Cosmic depth and detail: advancing LOFAR imaging workflows to unveil the deep high-resolution universe

Jong, J.M.G.H.J. de

6

Advanced strategy for deep sub-arcsecond wide-field facet calibration with LOFAR

J.M.G.H.J. de Jong, L. Veefkind, R.J. van Weeren, R.J. Schlimbach, J.B.R. Oonk, D.N.G. Kampert, H.J.A. Röttgering, M. van der Wild

Abstract

We address key bottlenecks that limit the processing of a large number of observations of the same sky area for sub-arcsecond wide-field imaging with the International Low-Frequency Array (LOFAR). Our improvements are focused on three critical aspects: First, we establish a framework to streamline the data reduction of LOFAR datasets on HPC systems. Second, we refine and further automate the direction-dependent calibration by employing performance metrics, linking S/N to solution intervals, and integrating a neural network to identify image artefacts. Third, by revisiting the imaging strategy, we reduce the total required computational resources up to a factor of 3 to 4, depending on whether processing a single 8-hour observation or combining hundreds of hours of data. We demonstrate the robustness of our new data reduction strategy through sub-arcsecond imaging on 1 already successfully reduced dataset and 1 new challenging dataset from the ELAIS-N1 deep field. This work enables the processing of hundreds of hours of LOFAR data of the same pointing, paving the way for the creation of the deepest high-quality sub-arcsecond wide-field images with sensitivities on the order of a few μJy beam^{-1} at 144 MHz. We also outline additional development steps needed to establish an automated survey pipeline for sub-arcsecond wide-field imaging of other sky regions using LOFAR.

6.1. Introduction

Deep high-resolution wide-field surveys at low radio frequencies provide us with essential information about the evolution of our universe. These surveys can be done using the International LOw Frequency ARray (LOFAR) Telescope (ILT van Haarlem et al., 2013). This radio telescope is sensitive to detecting low-frequency radio waves between 10-80 MHz with the low-band antennas (LBAs) and between 110-240 MHz with the high-band antennas (HBAs). Recent efforts to automatically calibrate and image LOFAR observations with only the Dutch LOFAR HBAs have led to the LOFAR Two-metre Sky Survey (LoTSS; Shimwell et al., 2017, 2019, 2022; Williams et al., 2019) and the LoTSS-Deep Fields (Kondapally et al., 2021; Duncan et al., 2021; Tasse et al., 2021; Sabater et al., 2021; Best et al., 2023; Bondi et al., 2024; Shimwell et al., 2025), providing us with wide-field images of the northern sky at 144 MHz and 6″ resolution. The deepest of these maps is created with over ~500 hrs of observations of the ELAIS-N1 deep field, reaching a central sensitivity of about 11 μJy beam^{-1} (Shimwell et al., 2025). However, 40% of the image noise in this map is due to confusion noise. This can be mitigated by producing higher-resolution images through the inclusion of visibility data from all LOFAR stations across Europe. This extends baselines to ~2,000 km, enabling 20 times better resolutions, and increasing the telescope's collecting area as well (Varenius et al., 2015; Morabito et al., 2022a). The resulting sub-arcsecond resolution resolves much smaller angular structures, which unlocks a wealth of scientific opportunities to for instance study supernovae (e.g. Venkattu et al., 2023), AGN (e.g. Mahatma et al., 2023; Jurlin et al., 2024), galaxy clusters (e.g. van Weeren et al., 2024), or separate AGN activity from star-formation (Morabito et al., 2022b, 2025b). Nonetheless, the scalability of calibrating and imaging many observations with all LOFAR stations for ultra-deep wide-field imaging is currently constrained by the lack of a framework for automatic processing, the high computational costs, and challenges in achieving high-quality calibration solutions for both short and long baselines simultaneously.

Building on earlier successful work to process LOFAR data for sub-arcsecond resolution imaging (e.g. Moldón et al., 2015; Varenius et al., 2015, 2016; Jackson et al., 2016; Harris et al., 2019), the next step to advance the development of a sub-arcsecond imaging strategy for LOFAR was introduced by Morabito et al. (2022a), who focused on the general direction-independent calibration of the international stations followed by postage stamp imaging of targets. Sweijen et al. (2022c) extended this work to wide-field imaging, producing the first image that captured thousands of sources within a 2.5×2.5 deg field of view at a resolution of 0.3″. This was followed by de Jong et al. (2024), who conducted imaging with four observations of the same pointing of the ELAIS-N1 deep field, resulting in the detection of four times more sources within the same sky area compared to imaging only one

observation. In addition to increasing depth, they introduced automated direction-dependent (DD) calibrator selection, aiming to reduce the need for manual intervention. They also demonstrated that the current DD calibration strategy required additional attention since the calibration solutions, corresponding to the Dutch LOFAR stations and hence the shorter baselines, were of insufficient quality to enable high-resolution imaging. This issue was mitigated by introducing an additional ad hoc calibration step specifically for the Dutch LOFAR stations after completing all other calibration steps. While this solution reduced the problem, image artefacts across different facets of the images persist and therefore a refinement of the calibration strategy is required. Moreover, Sweijen et al. (2022c), Ye et al. (2024), and de Jong et al. (2024) demonstrated that the final imaging steps consume the majority of data processing costs, accounting for approximately 80% of the total expenses. Revising key aspects of the imaging strategy is therefore essential before scaling up sub-arcsecond wide-field imaging for deeper imaging. One promising approach to address these high computational costs when combining multiple observations of the same sky area is sidereal visibility averaging (SVA), which was recently revisited by de Jong et al. (2025). This approach takes advantage of the ability to average visibilities over repeating baseline tracks for each sidereal day, offering a solution to significantly compress data volume and thereby reduce computational demands.

Unlike for wide-field imaging at $6''$ resolution with LoTSS (e.g. Shimwell et al., 2017, 2019; Mechev et al., 2017, 2018; Tasse et al., 2021), a fully developed framework to automatically process LOFAR data for sub-arcsecond wide-field imaging has not yet been realised. The lack of such a comprehensive framework is largely due to the exploratory nature of previous studies on sub-arcsecond wide-field imaging, which focused efforts on creating and refining individual processing steps and software. However, with most of the essential main building blocks for data processing for sub-arcsecond wide-field imaging in place, it is now possible to take the next steps in the development. Processing the large data volumes – approximately 4 TB of compressed data per observation using Dynamical Statistical Compression (Dysco; Offringa, 2016) – along with the intensive computational demands of hundreds of thousands of CPU core hrs (Sweijen et al., 2022c; de Jong et al., 2024), requires the implementation of automated data processing pipelines and workflows. Here, a pipeline is defined as a structured sequence of processing steps where the output of one step directly feeds into the next, whereas a workflow refers to a flexible, organised arrangement of interconnected tasks that manage the coordination and orchestration of those tasks. To achieve feasible wall times, these processes are best executed on a high-performance computing (HPC) cluster, capable of distributing the workload efficiently across multiple interconnected computing nodes, each equipped with numerous CPUs.

In this work, we establish the next step towards a framework to streamline the data reduction of LOFAR datasets on HPC clusters, revisit and automate the DD calibration, and develop strategies to reduce the required computational resources. We demonstrate the new calibration and imaging strategy on 2 LOFAR datasets from the ELAIS-N1 field, of which one has already been reduced by de Jong et al. (2024) and one new dataset. With hundreds of hrs of data available for this field with international LOFAR stations, we pave the way for ultra-deep imaging of a single pointing, reaching a sensitivity on the order of a few μJy beam^{-1} at 144 MHz.

In Section 6.2, we introduce a framework designed to manage large-scale data processing on high-performance computing infrastructures. This is followed by Section 6.3 where we discuss steps to further automate the self-calibration of LOFAR data with long-baselines, setting the stage for Section 6.4 where we present the optimised DD calibration and imaging strategy. In Section 6.5 we outline the data used to implement the new data reduction strategy. Section 6.6 presents our results, which are discussed along with future prospects in Section 6.7. Finally, we end our work with a summary and conclusions in Section 6.8.

6.2. Data processing framework

Efficient management of processing large data volumes in the order of a few to hundreds of TB with many individual steps on high-performance computing (HPC) systems requires a well-designed framework with workflows that orchestrate task execution, handle dependencies, and utilise computational resources across multiple nodes and CPUs and their cores efficiently. For this purpose, we use the combination of the Common Workflow Language (CWL; Amstutz et al., 2016; Crusoe et al., 2022), Toil (Vivian et al., 2017), and the Simple Linux Utility for Resource Management (SLURM; Yoo et al., 2003). This approach is effective because CWL, Toil, and SLURM seamlessly integrate to facilitate efficient workflow execution:

1. CWL defines how the tasks in a workflow are structured, including their inputs, outputs, computing resources, and dependencies.

2. Toil interprets the CWL workflow, orchestrates the execution of tasks, and interacts with the computing infrastructure.

3. SLURM handles the actual job scheduling and resource allocation on an HPC cluster.

For the data reduction in this work, we utilise Spider[1], a high-throughput data

[1] `https://doc.spider.surfsara.nl`

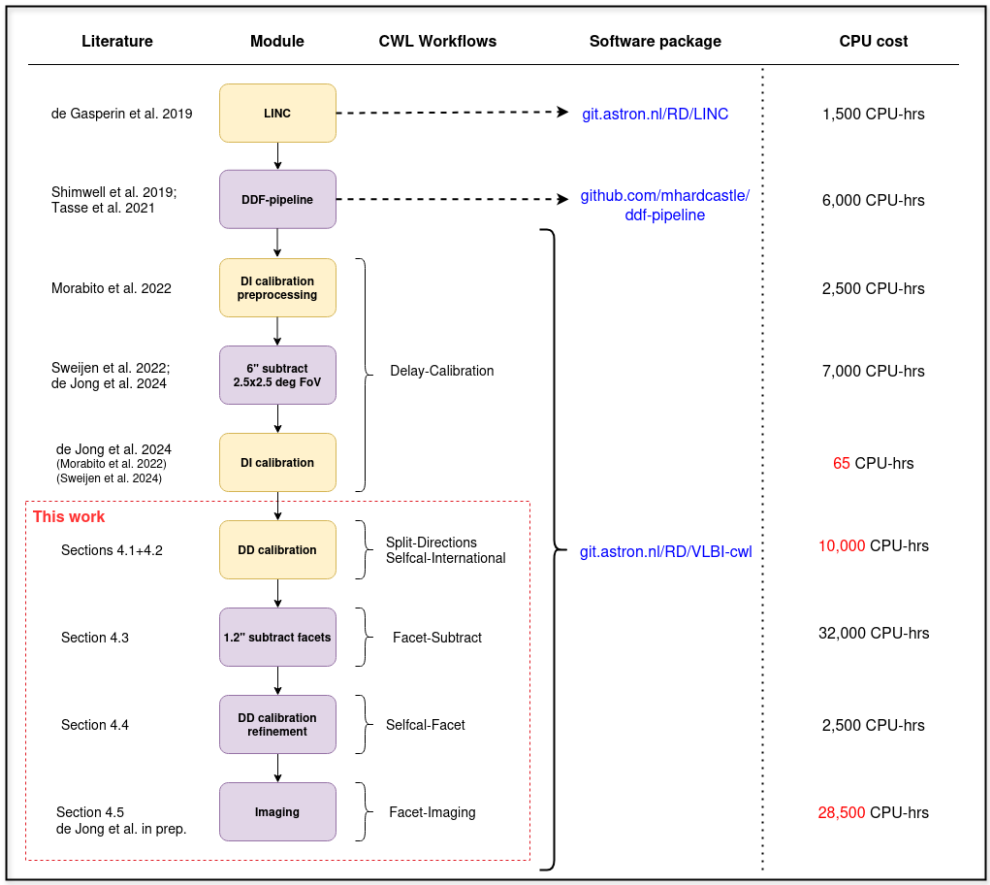

Figure 6.1: Diagram outlining the framework consisting of different modules with corresponding CWL workflows, software `git` directories, and computing requirements. The yellow modules are essential for postage stamp imaging, while the purple modules are additional modules necessary for wide-field imaging. The first 2 modules, the LOFAR Initial Calibration (`LINC`) and `DDF`-pipeline, are independent pipelines but are required to process LOFAR data for calibration of the international stations and for sub-arcsecond wide-field imaging. For the DI calibration step, we have put Morabito et al. (2022a) and Sweijen et al. (2022c) between brackets, since we refer here to the optimised step by de Jong et al. (2024) for ELAIS-N1. The CPU core hrs given in red are variable, as these will be smaller when more observations are combined for (ultra-)deep imaging (see Section 6.6.2.2) or when processing a different dataset with a different DI calibration strategy. The modules enclosed by the dashed red box represent the parts that we optimise in this work, while other steps are referenced in related literature. A more in-depth explanation of the different CWL workflows for postage stamp imaging will be discussed in van der Wild (in prep.).

processing platform provided by SURF[2]. We use AMD EPYC 7551 and AMD EPYC

[2] `https://www.surf.nl/`

7702P processor nodes, each with 60 cores, 0.5 TB and 1 TB RAM respectively, and about 4 TB of scratch storage on a local solid-state drive (SSD). A shared Ceph file system (CephFS) serves as central storage, from which data is copied to the local SSDs to increase the processing speed of I/O-heavy jobs. Once jobs are complete, the output is transferred back to CephFS to proceed to the next step, as Toil and CWL require a central shared point for communication between jobs with input and output dependencies.

The framework's overarching data processing workflow, reducing data from raw to image, is structured into distinct, self-contained modules. Each module can operate autonomously, enabling straightforward development, testing, and maintenance. This modular design enhances the workflow's scalability by allowing modules to be added, modified, and replaced independently, without having to directly affect the input of the next module. These modules are chained together as outlined in the second column in Figure 6.1, and can correspond to an entire pipeline (e.g. the `DDF`-pipeline[3]) or one or multiple CWL workflows. This modular approach allows users to customise their entire data reduction workflow by selecting the specific steps needed to achieve their desired output. The yellow modules in this figure correspond to the modules typically required to create postage stamp images (Morabito et al., 2022a), while for wide-field imaging, we use the additional purple modules (Sweijen et al., 2022c; Ye et al., 2024; de Jong et al., 2024). In Figure 6.1, we have provided an overview of the corresponding CWL workflows with the currently advised order for wide-field imaging, together with software packages and an estimate of the CPU requirements. We developed multiple CWL workflows because CWL has the flexibility to implement these as sub-workflows within an automated larger overarching workflow. This design also allows users to execute workflows step by step, providing the flexibility to manually inspect or modify the data output as needed. The modules and workflows enclosed in the red dashed box are the parts that are further improved in this work. In the discussion in Section 6.7.1 we further discuss opportunities to reorder some of the modules.

It is important to note that we measure CPU core hrs in this work as the number of allocated CPUs requested for a job per hour. This measure does not account for idle CPUs, such as when a job temporarily uses only one CPU for data transfer between nodes during a job. This way of measuring may therefore be less accurate than tracking actual CPU usage.

The required software for data processing is distributed using Singularity containers (Kurtzer et al., 2017).[4] These containers include standardised tools such

[3] `https://github.com/mhardcastle/ddf-pipeline`
[4] `https://github.com/tikk3r/flocs`

as `casacore`[5] (Casacore Team, 2019; CASA Team et al., 2022) for working with measurement sets – the standard data format for radio interferometric datasets; `astropy`[6] for general-purpose astronomy utilities (Astropy Collaboration et al., 2013, 2018); and `DP3`[7] for preprocessing LOFAR datasets (van Diepen et al., 2018a; Dijkema et al., 2023), including tasks such as flagging, averaging, and calibration. These Singularity software containers can be compiled for different AMD and Intel architectures, optimizing performance and compatibility across a wide range of HPC systems. This ensures efficient utilisation of hardware resources.

6.3. Automated long-baseline self-calibration

The ionosphere induces DD effects (DDEs) that lead to artefacts on scales from arcseconds to arcminutes (e.g. Intema et al., 2009; Smirnov, 2011b). To address these issues, we apply self-calibration on various calibrators distributed across the field of view. These sources correspond to facets in a Voronoi tessellation, where the calibration solutions are assumed to be constant (Schwab, 1984; van Weeren et al., 2016b). We aim to achieve sub-arcsecond resolution images, which necessitates finding calibrators with high enough S/N ratios at the longest baselines. This is more challenging towards higher resolutions because fewer sources have sufficient flux densities at $0.3''$ scales.

In this section, we revisit the calibrator selection from de Jong et al. (2024) and connect it to a metric for determining solution intervals, aimed at enhancing automatic self-calibration. Additionally, we introduce a neural network to assess self-calibration convergence. This helps to refine the automation of calibration strategies by identifying the optimal self-calibration cycle and set of automatically determined parameters.

6.3.1. Calibrator selection

de Jong et al. (2024) implemented a source selection method that exploits the fact that circular polarisation is a rare phenomenon at low radio frequencies, as demonstrated by Callingham et al. (2023). Therefore, the difference between right- and left-handed polarisation should be minimal for compact high S/N calibrators.[8] Consequently, the circular standard deviation, accounting for phase wrapping around π (e.g. Mardia, 1972; Fisher et al., 1993), is applied to the solutions that correct for the

[5] `https://casacore.github.io/python-casacore/`
[6] `https://www.astropy.org`
[7] `https://dp3.readthedocs.io`
[8] This is a simplified assumption ignoring the effects of polarisation leakage variations and the minor impact of differential Faraday rotation for different directions.

phase difference between both polarisation directions. The resulting values for the most distant international stations (excluding Dutch and German stations) provide a measure of the S/N on the smallest angular scales, as these stations correspond to the longest baselines. Having enough S/N is essential to correct for DDEs. de Jong et al. (2024) selected a solution interval of $\hat{\delta}_t = 10$ min for these corrections, since this value worked well for all their sources. Having a fixed solution interval helps to compare the S/N across different sources within the same field of view. In this work, we refer to this metric with the phasediff-score (denoted by variable $\hat{\sigma}_c$). While de Jong et al. (2024) used a phasediff-score threshold of $\hat{\sigma}_c < 2.3$ rad, we opt to split our calibrators in two groups:

- ☐ *Main facet calibrators*: $\hat{\sigma}_c < 2.0$ rad.
- ☐ *Weak facet calibrators*: $2.0 \leq \hat{\sigma}_c < 2.6$ rad.

The main facet calibrators define our facet layout (see Section 6.4.2 and Figure 6.2), while the second group consists of weaker 'secondary' DD-calibrators, which can be used to refine DD calibration (see Section 6.4.4).

The threshold of $\hat{\sigma}_c = 2.0$, which separates strong and weak facet calibrators, is based on a re-evaluation of the calibration quality in de Jong et al. (2024). This work showed that the selected DD-calibrators with scores above this threshold tend to exhibit less stable calibration behaviour. Therefore, a more conservative approach ensures the selection of generally more stable main DD-calibrators, with the option to incorporate weaker secondary sources to enhance image quality by cleaning more local DDEs after facet subtraction (see Section 6.4.3). An additional minimal distance criteria between calibrator candidates prevents having neighbouring facets with calibrator sources near the edge of the other facet. This potentially leads to DD artefacts crossing the facet boundary (this will be later demonstrated in an example in Section 6.6.1.1) or to having sources obtaining higher $\hat{\sigma}_c$ values due to picking up flux from bright neighbours. DDEs do not vary as much on small angular scales, which made us decide to take 0.15 degrees as a minimal distance value. If two or more sources are neighbouring within these distance thresholds, the source with the lowest phasediff-score is kept and the other one(s) is removed from the selection. The calibration of weak facet calibrators is the final step of our data processing strategy (as will be outlined in Section 6.4.4). This is not required in order to perform the final wide-field imaging, when for instance these weak calibrator are near an already properly calibrated main calibrator, or when the ionospheric effects on the data are mild. Therefore, users can decide to skip this step or manually enhance it by, for example, adding additional sources.

For the datasets used from the ELAIS-N1 field, we removed 2 sources from the main facet calibrators because they were located within 0.15 deg of another

calibrator. From the weak facet calibrators, we removed 7 calibrators, since those were within 0.2 deg from another calibrator. This yields 23 strong calibrators next to the already calibrated in-field calibrator, next to 8 weak facet calibrators. This is different to the 30 calibrators selected by de Jong et al. (2024). The facet layout corresponding to the main DD-calibrators, along with the positions of the main and weak facet calibrators, is shown in Figure 6.2.

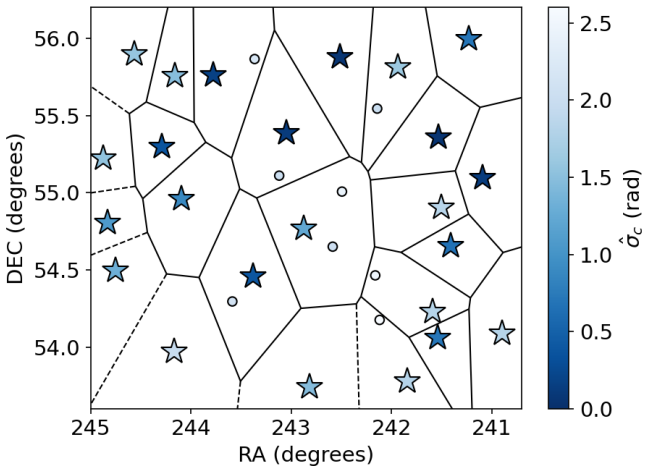

Figure 6.2: Main DD calibrator sources (stars) and weak facet calibrators (circles) for ELAIS-N1 with their respective phasediff-scores (σ_c) and corresponding facet boundaries in black.

6.3.2. Solution interval metric

Since the phasediff-scores are linked to the S/N of the longest international baselines, they also link to the optimal width of the solution interval to calibrate international stations. This enables us to derive automatically an appropriate value for the solution intervals for each calibrator source.

By varying the solution intervals between 1 min and 20 min for different calibrator sources when performing phase calibrations to correct for the differences between right-handed and left-handed polarisations, we empirically derived the following relation between the phasediff-scores (σ_c) and the corresponding solution intervals (δ_t):

$$\sigma_c \approx \pi \sqrt{1 - \exp\left(-\frac{\varsigma}{\delta_t}\right)}, \tag{6.1}$$

where ς is a unique constant value for each calibrator source in the data, which must be derived individually for each source and each observation. Increasing the solution intervals leads to a reduction in the variance of the phase solution data, while shorter intervals result in higher variance. The value for ς is a measure corresponding to the S/N of the source and therefore indicates the extent to which the solution interval can be narrowed to achieve a desired phase solution variance on the correction solutions.[9] Since the phasediff-scores are determined using calibration corrections on the data, we are effectively assessing the calibratability of the data and thus accounting for systematic effects, such as ionospheric variations unique to each observation. Consequently, Equation 6.1 provides a more precise method for determining solution intervals compared to theoretical approaches (e.g. Sob et al., 2021), which do not include direct information about the calibratability of a dataset.

For each source, we derive ς by evaluating $\sigma_c = \hat{\sigma}_c$ at a fixed solution interval of δ_t =10 min, such that we can derive a general expression for a metric for the solution intervals for different σ_c. By experimenting with self-calibrations of calibrator sources with different solution intervals, we found $\sigma_c = 1.75$ rad to correspond to δ_t values indicative of stable calibration. Using $\hat{\delta}_t = 10$ min and the fact that ς is constant, we derive

$$\begin{aligned} \delta_t &= \hat{\delta}_t \frac{\ln\left(1 - \left(\frac{\hat{\sigma}_c}{\pi}\right)^2\right)}{\ln\left(1 - \left(\frac{\sigma_c}{\pi}\right)^2\right)} \\ &\approx -26.92 \ln\left(1 - \left(\frac{\hat{\sigma}_c}{\pi}\right)^2\right). \end{aligned} \tag{6.2}$$

This expression functions as the metric for deriving solution intervals for different types of calibration (see Section 6.4.2). In Figure 6.3, we illustrate for two calibrator sources how the curve from Equation 6.1 fits well to simulated normally distributed data. The red star corresponds to the solution interval determined with Equation 6.2. The left source has a higher S/N than the right source, leading to a smaller δ_t for the left source compared to the right. The difference in S/N also accounts for the increased noise in the simulated data for the right source compared to the left source.

[9] This relation becomes particularly evident for $\delta_t \gg \varsigma$, where a Taylor expansion allows us to approximate the inverse square-root relationship $\sigma_c \propto \sqrt{\frac{\varsigma}{\delta_t}}$.

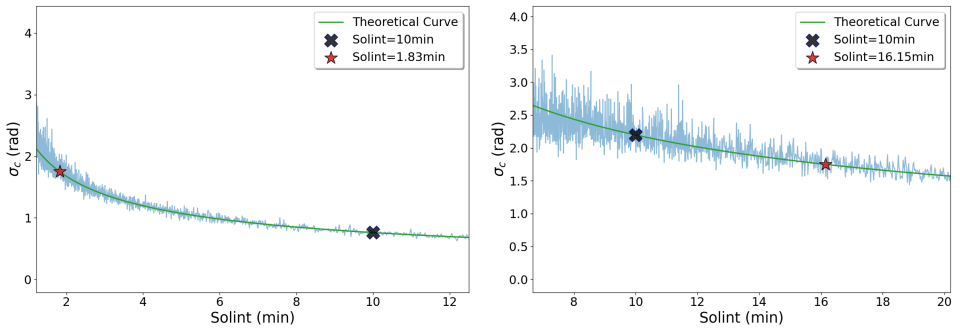

Figure 6.3: Phasediff-scores (σ_c) as a function of solution interval (δ_t) for two different calibrators. The black cross is the phasediff-score as calculated and described in de Jong et al. (Section 3.3.1; 2024) for $\delta_t = 10$ min, the green curve corresponds to the fit from Equation 6.1, and the red star is the best solution interval corresponding to $\sigma_c = 1.75$, according to Equation 6.2. The blue scattered line corresponds to determining circular standard deviations on randomly drawn normally distributed data with different solution intervals.

6.3.3. Automatic self-calibration assessment

By utilizing metrics that combine commonly used image metrics, such as the RMS background noise, the dynamic range, and by visually assessing images and solution inspection plots, Ye et al. (2024) and de Jong et al. (2024) found that many of the calibrator sources required fewer or in a few cases also more self-calibration cycles than the often selected 10 or 12 cycles. In certain instances, self-calibration even began to diverge after just 5 or 6 cycles. While the RMS and dynamic range provide useful information, they are not sufficient to fully determine if self-calibration has converged. For some calibrators, the RMS may for instance slightly increase during self-calibration when solving for amplitudes. In some cases, this is due to unstable diverging behaviour when a weaker calibrator absorbs signal from a strong nearby calibrator, while in other cases these are valid corrections to correct for small amplitude offsets. Similarly, the dynamic range may converge to favourable values during self-calibration, yet subtle image artefacts can persist and remain visible to the human eye. de Jong et al. (2024) also examined calibration solution stability, which indicates whether the calibration solutions converge across different cycles. While this confirms that self-calibration has indeed stabilised, it does not necessarily reflect whether the result is of sufficient quality, particularly if overly conservative self-calibration parameters are used.

To address this, we automate the decision-making process to remove the need for visual assessment. For this purpose, we adopt a pre-trained DINOv2 model (Oquab et al., 2023), utilizing register-based methods (Darcet et al., 2023) and with a vision transformer (ViT; Dosovitskiy et al., 2020) as the backbone. Specifically, we use

DINOv2 with the ViT-L backbone, comprising 300 million parameters, which has been trained using knowledge distillation on a larger DINOv2 model with a ViT-g backbone containing 1.1 billion parameters. The DINOv2 pre-training dataset (LVD-142) is a combination of existing curated datasets, including different ImageNet datasets (Russakovsky et al., 2015), Google Landmarks (Weyand et al., 2020), and images scraped from the internet by examining their similarity to the already existing images (without being duplicated).[10] The total combined dataset consists of 142 million images. The original DINOv2 model was trained in a self-supervised manner on this comprehensive dataset. We leverage this pre-trained model by substituting only the classification sub-model with a custom classifier. This classifier is a two-layer multi-layer perceptron (Rumelhart et al., 1986). In this classifier, a dropout is first applied to the feature extractor's output for regularisation. Subsequently, two linear layers are applied: the first layer, activated by a rectified linear unit (ReLU; Nair & Hinton, 2010), maps the feature extractor's output to the same dimensionality. The second layer reduces this output to a single value, followed by a sigmoid function to produce a pseudo-probabilistic output.

While the original DINOv2 model has mostly been trained on generic natural images curated from the internet, the self-supervised training scheme enables more generic and widely applicable features compared to feature extractors trained with a supervised training scheme (e.g. Huang et al., 2021). The primary advantage of transfer learning from the DINOv2 model is that, as long as the data modalities remain relatively similar, the extracted features are likely to be transferable (e.g. Gerace et al., 2022; Tahir et al., 2024). This allows us to achieve high performance by utilizing the well-trained feature extractor of the DINOv2 model, which is pre-trained on a large dataset, rather than training the feature extractor from scratch on a smaller dataset. By training a classifier that acts on these features (and potentially fine-tuning the feature extractor) on our domain-specific self-calibration images, we obtain a model that is robust for detecting artefacts in these images. This can be assessed by evaluating the model on test images that are unseen during the training stage. The effectiveness of utilizing DINOv2 as a base model on 'new' images has been demonstrated across various domains, including applications in medical imaging (e.g. Kundu et al., 2024; Song et al., 2024) and geological imaging (e.g. Brondolo & Beaussant, 2024).

We trained the model on 2,360 binary-labelled self-calibration images from different observations, including all calibrator sources considered by de Jong et al. (2024), Bondi et al. (in prep.), and Escott et al. (in prep.). These images include both high-S/N calibrator sources, weaker sources that were not selected for final DD

[10]This similarity has been determined using another pre-trained ViT, which embeds these images.

calibration, and sources that are resolved out and thus correspond to images dominated by noise. We further extend the dataset by applying data augmentations, specifically using random mirroring and rotations in multiples of 90 degrees. This enhances the model's capability to train on images associated with more unique and complex extended calibrator sources. The label *continue* ($P = 1$) is assigned to an image when significant artefacts and continued calibration is required. In all other cases, the images are labelled as *stop* ($P = 0$), which covers the following scenarios:

- ☐ Self-calibration has converged.
- ☐ The source is resolved out at high resolution, resulting in insufficient S/N for calibration.
- ☐ Artefacts are present but are not caused by the calibrator source in the image.

Figure 6.4 presents examples of images with labels from our training dataset. To improve regularisation we apply label smoothing as well (Szegedy et al., 2015). This is a technique that slightly reduces the *continue* class $P = 1$ to for instance $P = 0.9$ and adds the remaining probability (e.g., 0.1) to the *stop* class. This reduces overconfidence in the model's predictions, helping it generalise better by not focusing too strictly on exact class labels.

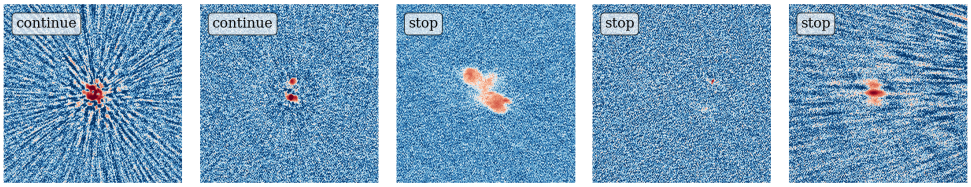

Figure 6.4: Images of sources at various stages of self-calibration. *Left:* Labelled *continue*, as no self-calibration has been applied and significant artefacts are present. *Middle left:* Labelled *continue*, after several cycles of self-calibration, though noticeable artefacts remain. *Middle:* Labelled *stop* since self-calibration has successfully converged. *Middle right:* Labelled *stop* because the source is resolved, and further self-calibration is unnecessary. *Right:* Labelled *stop* as artefacts from a nearby bright calibrator leak into the calibrator in the centre.

After training the model, we achieved an accuracy of 0.94, which corresponds to the confusion matrix given in Figure 6.5. This indicates that the model maintains high precision in correctly labelling the sources in our validation dataset and according to our visual inspections. Note, however, that the labelling of when to stop self-calibration remains subjective, as calibration is often never entirely perfect and some artefacts may persist in the image. Although expert labelling is involved, a certain level of bias is therefore inevitable.

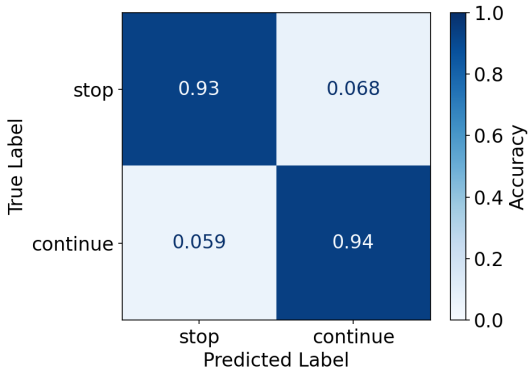

Figure 6.5: Confusion matrix with the fraction of true positives, true negatives, false positives, and false negatives.

Figure 6.6 displays examples of label predictions for calibrators across various self-calibration cycles. The sources in the top three rows all achieve convergence by the 5th cycle, with P values below 0.5. In each of these cases, we find that the model recognises subtle residual amplitude artefacts (4th column) that could still be corrected. In the fourth row, convergence is reached by the 3rd cycle, after which the label briefly increases, before reaching its lowest value in the 5th cycle. While the image shows notable improvement compared to the uncalibrated version, this example highlights the subjectivity of the labelling, as other astronomers might still find minor amplitude artefacts around the hotspots of this FRII source (Fanaroff & Riley, 1974) unsatisfactory. In the fifth row, we observe a source gradually approaching convergence by the 5th cycle, though it has not yet met the early-stopping criteria for self-calibration. The final row illustrates a point source that fails to converge to an artefact-free image, although the model detects slight improvements compared to the initial uncalibrated image.

Since the model is trained purely to classify individual images, it does not account for amplitude drifts or information from prior self-calibration cycles. Therefore, we set our self-calibration stopping criteria on the requirements that the image is labelled with $P < 0.5$, an RMS increase compared to the initial uncalibrated image of no more than 5%, and an improvement in dynamic range compared to the uncalibrated image. We also require the solutions to be stabilised between cycles as well. This is determined by subtracting the current phase solutions from the previous one and calculating the circular standard deviation of the difference, which must be less than 0.1 rad. This threshold is based on the results from the self-calibration inspections from de Jong et al. (2024).

Our stopping criteria not only reduce the computational cost of self-calibration but also aid in identifying optimal calibration settings for our sources. This parameter optimisation allows us to fine-tune the automatic settings, as described in Section 6.4.2. A potential future application is to integrate these automatic self-calibration stopping criteria to dynamically adjust parameters during self-calibration, using the different stopping criteria to decide to refine settings after each cycle. As this approach is still in its early stages and requires validation of data from other observations, further potential applications are discussed in Section 6.7.2.

6.4. Improved DD calibration and imaging strategy

We follow the same data reduction strategy up to the primary in-field calibration as described in de Jong et al. (2024, Section 3.1 and 3.2). We only replaced the software from the first calibration step using `prefactor` (van Weeren et al., 2016b; de Gasperin et al., 2019b) by its successor the LOFAR Initial Calibration (`LINC`[11]). `LINC` uses CWL and therefore integrates well into our data processing framework. In this section, we detail DD calibration and imaging improvements, aimed at enhancing image quality and reducing computing costs to enable scalable data reduction for ultra-deep imaging. The full new data processing strategy discussed in this Section is presented in Figure 6.7 and will be referred to throughout the text.

The different calibration steps utilise `facetselfcal`[12] (van Weeren et al., 2021), which is an advanced flexible low-frequency self-calibration software package. This software integrates the Default Preprocessing Pipeline (`DP3`; van Diepen et al. 2018a; Dijkema et al. 2023) and `WSClean` (Offringa et al., 2014) for (self-)calibration on a source. `facetselfcal` executes multiple self-calibration cycles, adjusting the source model iteratively before proceeding to the next cycle.

6.4.1. Dutch station calibration

For the final imaging, de Jong et al. (2024) created datasets for each individual facet, as this allows to perform parallel imaging across facets and reduces overall computing wall-time. Before imaging, sources outside each facet were subtracted using $1.2''$ resolution model images with corresponding calibration solutions. However, during imaging, they found that the DD calibration solutions for the Dutch core and remote stations were of poor quality at lower resolutions. To address this, an additional calibration refinement step was introduced on the datasets corresponding to each facet, selecting only the Dutch core and remote stations. While

[11] `https://linc.readthedocs.io`
[12] `https://github.com/rvweeren/lofar_facet_selfcal`

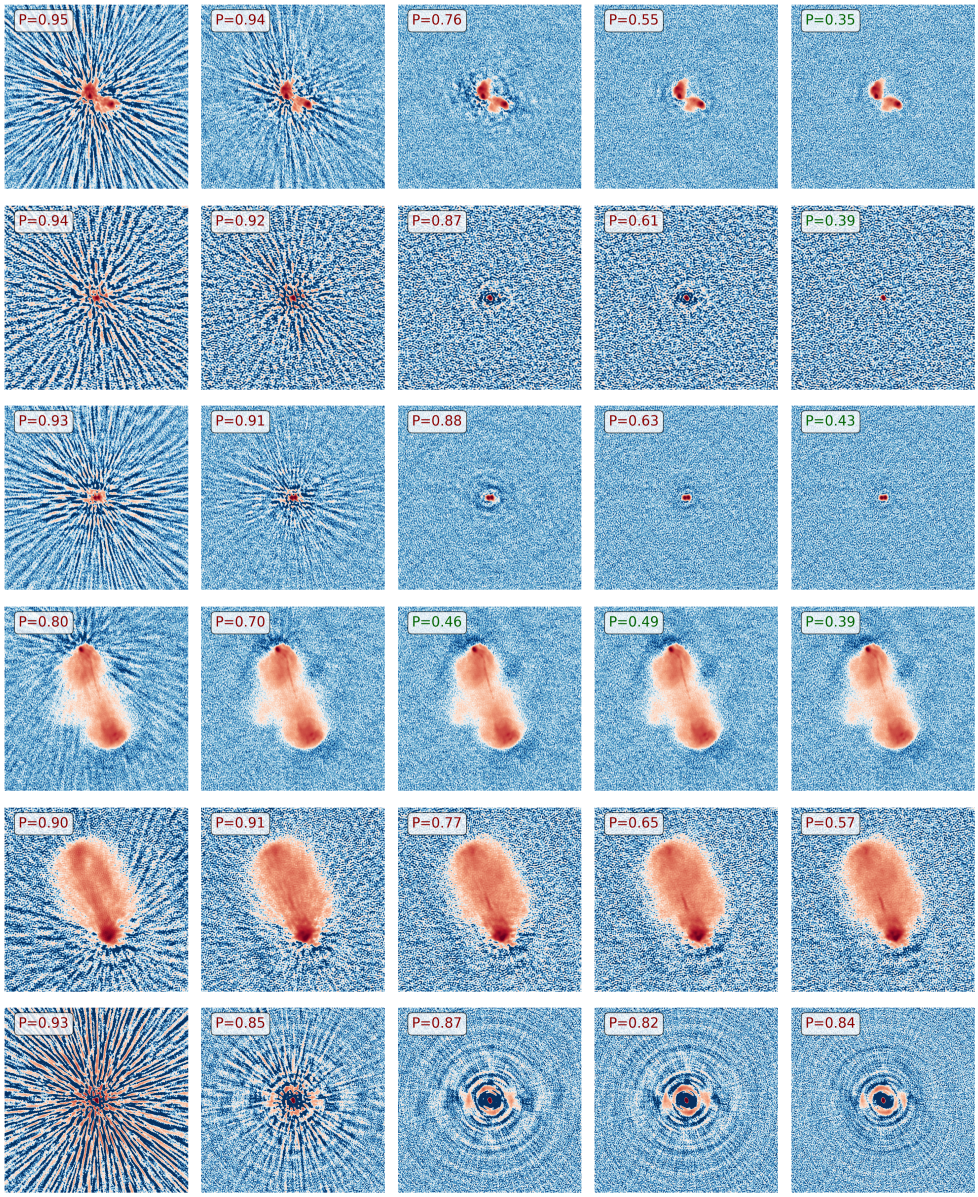

Figure 6.6: Examples of self-calibration images with early-stopping scores from our neural network. The *rows* represent different sources, while the *columns* correspond to self-calibration cycles. $P < 0.5$ values in green indicate successful convergence and will make self-calibration stop, while $P \geq 0.5$ in red suggests poor image quality, indicating that further self-calibration is required.

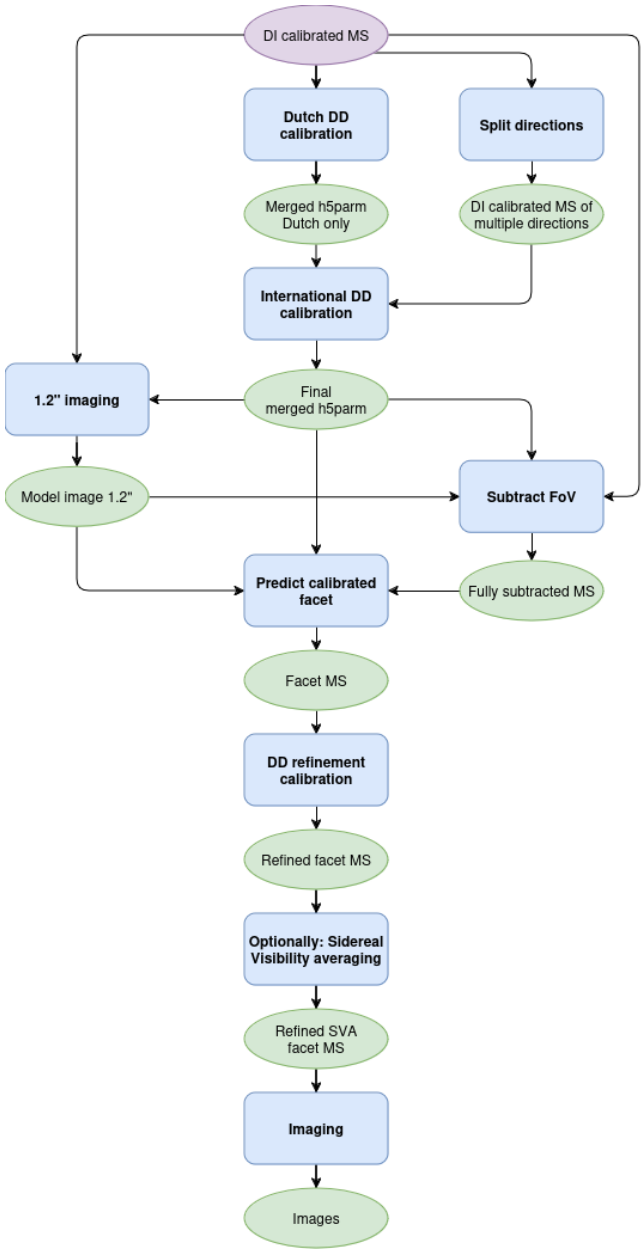

Figure 6.7: Flowchart corresponding to the full DD calibration discussed in Section 6.4. The workflow starts with DI-corrected *uv*-data and ends with images. Purple ovals are input data, blue boxes are operations on the data, and green ovals are output data.

this improved image quality around the calibrator source of the facet, it did not eliminate DDEs originating from neighbouring facets. This was because the refinement was applied after splitting the facets and subtracting external sources using the 1.2″ resolution model images, which were created with poor Dutch calibration solutions. Hence, it is essential that poor Dutch calibration solutions are corrected before any subtraction is performed.

To enhance the calibration solutions for short baselines, we introduce an additional DD calibration step for the Dutch core and remote stations before performing subtraction of sources outside the facets. We already obtained DD solutions from the `DDF`-pipeline[13], which utilises `KillMS`[14] (Tasse, 2014a,b; Smirnov & Tasse, 2015) to obtain phase and amplitude corrections. However, we have already corrupted our data with the DI in-field calibration solutions for both the Dutch and international stations, which makes pre-applying Dutch DD solutions from the `DDF`-pipeline to the closest calibrator selected in Section 6.3.1 no longer valid. An alternative approach for incorporating the `DDF`-pipeline DD solutions is discussed in Section 6.7.1.

We opt instead for using a new joint-solve feature from `facetselfcal`, which performs self-calibration in multiple directions simultaneously using `DP3`. For this, we average our datasets to 16 sec and 195.36 kHz after removing the international stations. This reduces the data volume by a factor ∼300 and therefore reduces the computing resources required for a multi-directional joint-solve while having enough time and frequency resolution to calibrate for the changing ionospheric effects. Recent updates in `DP3` and `WSClean` have also enabled Stokes I-only data processing, which implies that 4 times less data has to be processed, saving a significant amount of RAM and computational time. Since our selection (see Section 6.3.1) is based on the S/N for the international stations, we select calibrators that have a peak intensity above 85 mJy beam^{-1} in the 6″ ELAIS-N1 catalogue from Sabater et al. (2021), as we find this threshold to correspond to stable calibrators. This leaves us with 15 calibrators which we calibrate with the following strategy:

1. We first perform phase calibration focused on correcting the Dutch remote stations by taking solution intervals of 32 seconds for sources with peak intensities below 300 mJy beam^{-1}, and 16 seconds for sources with peak intensities above this threshold. We also use a frequency smoothness kernel of 20 MHz, which from experience has shown to result in stable solutions at 6″. After this step, we reset the Dutch core station solutions to amplitudes equal to 1 and phases to 0, such that only the Dutch remote stations get calibration corrections.

2. We then carry out phase calibration for slower phase variations from the Dutch

[13] `https://github.com/mhardcastle/ddf-pipeline`
[14] `https://github.com/saopicc/killMS`

core stations, using longer solution intervals of 64 seconds and a broader frequency smoothness kernel of 40 MHz. This approach works well for the Dutch core stations because their shorter baselines are less sensitive to rapid phase variations and primarily capture larger-scale structures.

3. After doing 3 cycles, we have calibrated the phases well enough to also include calibration for the combination of phases and amplitudes as well. This applies longer solution intervals, compared to the previous phase calibration steps, since phases tend to vary on much shorter time scales compared to amplitudes. The solution intervals are 40 min for sources with peak intensities below 300 mJy beam^{-1}, while 20 min for sources with peak intensities above this threshold. We use a frequency smoothness kernel of 10 MHz.

All calibration steps are polarisation-independent, as polarisation corrections have already been applied through the `DDF`-pipeline full-Jones DI solutions and during a full-Jones in-field calibration step (see Section 3.2.3; de Jong et al., 2024). We also use a *uv*-cut of 10λ which corresponds to largest angular scales of about ~5.7 deg. The used settings are optimised on an empirical basis for ELAIS-N1. We expect these to work well in the general case, but for further optimisation, these settings can be adjusted, using a configuration file. This step corresponds to the 'Dutch DD calibration' box and its merged output solutions in Figure 6.7.

6.4.2. International station calibration

With the improved Dutch DD calibration solutions, we proceed to calibrating the international stations. We first create datasets of each of the 24 calibrator sources, selected by the metrics in Section 6.3.1, by phase-shifting to the centre of the source. We then average the datasets to 32 sec and 390.72 kHz. These averaging settings were shown by de Jong et al. (2024) to be effective in averaging out signal from nearby sources, while providing sufficient time and frequency resolution to correct fast phase variations for the brightest calibrator sources, without introducing bandwidth or time smearing in the self-calibration images. Subsequently, we apply on each dataset the Dutch calibration solutions from the nearest of the Dutch calibrators from Section 6.4.1. To suppress the signal of nearby sources at short baselines and reduce the computational cost, we adopt, similar to previous works (e.g. Moldón et al., 2015; Morabito et al., 2022a), a phase-up of the Dutch core stations into a superstation. Since the calibration solutions for the Dutch core and remote stations have already been applied, the goal is now fully focused on calibrating the international stations. For self-calibration, we use again `facetselfcal`.

During our analysis and testing, we observed that when performing self-calibration with all stations, including the Dutch phased-up stations, the Dutch core and re-

mote stations drifted, despite having their initial optimised calibration solutions applied, to the same poor solutions previously identified in de Jong et al. (2024). This drift is likely caused by flux from other high S/N sources in the field leaking into the signal from the target, thereby contaminating and corrupting the self-calibration process. This effect is particularly pronounced on the short baselines from the Dutch stations, despite the use of the phase-up of the core to partially mitigate this effect and aid convergence of the calibration for the international stations (Morabito et al., 2022a). While alternative solutions exist, by for instance drawing boxes around calibrators and predicting and extracting all sources outside these boxes (e.g. van Weeren et al., 2021; de Jong et al., 2022), these methods are computationally far too expensive for our large data volumes. Instead, we need to address this by ensuring that the already obtained solutions for the Dutch core stations and some of the Dutch remote stations change as little as possible after each self-calibration cycle and by applying large *uv*-cuts of $\geq 20,000\lambda$, corresponding to angular scales smaller than $10.3''$. These baseline cuts correspond to smaller angular scales, more compact source models during self-calibration, and focus on corrections for the longest baselines.

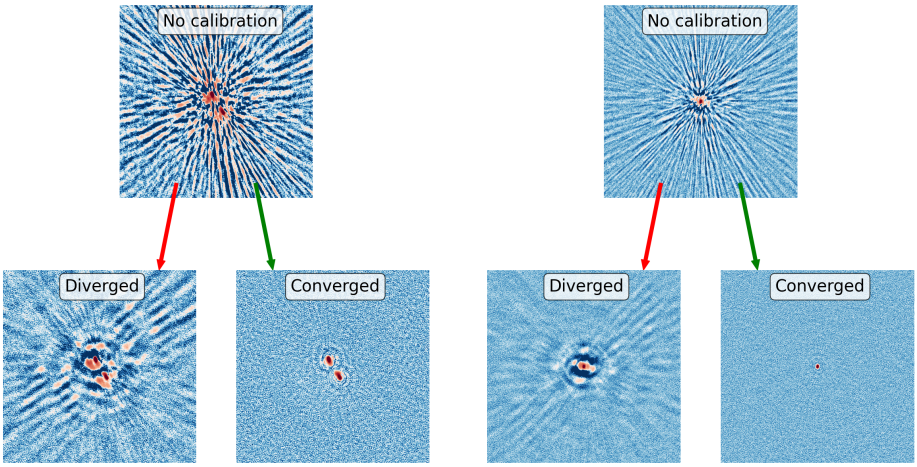

Figure 6.8: Self-calibration stability for various calibration parameters. The top row displays two images of uncalibrated sources. Below, a left and right image illustrate the corresponding images after performing self-calibration on the uncalibrated sources, where the calibration either diverged or successfully converged based on different parameter settings. The difference in the settings between the diverging and converging images was dependent on changing the *uv*-cut settings and by incorporating an additional round of polarisation-independent phase calibration for stations and resetting the Dutch core and a few additional remote stations, as we further explain in Section 6.4.2.

Starting with the parameter settings from de Jong et al. (2024) and utilizing

the solution interval metric δ_t (Equation 6.2) along with the image quality-based stopping criteria from the neural network discussed in Section 6.3.3, we optimised our calibration settings for different calibrator sources. To preserve the good calibration solutions from the DD Dutch calibration, we first simply reset all Dutch calibration solutions to the original pre-applied solutions obtained in Section 6.4.1, after self-calibration. However, instabilities during calibration and the emergence of spots with negative signals around compact high S/N sources highlighted the need to allow more flexibility for Dutch stations in adjusting their solutions. This is likely due to the fact that the Dutch stations have only a quarter of the collecting area compared to the international stations. Hence, for the most compact sources, the self-calibration process may run into the situation that S/N gets too low for the short baselines from the Dutch stations, while still being sufficient for calibrating the long baselines with the international stations. Two examples of different diverging and converging behaviour during self-calibration for different parameter settings are presented in Figure 6.8 and demonstrate the importance of careful parameter tuning.

The self-calibration divergence issues led us to explore an approach that involves calibrating with multiple cycles for different station groups linked to the brightness of the source. The options from `facetselfcal` provide the possibility for calibrating specific stations by resetting the solutions from station groups after each self-calibration cycle. The reset ensures that solutions for these stations are unchanged compared to the starting solutions. This provides a workaround for the limitation of the `DP3` solvers, which cannot apply different time intervals for different stations. This enables us to obtain accurate calibration solutions while minimizing the degrees of freedom by reducing the number of solutions needed to correct the data effectively. After experimenting with various parameter settings in a grid search and selecting the best-performing configuration based on the automatic image and solution quality-based stopping criteria, we identified the following strategy as the most effective:

1. We perform the first calibration for fast wrapping phases of the most distant international stations by calibrating these using solution intervals with size $\sqrt{\delta_t}$ (using Equation 6.2) and frequency smoothness kernels of 8 MHz. The solution intervals have a minimal size equal to the time resolution of the datasets at 32 sec and a maximum size of 3 min. After this calibration step, we reset the solutions of all Dutch core and remote stations and German stations closest to the Dutch stations (DE601HBA and DE605HBA).

2. With the first phase solutions for most of the international stations, we add another round of phase calibration but now with larger solution intervals with

size $\sqrt{2\delta_t}$ and frequency smoothness kernels of 10 MHz. The solution intervals have a minimal size equal to the time resolution of the datasets at 32 sec and a maximum size of 5 min. After calibration, we reset the solutions for all Dutch core and remote stations.

3. The following step only continues for sources with solution intervals $\delta_t < 3$ min, where we add another round of phase calibration for a specific (sub-)group of Dutch remote stations. Since this step involves solving for some of the Dutch stations, which correspond to shorter baselines, we use larger solution intervals and frequency smoothness kernels compared to the previous steps. The solution intervals are set to $2\sqrt{\delta_t}$ and the smoothness kernel to 15 MHz. For the brightest sources, with $\delta_t < 0.3$ min, the best results were obtained by resetting only the solutions of the Dutch core stations, which implies that all remote stations can be freely adjusted during self-calibration. For sources with $0.3 \leq \delta_t < 1$ min, we limit the adjustments for remote stations by resetting the solutions of all Dutch core stations and the five remote stations closest to the Dutch core (RS106HBA, RS205HBA, RS305HBA, RS306HBA, RS503HBA). Lastly, for sources with $1 \leq \delta_t < 3$ min, we reset the same remote station, including a few more distant remote stations (RS307HBA, RS406HBA, RS407HBA). This approach helps balance image quality and solution stability by including only specific groups of Dutch remote stations, selected based on their solution interval, which directly relates to their S/N.

4. Finally, after performing 3 cycles with only phase calibration, we also calibrate for both phases and amplitudes together but with larger solution intervals of $20\delta_t$ with a minimum of 18 min, since phases vary on much shorter time scales than amplitudes. This step is entirely skipped when $20\delta_t > 4$ hrs. The smoothness kernels depends on the δ_t, as we set this to 8 MHz if $\delta_t < 1$ min, 10 MHz if $1 \leq \delta_t < 3$ min, and 12 MHz otherwise. We reset in this step the solutions from all Dutch core stations and the same specific sub-group of remote stations as for the previous step if $\delta_t \geq 3$.

All these calibration steps are polarisation-independent, as we have already performed polarisation corrections during in-field calibration (see Section 3.2.3. in de Jong et al. (2024)). Since calibration for the Dutch stations has already been performed, the short Dutch-only baselines have already been corrected, which allows us to use large *uv*-cuts to focus on calibrating against a more compact sky model corresponding to long baselines (e.g. the international stations). We found by experimenting, using the automatic self-calibration quality stopping criteria, that varying the *uv*-cut based on their S/N led to better results as well. This is likely because when a source has a high S/N at long baselines, it is sufficient to construct

the sky model during self-calibration with strong signal at smaller angular scales. By doing so, we mitigate the effects of signal from other high S/N sources in the field leaking into the target source, which we already identified to corrupt the calibration solutions in particular at shorter Dutch baselines. While for less compact high S/N calibrators, we need more baselines with enough signal to reach convergence, which requires a smaller *uv*-cut. Linked to the freedom we give to the Dutch remote stations to vary in the 3rd step of our calibration strategy, we found the following relation between the *uv*-cut and δ_t to work well:

$$uv\text{-cut} = \begin{cases} 40,000\lambda & \text{if } \delta_t < 0.3 \\ 35,000\lambda & \text{if } 0.3 \leq \delta_t < 1 \\ 25,000\lambda & \text{if } 1 \leq \delta_t < 3 \\ 20,000\lambda & \text{if } \delta_t \geq 3 \end{cases}$$

A *uv*-cut of 20,000λ corresponds to a largest angular scale of approximately 10.3″, while a *uv*-cut of 40,000λ excludes most remote-remote baselines and corresponds to a largest angular scale of around 5.2″. Thus, even though we do not reset the remote stations for the brightest sources in step 3, we constrain their source model by applying a larger *uv*-cut to focus on correcting smaller angular scales. In this way, we balance, based on the S/N at the longest baselines, the freedom of the Dutch remote stations to be adjusted with a constraint on the baseline length.

For deciding on early-stopping during self-calibration, we utilise the neural network in combination with the other quality metric assessments discussed in Section 6.3.3. We use between 3 and 20 self-calibration cycles, providing a suitable range to ensure self-calibration convergence before proceeding to the next step in the data reduction process. In Section 6.4.4, we further refine calibration and provide a larger maximum number of self-calibration cycles to reach complete convergence. Unlike de Jong et al. (2024), we do not need to perform self-calibration with all observations combined since this will be carried out in the final self-calibration step. This allows in this intermediate calibration step for embarrassingly parallel processing over each calibrator source and each observation, reducing the wall time when enough CPU cores are available. The step discussed in this Section corresponds to the 'International DD calibration' box and its merged output solutions in Figure 6.7.

6.4.3. 1.2″ facet subtraction

Producing a typical 0.3″ wide-field image requires a wall-time of about 1.5 months per 8-hrs observation when using the 'standard' facet-mode from `WSClean` on a

single node. To reduce wall-time, we divide, similar to Sweijen et al. (2022c) and de Jong et al. (2024), the full dataset into smaller subsets, each corresponding to a single facet, enabling parallel processing of all facets across an HPC cluster up to the final single-facet imaging stage. This approach enables more data averaging and therefore accelerates imaging through parallelisation, while also allowing for SVA as this can only be performed with fully calibrated facets across multiple observations (see Section 6.4.5). To avoid calibration errors introduced by sources outside of the facet, it is crucial to subtract all sources that do not correspond to the current facet. de Jong et al. (2024) associated this step as part of the final imaging, as it utilises imaging software for prediction of sources within the field of view and is the final part of the pre-processing stage before the actual imaging. This step accounted for 76% of the total imaging costs, which represents 62% of the overall computational costs, including all calibration steps. The remaining 24% of the imaging costs were attributed to imaging the facets using all observations combined, as discussed in Section 6.4.5. This implies that reducing the computational cost at this stage can significantly lower the overall computational cost of the total data processing.

To enhance the computational efficiency of the source subtraction step before imaging, we introduce an improved method to make datasets of each facet. Instead of copying the full datasets for each facet and masking a facet and subtracting all sources outside this facet, we first create a dataset for each observation where we subtract all sources in the entire field of view. For this subtraction step, we use the improved $1.2''$ resolution model images of each observation, which result from imaging with our improved Dutch calibration solutions. This 'empty' dataset is then copied for each facet, whereafter we add back the sources corresponding to that facet, as illustrated in Figure 6.9. This approach eliminates the need to repeatedly predict the same sky, as was done with the original method from de Jong et al. (2024). All these steps can be done in parallel for each facet, each frequency subband, and each observation. To minimise I/O overhead, the data volume may be further reduced using stricter Dysco compressions (Offringa, 2016), as outlined in Section 6.6.2.1. In Section 6.6.2.2, we also discuss the significant computational cost improvements from our new approach. The subtraction steps correspond to the 'Subtract FoV' and 'Predict calibrated facet' boxes and their corresponding output in Figure 6.7.

6.4.4. Final calibration refinement

Since the sources with low phasediff-scores ($\sigma_c < 2.0$ rad), indicating high S/N, are now contained in separate datasets corresponding to individual facets, we have mitigated the interfering effect that high S/N calibrators have on each other when they are not corrected for DDEs and are contained in the same dataset (see Section

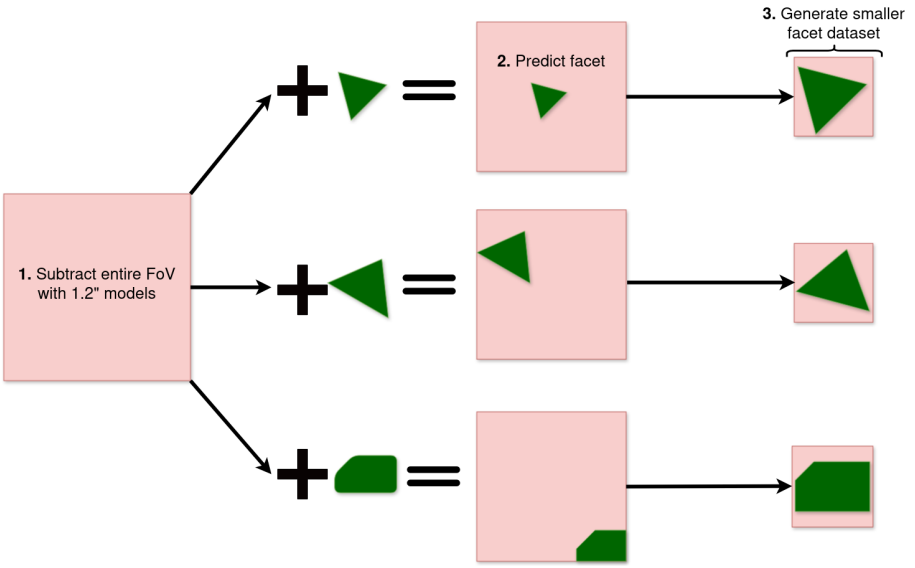

Figure 6.9: Updated subtraction and prediction strategy. The first large box represents a fully subtracted dataset, which has been created after predicting and subtracting sources with a DD-corrected 1.2″ model image. Next, sources within a polygon-shaped region are predicted and added back, creating a dataset with sources inside the facet while sources outside are subtracted. Finally, the data is phase-shifted to the centre of the polygon, and after applying beam corrections, solutions, and averaging, the smaller facet data is imaged.

6.4.2). This enables us to apply a final calibration step to improve calibration accuracy for all baselines and all observations together.

Similar to de Jong et al. (2024), we calibrate the Dutch stations by performing self-calibration at 6″ on the entire facet. Doing this at 0.3″ would in theory also be possible but is currently computationally too expensive, while at 6″the data can be averaged by a factor of 8 in both time and frequency and the international stations can be ignored, which reduces the computational cost significantly. Given that for facets with calibrators with solution interval metrics of $\delta_t > 3$ min the corrections on the Dutch and remote were left unchanged (see Section 6.4.2), we do not have to perform this step for those facets. Adopting the same recipe as de Jong et al. (Section 3.3.5; 2024), we proceed as follows, using a uv-cut of 750λ.

1. The first calibration step focuses on solving for 'fast' phase changes for the Dutch remote stations, using a solution interval of 1 min and a frequency smoothness kernel of 10 MHz, after which we reset the Dutch core stations.

2. Next, we apply another solve for phases, using a solution interval of 5 min and a larger frequency smoothness kernel of 20 MHz, but without resetting solutions.

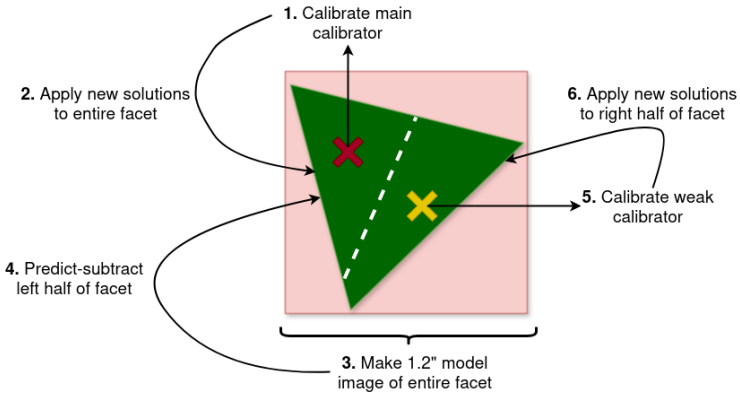

Figure 6.10: Additional facet calibration refinement for facets with weaker calibrators with low phasediff-scores ($2.0 \leq \sigma_c < 2.6$ rad). This calibration involves several steps. First, the main facet calibrator is calibrated using all observations combined. The derived solutions are then applied to the facet, and a $1.2''$ model image of the entire facet is created. With this model, the left half of the facet is predicted and subtracted, where both halves are defined with a Voronoi tessellation. This allows for the calibration of a secondary, weaker calibrator, after which the new solutions are applied to the remaining half of the facet. Finally, both halves of the facet are imaged.

3. After having done 2 rounds of self-calibration with only phase calibration, we perform a phase and amplitude correction with a solution interval of 30 min and a frequency smoothness kernel of 15 MHz as well.

After pre-applying the resulting solutions to the data, we employ a final round of self-calibration, where we first perform phaseshifts back to the centre of the calibrator source corresponding to each facet before averaging the datasets back to 32 sec and 390.72 kHz. Our calibration strategy follows a similar approach to what we described in Section 6.4.2, but with slightly modified parameters and different station resets.

1. We begin by calibrating the fast-wrapping phases of the international stations using solution intervals of size $\sqrt{\delta_t}$ and frequency smoothness kernels of 5 MHz. The solution intervals range from a minimal size equal to the time resolution of the datasets at 32 sec to a maximum size of 3 min. After this calibration step, we reset the calibration solutions for all Dutch core and remote stations.

2. We now perform phase calibration with all stations together, using a solution interval of $2\sqrt{\delta_t}$, and a frequency smoothness kernel of 20 MHz. The solution intervals range from a minimum of 64 sec to a maximum of 10 min.

3. Finally, after performing two cycles of phase-only calibration, we proceed with calibrating both phases and amplitudes using solution intervals of size $20\delta_t$, with a minimum interval of 18 min. This step is skipped if $20\delta_t > 4$ hrs. The smoothness kernel depends on δ_t, set to 5 MHz if $\delta_t < 1$ min, 10 MHz if $1 \leq \delta_t < 3$ min, and 15 MHz otherwise.

All steps are performed using polarisation-independent calibration, with the uv-cut constrained to 20,000λ. This helps to avoid potential issues from an incomplete high-resolution sky model, as remaining sources within the full facet but outside the smaller imaging box are not imaged or included in the model. We assess self-calibration convergence using our neural network model in combination with the convergence criteria described in Section 6.3.3, setting a minimum of 5 cycles and a maximum of 20. This is the first step where for deep imaging, the data from all observations are combined together, ensuring proper alignment before final imaging. This alignment is essential for correcting amplitude scale differences and preventing image blurring from small astrometric offsets between observations.

Finally, there is an optional step to include calibration of the weaker 'secondary' DD-calibrators, which were defined in Section 6.3.1 as sources with phasediff-scores of $2.0 \leq \sigma_c < 2.6$ rad. These sources may not have high enough S/N to have their own facet during the previous stages of the calibration, but they may still suffer from DDEs that can be corrected through an additional refinement step. This depends largely on the severity of the ionospheric conditions, which varies per observation. However, because the facet datasets are smaller at this stage, these refinement steps are significantly less computationally expensive than earlier calibration and prediction steps. In this refinement, we first create a $1.2''$ mode image of the entire facet and apply a Voronoi tessellation, using the positions of the main facet calibrator (which has already been corrected and solutions have been applied to the data) and the weaker calibrators. The area corresponding to the main facet calibrator is predicted and subtracted, followed by a phase-shift to the weaker calibrator before further calibration, using the same method as for the main facet calibrators, skipping the $6''$ full facet calibration. This process splits the facet into two (or more) sub-facets, where the weaker calibrator benefits from the pre-applied solutions of the primary calibrator, allowing for further refinement of local DDEs. Figure 6.10 illustrates this process, beginning with the calibration of the primary calibrator and showing the subsequent steps leading to the final calibrated data. In the case of multiple weaker calibrators, this can be easily extended to multiple facet splits, where all weaker calibrators have their own local solutions. The entire calibration refinement step, as described in this subsection, corresponds to the 'DD refinement calibration' box and its dataset output in Figure 6.7.

6.4.5. Final imaging

Imaging with multiple observations of the same field is typically done by concatenating the time axis from all observations, leading to an increase in data volume proportional to the number of observations. Consequently, the computational cost of imaging increases approximately linearly with the number of observations. However, we can exploit the fact that baseline tracks repeat every sidereal day. Since each observation is conducted on a different sidereal day, we can apply SVA, as described by de Jong et al. (2025) after having fully calibrated the facet datasets. They highlight that when combining observations separated over several years and conducted at different times of the year, offsets in the baseline coordinates must be accounted for by selecting a higher time resolution than the input datasets. These offsets are influenced by celestial motions, including aberration, precession, and nutation. The algorithm includes functionality that determines the optimal time resolution by considering image size, resolution, and baseline coordinate offsets. This minimises time smearing while significantly reducing the data volume compared to the total input. Although frequency effects like Doppler shifts are present, de Jong et al. (2025) found these to be negligible for ELAIS-N1 and therefore not a concern. The imaging step corresponds to the optional 'sidereal visibility averaging' and 'imaging' boxes with the images as output product in Figure 6.7.

For the final imaging, we follow the same approach as de Jong et al. (2024), using `WSClean`'s `wgridder` module (Offringa et al., 2014; Arras et al., 2021; Ye et al., 2022), with Briggs weighting set to -1.5 (Briggs, 1995). We adopt a minimum uv-value of 80λ and use pixel sizes of $0.1''$, $0.2''$, and $0.4''$, with corresponding Gaussian tapers of $0.3''$, $0.6''$, and $1.2''$. For efficient cleaning, we employ 'auto' masking, multi-scale deconvolution, and set an RMS box size 50 times the synthesised beam (Cornwell, 2008; Offringa & Smirnov, 2017). We apply a correction to account for the primary beam attenuation as well. We also use baseline-dependent averaging (BDA; e.g. Cotton, 1986, 2009; Wijnholds et al., 2018) during imaging.

6.5. Data

Observ. ID	Project	Calibrator ID	Observ. date	Pointing centres	Stations (int)	Frequencies
`L686962`	`LT10_012`	`L686958`	26-11-2018	16:11:00, +54.57.00	49 (11)	120-166 MHz
`L833466`	`LT14_003`	`L833474`	09-10-2021	16:11:00, +55.00.00	50 (13)	118-166 MHz

Table 6.1: Metadata from the 2 ELAIS-N1 observations used in this work. Observations used calibrator 3C 295 for calibration with `LINC`. The number of stations and frequencies are recorded after flagging. International stations are given between brackets.

ELAIS-N1 is a famous deep field, which has been explored across various wave-

lengths. This includes X-rays (e.g. Manners et al., 2003), ultraviolet (e.g. Martin et al., 2005), optical (e.g. McMahon et al., 2001; Aihara et al., 2018), infrared (e.g. Lawrence et al., 2007; Mauduit et al., 2012), and radio (e.g. Sirothia et al., 2009; Ocran et al., 2020; Best et al., 2023). This extensive multi-wavelength coverage has established ELAIS-N1 as an important field for extragalactic studies. Over 500 hrs of observing time are available for this field (Shimwell et al., 2025). About 200 hrs of these observations include only a few international stations or are heavily averaged, making them unsuitable for sub-arcsecond imaging. However, with hundreds of hrs of data still remaining, this field presents a great opportunity to produce the deepest LOFAR image to date. With the work done by Ye et al. (2024) and de Jong et al. (2024), we also have extensive prior knowledge of this field for calibrating and imaging data at arcsecond and sub-arcsecond resolutions. This gives us an advantage in experimenting with different calibrator sources and automated imaging settings to enhance the already existing images.

For the purpose of this work, we have selected 2 datasets from the ELAIS-N1 deep field to serve as test cases for our upgraded data reduction strategy. These datasets are detailed in Table 6.1. The first dataset, `L686962`, is selected from project `LT10_012`, representing their best observation and for us a benchmark for validating the improved calibration and imaging strategies. We used the already processed dataset up to the in-field calibration. The second dataset, `L833466`, comes from project `LT14_003` and features data recorded with a 2-sec integration time, compared to the default 1-sec, which leads to additional smearing at the edges (de Jong et al., 2024). Additionally, this observation was taken closer to solar maximum and shows stronger ionospheric effects compared to `L686962`, which is reflected in its `LINC` solution plots. This makes it an ideal candidate to test the robustness of our upgraded strategy in processing challenging observations. By combining these two datasets, we also evaluate the effectiveness of joint-calibration across multiple observations (see Section 6.4.4) and assess the impact of SVA (see Section 6.4.5).

6.6. Results

We have in this work, implemented enhancements in calibration quality and computational efficiency, enabling ultra-deep imaging at sub-arcsecond resolutions with LOFAR. To demonstrate these improvements, we have applied the improvements on two datasets. In this section, we highlight the resulting improvements in terms of calibration and computing costs.

6.6.1. Image quality improvements

The calibration strategy has been improved by addressing challenges highlighted in previous work (de Jong et al., 2024), and incorporating the solution interval metric and neural network described in Section 6.3. This has led to significant improvements in image quality, as outlined in this subsection.

6.6.1.1. Facet boundary leakage

In Section 6.4.1, we introduced an additional step to perform DD calibration for the Dutch core and remote stations before proceeding with DD calibration for the international stations. This improves the Dutch calibration solutions, which is in particular important for producing the $1.2''$ model images required for subtracting sources outside each facet before imaging (see Section 6.4.3). To illustrate the improvements in the $1.2''$ resolution image, Figure 6.11 shows a cutout of a challenging region. In the left panel of this figure, DD artefacts from two neighbouring calibrators were previously leaking into each other's facets. It is clear that in the new image in the right panel, this effect has been significantly reduced, which enables performing an improved subtraction of source signal at short (Dutch) baselines. This improvement also facilitates the final calibration refinement step (see Section 6.4.4) by minimizing interference from residual artefacts leaking from neighbouring facets.

6.6.1.2. Facet refinements

With better source subtraction before splitting out datasets for individual facets of our full mosaic, we mitigate the negative effect of high S/N sources affecting the self-calibration of the calibrators corresponding to each facet. This allows us, as outlined in Section 6.4.4, to perform a final self-calibration refinement step.

We tested the final calibration step for both the Dutch and international stations, with and without phasing up the Dutch core stations. For some facets, we found that the phase-up was not required, as most of the bright calibrator sources had been removed from the data, minimizing their impact on the calibration of shorter baselines. However, in a few cases, sources within the facet, other than the calibrator source, contained enough signal to still disrupt the calibration of short baselines. Therefore, in particular for automated approaches, it may be advisable to include a phase-up of the Dutch core stations to ensure robust calibration in the final refinement step.

Figure 6.12 illustrates the image quality enhancements that we achieve with the final refinement step, for calibrating one of the most challenging calibrator sources in the ELAIS-N1 field. This source introduced the most artefacts across the

full wide-field image from de Jong et al. (2024). At 1.2″ resolution we find most of the spike-like structure originating from the calibrator source to have reduced significantly from the left to right panel. At the 0.3″ resolution, all circular-type artefacts appear to be completely mitigated.

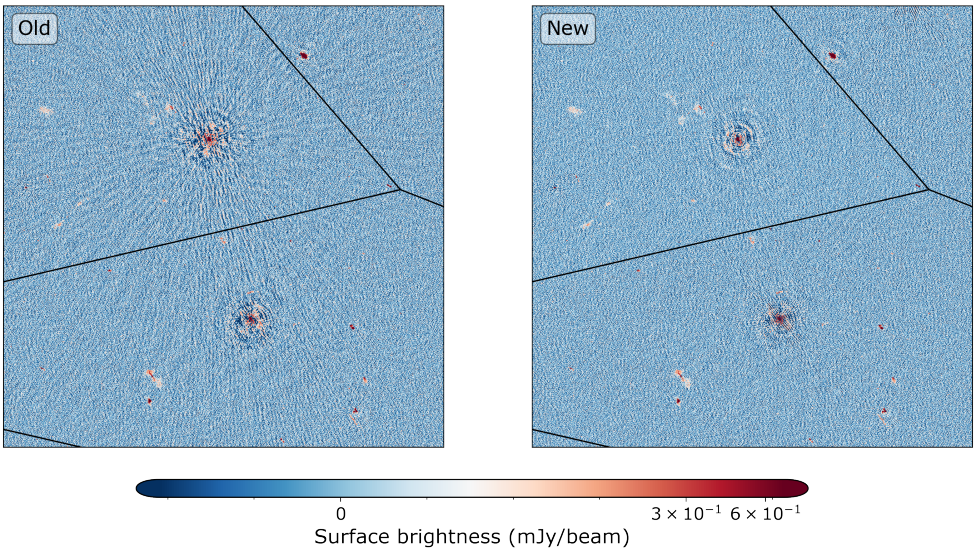

Figure 6.11: Improvements in direction-dependent (DD) calibration for 1.2″ imaging. The *left* image displays the image quality using the calibration strategy from de Jong et al. (2024), while the *right* image shows the results employing the new calibration strategy where we first apply Dutch-only DD calibration before calibrating the international stations for DDEs. This example corresponds to a challenging observation, which was previously difficult to calibrate. The black lines represent the facet boundaries.

6.6.2. Computing costs

Reducing LOFAR data for sub-arcsecond wide-field imaging is computationally intensive, due to its substantial storage demands, leading to significant CPU core hr requirements. In this subsection, we discuss the computational improvements we have made with the new data processing strategy.

6.6.2.1. Data volume

The large data volumes required for processing LOFAR data pose a significant bottleneck, especially for sub-arcsecond wide-field imaging, as the visibility data cannot be further averaged due to time and bandwidth smearing constraints. The initial unaveraged and uncompressed input data is approximately 16 TB. However, Dysco

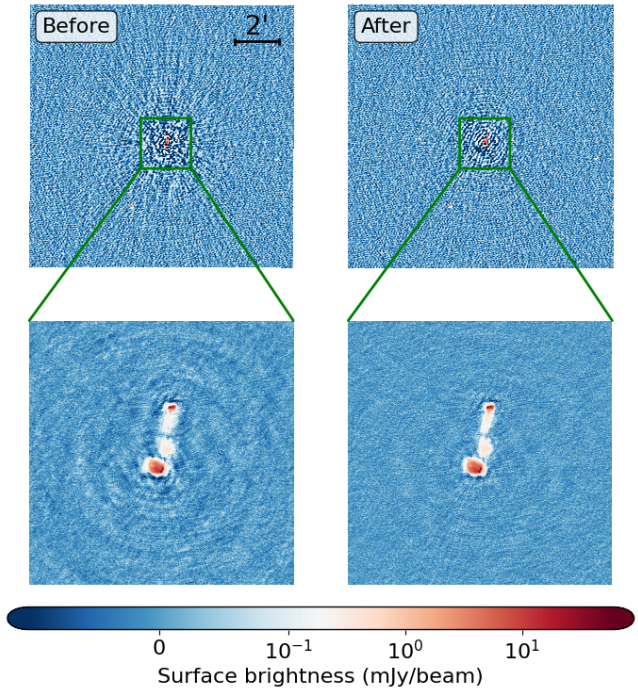

Figure 6.12: Final image quality improvements at 1.2″ resolution (above) and 0.3″ resolution (below) for one of the most challenging calibrators in the field. The *left* panels display the image quality before refinement (up to the calibration from Section 6.4.2), while the *right* panel presents the image after refinement (including calibration discussed in Section 6.4.4). We included a 2′ scale bar to illustrate the size of the image.

compression (Offringa, 2016) offers a solution to bring this number down with lossy compression. In the LOFAR Long-Term Archive (LTA), data are compressed by default at a factor of 4, with visibility data stored at 10 bits and their associated weights at 12 bits. These settings do generally not lead to a significant loss of information in LOFAR data. As noted by Offringa (2016), at high time and frequency resolutions, the data are expected to be noise-dominated. Therefore, compressing with higher bit rates may be possible for datasets for sub-arcsecond wide-field imaging.

To determine the extent to which we can further compress our data, we selected two facets from the ELAIS-N1 data after the 1.2″ subtraction, where one of the facets corresponds to a high S/N calibrator and the other to a low S/N calibrator. We applied various levels of Dysco compression by adjusting the bit rate used to store the visibilities and their corresponding weights and compared the results after imaging with the standard settings described in Section 6.4.5. Visibility weights

are already compressed by a factor of 10 compared to uncompressed data (Offringa, 2016), by storing only one polarisation and using a 12-bit compression, leaving limited potential for further compression. Changing the bit-rate reduces the total data volume only in the order of a few per cent. In contrast, much larger data volume reductions can be achieved beyond the default Dysco settings by adjusting the number of bits used for storing visibilities. This is because visibilities are stored as complex values and are more noise-dominated compared to weights. For varying compression levels expressed in the number of bits, we show in Figure 6.13, the RMS increase in the images as a function of compression level, given the data volume compressed size compared to the default 10-bit visibility and 12-bit weight compression with Dysco. This figure indicates that with 6-bit compression, the RMS background noise remains unchanged, and only a an increase in the order of a few per cent is observed for 4-bit compression, with more pronounced noise increases at higher compression levels, reaching about 100% at 2-bit compression. We also find that the image quality at 1.2″ imaging is more affected at higher bit rates to RMS increases compared to the 0.3″ resolution imaging. This is because the data is more averaged over time and frequency, making the data less noise-dominated. We verified that the residual images at these bit rates – obtained after subtracting the image corresponding to the original 10-bit stored data – remain purely noise-dominated and show no unusual artefacts. We also performed 1.2″ wide-field imaging with datasets stored at different bit rates and found similar results to the 0.3″ low S/N case, which is probably because these datasets are less averaged compared to the facet-subtracted data. Additionally, we confirmed that for bit rates below 6, the peak fluxes remain unaffected. Our results demonstrate that for the ELAIS-N1 dataset, and likely for many other LOFAR pointings, the visibility data can be compressed to 6-bits, resulting in a 40% data volume reduction. In cases where data volume presents a significant bottleneck, such as large-scale data processing with limited storage, 4-bit compression may also be viable.

The above applies to the data compression of one observation. For deep imaging using multiple observations, data compression is also achieved by reducing the number of visibilities through SVA, as discussed in Section 6.4.5 and by de Jong et al. (2025). For example, when combining 500 hrs of data without significant precession or aberration effects due to large observation time offsets, the number of visibilities can be reduced by a factor of 29. However, applying Dysco compression on top of SVA must be done with caution, as the combined data has lower noise levels depending on the number of observations merged.

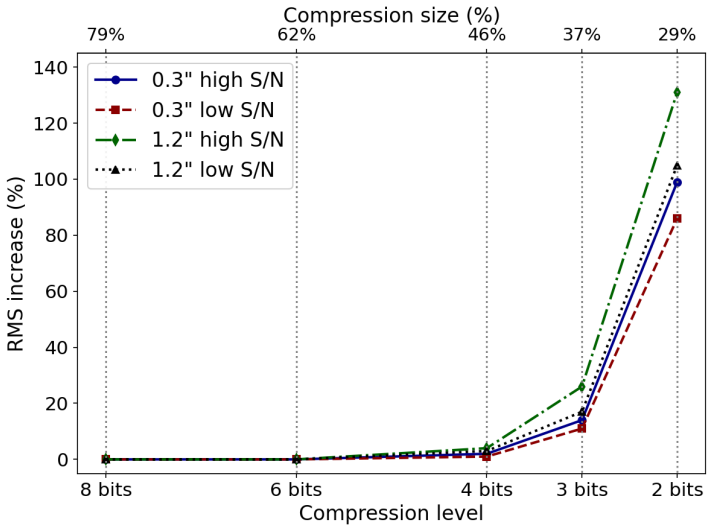

Figure 6.13: Comparison of RMS noise increase in image space as a function of the number of bits the visibilities are stored and compression size for two different facets, imaged at two different resolutions. The 1.2″ images are 4 times more in time and frequency averaged compared to the 0.3″ images. The compressed size is compared with Dysco's default settings (10-bit visibility storage and 12-bit weight storage).

6.6.2.2. CPU core hours

The computing costs reported in de Jong et al. (2024) summed to 680,000 CPU core hrs for calibrating and imaging 32 hrs of data, which corresponded to 170,000 CPU core hrs for 8 hrs of data. Since half of the data was before data reduction already averaged to 2 sec, this corresponds to about 250,000 CPU core hrs for one 8 hrs at 1 sec. With the strategy improvements presented in this work, we have achieved a significant reduction in these high costs, as outlined in this subsection. The costs for the different modules from our upgraded data processing strategy are given in Figure 6.1. The cost reductions enable us to scale data processing of LOFAR data for ultra-deep wide-field imaging and achieve sensitivities in the order of μJy beam^{-1}.

A new step, as part of the 'DD calibration' module in Figure 6.1, is the pre-self-calibration steps for the Dutch core and remote stations (see Section 6.4.1). This increases the total computational costs for every observation by about 8,000 CPU core hrs. However, for ultra-deep imaging, this cost can be partly mitigated by applying the Dutch calibration to for instance 6 observations to generate a sufficiently deep model, which can then be used as input for calibrating the remaining observations. This approach reduces the number of self-calibration cycles required for subsequent observations, lowering the overall cost by a factor of ~3 compared to

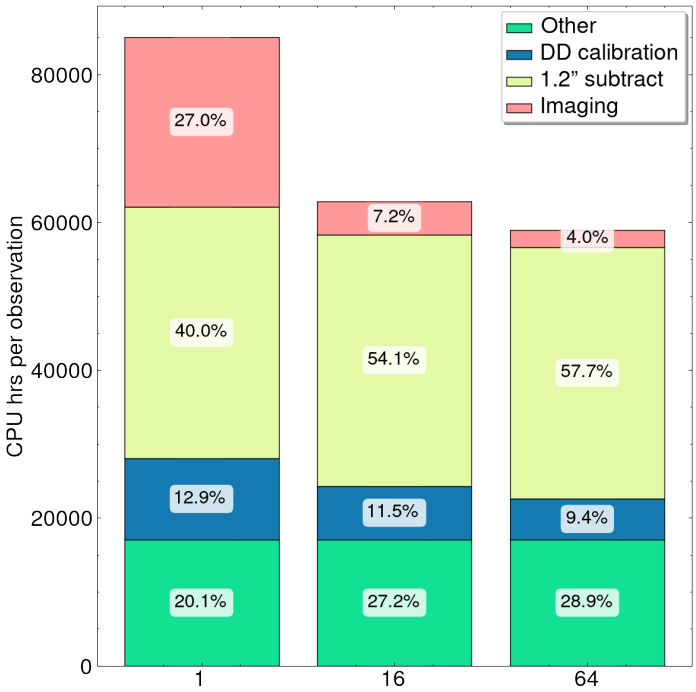

Figure 6.14: Breakdown of the computational cost across the major data processing steps when combining 1, 16, or 64 LOFAR observations of 8 hrs. We have separated the full cost for the DD calibration (outlined in Sections 6.4.1, 6.4.2, and 6.4.4) from the 1.2″ subtraction (outlined in 6.4.3), and imaging (outlined in Section 6.4.5). Percentages represent the fraction of the computational cost for full processing. For comparison, we normalised the total cost to process one 8 hrs observation. The lower computational cost when processing more observations is because using sidereal visibility averaging and model images during calibration reduces the required CPU core hrs. The computational cost for 'other' encompasses the sum of all additional steps to reduce the data in this work, which are associated with the modules not covered within 'This work' in Figure 6.1.

self-calibration without an initial model.

Employing early-stopping with the neural network from Section 6.3.3, allows us to self-calibrate the main facet calibrators to already reach convergence after 5 or 6 self-calibration cycles. Only in a few cases, this exceeds up to cycle 10 (in Section 6.4.2). This nearly halves the original computational cost for this calibration step compared to de Jong et al. (2024), which was about 1,500 CPU core hrs. However, the computational cost increases again with the introduction of an additional refinement step from Section 6.4.4, where we perform another round of self-calibration for all facets (see Section 6.4.4). Nonetheless, due to recent improvements in speed and data volume handling within `facetselfcal`, the total computational costs for

DD calibration of the international stations remain still comparable to that of de Jong et al. (2024). Thus, this step does not present a computational bottleneck.

The most computationally expensive module is the 1.2″ facet subtraction. While de Jong et al. (2024) included the facet subtraction with 1.2″ models as part of imaging, we have in this work separated this step from the imaging (see Section 6.4.3), as it is in the new strategy an essential part of performing the final high-quality self-calibration outlined in Section 6.4.4. We have significantly reduced the cost for this step by modifying the source subtraction strategy (see Section 6.4.3), resulting in a computational cost reduction of a factor of 5, bringing it down to approximately 32,000 CPU core hrs per observation. This is significant given that this step originally accounted for over 158,000 CPU core hrs for a single 8 hrs observation, which was 62% of the total data reduction. As a result, this improvement alone accelerates the entire overarching data processing workflow by about a factor of 2.

The final imaging module, utilizing SVA for deep imaging (see Section 6.4.5), introduces another significant speed improvement, especially when processing a large number of observations. This is because each observation has its own sidereal day. The CPU core hr reduction scales linearly with the data volume reductions mentioned in Section 6.6.2.1, with a slight compute overhead for applying the SVA algorithm itself. For 128 hrs of data, the reduction in CPU core hrs is approximately a factor of ~5 (equivalent to 5,700 CPU core hrs per observation), while for 500 hrs, the speedup increases to roughly ~10 (equivalent to 2,850 CPU core hrs per observation).

Beyond our calibration and imaging strategy changes, we are also benefiting from more efficient usage of computing resources, by combining CWL, Toil, and Slurm. This allows us to more efficiently distribute our jobs and optimise the number of CPU cores for each substep. In Figure 6.14, we compare the current total computational costs for reducing LOFAR data for wide-field 0.3″ imaging for combining 1, 16, and 64 observations. We normalised the cost to processing one 8 hrs observation. To simplify the plot and to focus on the steps optimised in this work, we have combined `LINC`, the `DDF`-pipeline, 6″ subtraction, and DI calibration in 'other'. We find a total computational cost for 1 observation of ~85,000 CPU core hrs, and estimate costs for 16 observations of ~1,000,000 CPU core hrs, and for 64 observations of ~4,000,000 CPU core hrs. This demonstrates that, depending on the number of combined observations, the total data processing cost has been reduced by a factor of 6 to 8 compared to Sweijen et al. (2022c, taking into account they averaged the data by a factor 2), and by a factor of 3 to 4 compared to de Jong et al. (2024, taking into account half of their data was averaged by a factor 2). This factor depends on how many observations are combined together with SVA.

6.7. Discussion

By enhancing our data reduction strategy, we have enabled the production of sub-arcsecond resolution images with sensitivities in the order of μJy beam^{-1} at 3 to 4 lower computational costs and with improved image qualities compared to what was previously possible, as we show in Section 6.6. In this section, we discuss possible remaining improvements and implications of our results for a potential future automated sub-arcsecond wide-field survey with LOFAR.

6.7.1. Modular data processing

The data processing framework presented in Section 6.2 and Figure 6.1 comprises different modules, each responsible for a specific data processing task. While all modules are essential for wide-field imaging, they function as individual blocks (as already highlighted by de Jong et al., 2024), which can be modified independently without directly affecting the input of the next module. This allows for flexible updates to the data processing strategy by improving an isolated single module or having the option of adding new modules.

Reordering of some of these modules may also be considered for performing DD calibration for the Dutch stations. This type of calibration happens currently both during the `DDF`-pipeline and after the DI in-field calibration of the international stations with `facetselfcal` (see Section 6.4.1). However, it may be more efficient to allow the option to pre-apply the DD solutions from the `DDF`-pipeline before performing the DD calibration of the international stations. This implicates that we could place the `DDF`-pipeline module in Figure 6.1 after the DI in-field calibration of the international stations, making the additional DD calibration step for the Dutch stations with `facetselfcal` (from Section 6.4.1) unnecessary. This may lead to computing cost reductions by approximately 10% over the entire data processing.

Given the calibration strategy of our in-field calibrator (as discussed in Section 3.2.3. from de Jong et al., 2024) with full-Jones corrections, we can also perform polarisation studies. However, we need for this to align the different observations by applying an additional correction to the polarisation angles across observations. An automatic approach for this correction may be implemented as a separate module prior to the final imaging step. Similarly, additional calibration refinement steps or automated inspections can be added between already existing modules as well.

A benefit of the new data processing strategy for deep imaging is that we fully process each observation embarrassingly parallel up to the 'DD calibration refinement' module in Figure 6.1. This is possible because only at the DD calibration refinement stage do the different observations need to be calibrated together, as discussed in Section 6.4.4. Parallel processing is in particular beneficial when pro-

cessing a large number of observations of the same field of view for ultra-deep imaging on different nodes or even different HPCs. It also enables full processing up to the calibrated facet datasets and removing all intermediate products before processing the next dataset. This may be essential when storage space is limited.

6.7.2. Advanced decision making

In Section 6.3.3, we introduced the use of neural networks to employ early-stopping during self-calibration and optimise our automatic DD calibration parameter settings in Section 6.4.2. This approach shows that incorporating steps for automatic decision-making in our data reduction strategy enhances performance by removing human intervention to select the optimal self-calibration cycle, while also reducing computational costs. In this subsection, we discuss a few other use-cases to incorporate automated decision-making during data reduction.

6.7.2.1. Calibratability

Not every dataset is calibratable, by which we mean that some datasets are dominated by RFI or bad ionospheric conditions and therefore too challenging to calibrate. It may therefore in some cases not be worth processing a dataset, given the high computing costs involved. To avoid wasting the use of computing resources on calibrating bad data, it is worth to assess the calibratability of data at an early stage. This could for instance be done before or after running `LINC`, which is the first step in processing LOFAR data (see Figure 6.1). This may be best done automatically by including various metrics on the calibration solutions or UV data directly, or perhaps by using external data with information on the ionospheric conditions during observations (e.g. Flisek et al., 2023). A final data quality score would not only aid in assessing our specific use case but also benefit other applications of LOFAR data across various scientific objectives, as `LINC` serves as a general data reduction step for imaging LOFAR datasets.

Similarly, the `DDF`-pipeline returns data products, including wide-field images, that indicate the calibratability of the data as well. An assessment of the remaining DDEs may indicate the severity of the ionosphere for a particular dataset. Providing postage stamps of bright sources in the field to our neural network from Section 6.3.3, can assist in deciding whether to continue or stop data processing. This would only require a re-training of the model including images at $6''$. Since the majority of CPU core hrs are spent after `LINC` and the `DDF`-pipeline (see Figure 6.1), this approach can save a considerable amount of computing time, allowing resources to be allocated to processing other observations with fewer data issues.

6.7.2.2. In-field calibration

Ye et al. (2024) and de Jong et al. (2024) devoted much attention to manually select and calibrate the in-field calibrator for ELAIS-N1 (identified as ICRF J160607.6+552135; Charlot et al., 2020; Sexton et al., 2022). Since for ELAIS-N1, we are aiming to calibrate the same field of view for many observations, we do not have to modify this calibration step. However, for a potential future wide-field imaging survey (e.g. ILoTSS) this process needs to be automated for different sky areas.

The first source of information for selecting potential in-field calibrators comes from the Long-Baseline Calibrator Survey (LBCS; Jackson et al., 2016, 2022). However, if no candidate sources from the LBCS catalogue are available, one could consider using the brightest sources above a specific mJy threshold from the LoTSS catalogues (Shimwell et al., 2017, 2019, 2022), which covers most of the northern sky. In the rare instance where a LoTSS catalogue is unavailable, one could also consider using source finder software such as `PyBDSF`[15] (Mohan & Rafferty, 2015), on the $6''$ map that the `DDF`-pipeline produced in the step before the international DI calibration (see Figure 6.1).

Having at least one calibrator source after the initial catalogue-based selection, allows us to assess the S/N for the longest baselines by employing the same calibrator selection method based on the phasediff-scores as described in Section 6.3.1, but with a different score threshold. To find the best calibrator and (self-)calibration parameters, a good starting point is using similar settings as those optimised by de Jong et al. (2024). Next to this, it may be worth applying a grid search with different (starting) parameters on the candidate(s). In this way, the neural network model from Section 6.3.3 and additional image and solution quality measures can be utilised to select the best parameters and/or candidate source. Given that the CPU cost is low for the DI calibration (see Figure 6.1), this step does not substantially increase the overall computing time. Over time, when many in-field calibrators have been optimised and the parameters from the different grid-searches have been collected, a more advanced new neural network can be trained to immediately set the best initial parameters corresponding to the potential in-field calibrator source that has been automatically selected.

6.7.2.3. Self-calibration parameter-tuning

In the DD self-calibration images from Figure 6.6, we observe that while the self-calibration results in the last two rows show minor improvements after the initial cycles, additional cycles do not lead to full convergence. In the fourth row, despite the neural network providing a positive assessment, there remains room for further

[15] `https://pybdsf.readthedocs.io`

refinement. In such cases, it may be beneficial to not only assess convergence and image quality but also adjust the calibration parameters to enhance results. A possible approach could involve beginning with a conservative set of parameters in the initial self-calibration round, then gradually shifting to less restrictive settings, such as shorter solution intervals and reduced frequency smoothness constraints, continuing until convergence is achieved according to our metrics. If self-calibration starts to diverge, parameters can be adjusted back to more conservative values.

6.7.3. Data compression and I/O

With our data reduction strategy enhancements and data volume reductions, we have significantly reduced the computing costs compared to previous works. Nonetheless, more gains in terms of computing cost reductions are still possible.

As outlined in Section 6.6.2.1, we use Dysco compression to reduce the size of our visibility data. For the ELAIS-N1 dataset, we found that stricter compression using fewer bits to store visibilities data, reduced data volumes without loss of image quality. This works well for ELAIS-N1 due to the absence of extremely bright sources in the field. However, for other observations containing brighter sources or datasets with more averaging applied, the optimal bit rates may differ, as the data is less noise-dominated. Additionally, we do not recommend using high compression levels for phased-up data, as the S/N is much higher on baselines involving the Dutch superstation. Since bit rates depend on the S/N across baselines in each dataset, optimizing compression for automated data processing would require a method to dynamically determine the maximum allowable compression level for each specific dataset. With higher bit rates, we can improve the I/O overhead in our data reduction, such as during the $1.2''$ subtraction, where the unaveraged subtracted data has to be copied multiple times to perform the prediction for each facet before averaging down in time and frequency.

A limitation of Dysco is that it cannot compress model data or be applied to BDA data. Recent work with MultiGrid Adaptive Reduction of Data (MGARD; Dodson et al., 2024) compression on high-resolution ELAIS-N1 data from de Jong et al. (2024) discuss the potential for combining lossy compression with BDA, while performing similar levels of compression as Dysco on the visibility data and weights. MGARD may also have the potential to further compress model data, which could help reduce data volume requirements during self-calibration, particularly in the multi-directional self-calibration process discussed in Section 6.4.1.

6.8. Summary and conclusion

In this work, we tackled several key challenges that limit the efficient processing of numerous observations for ultra-deep sub-arcsecond wide-field imaging with LOFAR. Our main improvements include:

- An efficient modular framework for data reduction on HPC systems with a scheduler supporting Toil and CWL, allowing for embarrassingly parallel processing of many observations of the same field for ultra-deep imaging and having the potential to perform ILoTSS surveys as well.
- An improved automated DD calibration strategy, with a focus on improving the calibration solutions for the Dutch stations and removing the need for human interventions, by combining already existing image and solution quality metrics with a neural network that assesses image quality.
- optimised strategy for reducing the computing costs of wide-field imaging by a factor of 3 to 4 compared to the most recent similar work. The exact factor depends on the number of observations that are being combined since more observations of the same sky area allow with our new strategy for more speedup.

Our new data reduction strategy has been validated using two observations from the ELAIS-N1 deep field, where one was used to compare results with de Jong et al. (2024), and the other for testing on a more challenging observation. We find clear improvements for in particular the calibration of the shorter baselines, compared to previous work, which enhances the quality for both the $1.2''$ and $0.3''$ resolutions.

We have also identified and discussed several steps to further improve the automated data reduction strategy:

- In the new data calibration strategy we perform DD calibration on the Dutch stations twice. This is inefficient from a workflow perspective. A possible solution is to perform DI in-field calibration before running the `DDF`-pipeline or `facetselfcal` for joint-calibration of the Dutch stations. With the modular design of our framework, this can be done by reordering the main processing steps, which may eventually save an additional 10% on total computing time.
- We have demonstrated the value of automated decision-making through a neural network for image quality validation. Additional steps earlier in the data reduction process could also be improved with more data quality assessments. This would help reduce the waste of computing resources on observations impacted by poor ionospheric conditions.

- ☐ To extend this work to other sky areas through automatic processing, an automated DI in-field calibration step is currently missing. We propose using similar techniques as we have presented in this work for DD calibration. However, this needs careful testing on more observations from sky areas with various conditions.

- ☐ Further improvements in data processing speed can be achieved by focusing on reducing data volumes, by incorporating data compression techniques to compress model data or by combining lossy data compression with BDA.

The next major goal is to apply our calibration and imaging strategy to reduce hundreds of hrs of ELAIS-N1 data available in the LTA and other deep surveys, aiming to reach sensitivities on the order of 5μJy beam^{-1}. This work also marks significant progress toward establishing a fully automated survey workflow for (deep) sub-arcsecond imaging with LOFAR.

Acknowledgements

This work is part of the project CORTEX (NWA.1160.18.316) of the research programme NWA-ORC which is (partly) financed by the Dutch Research Council (NWO). This work made use of the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-1287. This work is co-funded by the EGI-ACE project (Horizon 2020) under Grant number 101017567. RJvW acknowledges support from the ERC Starting Grant ClusterWeb 804208. LOFAR data products were provided by the LOFAR Surveys Key Science project (LSKSP; `https://lofar-surveys.org/`*) and were derived from observations with the International LOFAR Telescope (ILT). LOFAR (van Haarlem et al., 2013) is the Low Frequency Array designed and constructed by ASTRON. It has observing, data processing, and data storage facilities in several countries, which are owned by various parties (each with their own funding sources), and which are collectively operated by the ILT foundation under a joint scientific policy. The efforts of the LSKSP have benefited from funding from the European Research Council, NOVA, NWO, CNRS-INSU, the SURF Co-operative, the UK Science and Technology Funding Council and the Jülich Supercomputing Centre. The use of the national computer facilities in this research was subsidised by NWO Domain Science.*