501.09883, arXiv:2501.09883, Jan. 2025. DOI: 10.48550/arXiv.2501.09883. arXiv: 2501.09883 [astro-ph.GA].
- [26] N. Gupta *et al.*, “Discovery of odd radio circles and other peculiars in the first year of the emu survey using object detection,” submitted to PASA journal.
- [27] N. Gupta *et al.*, “Emuse: Evolutionary map of the universe search engine,” in preparation.
- [28] R. P. Norris *et al.*, “The Evolutionary Map of the Universe pilot survey,” vol. 38, e046, e046, Sep. 2021. DOI: 10.1017/pasa.2021.42. arXiv: 2108.00569 [astro-ph.CO].
- [29] Yew *et al.*, “Emu and the dragns paper i: A catalogue of dragns,” in preparation.
- [30] E. L. Wright *et al.*, “The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance,” vol. 140, 1868, pp. 1868–1881, Dec. 2010. DOI: 10.1088/0004-6256/140/6/1868. arXiv: 1008.0031 [astro-ph.IM].
- [31] R. M. Cutri *et al.*, “VizieR Online Data Catalog: All-WISE Data Release (Cutri+ 2013),” *VizieR Online Data Catalog*, II/328, pp. II/328, Feb. 2021.
- [32] F. Marocco *et al.*, “The CatWISE2020 Catalog,” vol. 253, no. 1, 8, p. 8, Mar. 2021. DOI: 10.3847/1538-4365/abd805. arXiv: 2012.13084 [astro-ph.IM].
- [33] T. M. C. Abbott *et al.*, “The Dark Energy Survey Data Release 2,” vol. 255, no. 2, 20, p. 20, Aug. 2021. DOI: 10.3847/1538-4365/ac00b3. arXiv: 2101.05765 [astro-ph.IM].
- [34] H. Zou *et al.*, “Photometric redshifts and stellar masses for galaxies from the desi legacy imaging surveys,” *The Astrophysical Journal Supplement Series*, vol. 242, no. 1, p. 8, 2019. DOI: 10.3847/1538-4365/ab1847. [Online]. Available: <https://ui.adsabs.harvard.edu/abs/2019ApJS.242...8Z/abstract>.
- [35] M. Bilicki *et al.*, “WISE × SuperCOSMOS Photometric Redshift Catalog: 20 Million Galaxies over 3π Steradians,” vol. 225, no. 1, 5, p. 5, Jul. 2016. DOI: 10.3847/0067-0049/225/1/5. arXiv: 1607.01182 [astro-ph.CO].
- [36] N. Gupta *et al.*, “Catalogue of radio galaxies for the first year of emu survey,” in preparation.
- [37] R. P. Norris *et al.*, “Unexpected circular radio objects at high Galactic latitude,” vol. 38, e003, e003, Jan. 2021. DOI: 10.1017/pasa.2020.52. arXiv: 2006.14805 [astro-ph.GA].
- [38] T. Kohonen, “Self-organized formation of topologically correct feature maps,” *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982. DOI: 10.1007/BF00337288.
- [39] M. Cherti *et al.*, “Reproducible scaling laws for contrastive language-image learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 2818–2829.
- [40] B. Zhou *et al.*, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.
- [41] J. Ahn *et al.*, “Weakly supervised learning of instance segmentation with inter-pixel relations,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2209–2218.