Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Deep microlocal reconstruction for limited-angle tomography

Héctor Andrade-Loarca [a], Gitta Kutyniok [a,b], Ozan Öktem [c,d,*], Philipp Petersen [e]

[a] *Department of Mathematics, LMU Munich, 80333 Munich, Germany*
[b] *Department of Physics and Technology, University of Tromsø, 9019 Tromsø, Norway*
[c] *Department of Mathematics, KTH - Royal Institute of Technology, SE-100 44 Stockholm, Sweden*
[d] *Department of Information Technology, Division of Scientific Computing, Uppsala University, SE-751 05 Uppsala, Sweden*
[e] *Faculty of Mathematics and Research Network Data Science, University of Vienna, 1090 Vienna, Austria*

A B S T R A C T

We present a deep-learning-based algorithm to jointly solve a reconstruction problem and a wavefront set extraction problem in tomographic imaging. The algorithm is based on a recently developed digital wavefront set extractor as well as the well-known microlocal canonical relation for the Radon transform. We use the wavefront set information about x-ray data to improve the reconstruction by requiring that the underlying neural networks simultaneously extract the correct ground truth wavefront set and ground truth image. As a necessary theoretical step, we identify the digital microlocal canonical relations for deep convolutional residual neural networks. We find strong numerical evidence for the effectiveness of this approach.

1. Introduction

Tomographic imaging aims at uncovering the interior 2D/3D structure of an object from a *sinogram*, which is data obtained by repeatedly exposing the object to a particle or wave from different directions. This is a key example of an *inverse problem* where one computationally attempts recover an unknown signal from data given as indirect observations. A reconstruction method refers to an algorithm performing the recovery.

* Corresponding author.
  E-mail addresses: andrade@math.lmu.de (H. Andrade-Loarca), kutyniok@math.lmu.de (G. Kutyniok), ozan@kth.se, ozan.oktem@it.uu.se (O. Öktem), philipp.petersen@univie.ac.at (P. Petersen).

Inverse problems, like those that arise in tomographic imaging, are often *ill-posed* which means that there can be multiple solutions consistent with the data or solution procedures that maximise consistency against measured data are sensitive to variations in data. Such high sensitivity is referred to as instability, and it appears when the *forward operator*, which models how a signal gives rise to corresponding noise-free data, is not continuously invertible. Stability properties are further degraded for sparse-view data, which is when data is under-sampled, and for limited-angle data, which refers to unevenly sampled data. *Regularisation* refers to mathematical theory and methods for stabilising the solution procedure of an ill-posed inverse problem. Many regularisers enforce stability by requiring consistency against a *prior model*. This prior should ideally encode known properties of the unknown signal one seeks to recover, and choosing an appropriate prior is an essential part of regularisation.

Most reconstruction methods in tomography assume that measurements are collected from views that are evenly distributed around the object. *Limited-angle tomography* refers to a case when this is not fulfilled. Such problems arise naturally in many applications, like digital breast tomosynthesis [60,6], dental tomography [43,44], electron tomography [68,63], transmission x-ray microscopy [42], nondestructive testing [66,71], geophysical prospecting [84], etc. This missing data significantly amplifies the instability in the corresponding reconstruction problem [21,58]. Hence, traditional reconstruction methods, like filtered back-projection (FBP) [10,78,36], that implicitly assume missing data is zero do not perform well in such situations. Overall, it has been very challenging to develop regularisation methods that handle this instability, as these methods somehow need to fill in the missing data without imposing too strong assumptions on the signal; see the brief survey in Subsection 1.3. As a consequence, it has been challenging to develop practically useful reconstruction methods for limited-angle tomography that provide sufficient improvement over traditional reconstruction methods. In this context, we consider a reconstruction method practically useful if it is computationally feasible and does not require a user to set multiple hyperparameters.

1.1. Main contributions

In this paper, we develop theory and algorithms for reconstruction in severely ill-posed inverse problems that arise in tomographic imaging with limited data. In particular, we develop a data-driven reconstruction method for limited-angle tomography that is microlocally consistent, which means 'filling in' missing data in a way that is consistent with how singularities in data are related to those in the signal.

Our approach relies on deep neural networks (DNNs) that integrate a handcrafted forward operator and a theoretical characterisation of how singularities in data are related to those in the signal. In this context, we model singularities through the concept of the *wavefront set*, that will be introduced in detail in Section 3 below. For the Radon transform, which is the forward operator underlying the tomography problem, it is known exactly how an application of it affects the wavefront set of underlying data. Indeed, the wavefront set of an image and its Radon transform are linked through so-called *canonical relations*. To use this a priori information on the wavefront set in the reconstruction problem, we set up the following method: Our algorithm consists of two workflows that act in parallel. First, a *reconstruction line* that maps the data to the (unknown) signal. This is a specific deep neural network using the *Learned Primal-Dual architecture*, [2]. Second, a *micro-local line* that identifies the wavefront set of the signal by using the following three steps: (a) A neural network that extracts the wavefront set of the data. This neural network has been established earlier in [4] under the name DeNSE and is based on the interaction of the shearlet transform with the wavefront set of a function. (b) An analytical computation relating the wavefront set in the data domain to the image domain in a way that mirrors the action of the Learned Primal-Dual architecture of the reconstruction line. The two lines are coupled via this step. (c) A neural network with the U-Net architecture [72] that inpaints the wavefront set in the image domain that was inferred from the incomplete data.

The neural networks in the microlocal line and the reconstruction line are now trained in parallel on an artificial training set consisting of 2D phantoms (signals) with corresponding noisy and incomplete sinogram data. The phantoms are made up of random shapes that are demarcated by piecewise smooth curves given by splines of degree at most four. Several examples from the data set are shown in Fig. 2.

The combined procedure, which we will be calling the *joint reconstruction algorithm*, now attempts to satisfy two objectives on that data set: (a) The reconstruction returned from the reconstruction line should agree as closely as possible with the ground truth. (b) The wavefront set returned by the microlocal line should resemble the ground truth wavefront set of the data accurately.

We refer to Section 5 for the complete description of the joint reconstruction algorithm and in particular Fig. 4 for an illustration of the underlying DNN architecture.

Note that, due to the simplicity of the training data set, we can analytically compute the true wavefront set for those signals, which in turn is needed for the aforementioned joint training of the DNNs for wavefront set inpainting and reconstruction. In contrast, our test data consists of images of brains and associated noisy incomplete sinograms. Therefore, the *training data is substantially different from the test data.* As such, the empirical numerical study also shows the generalisation properties of our approach.

We wish to emphasise that this approach applies in principle to any inverse problem where the forward operator is a Fourier integral operator, as is the case for most inverse problems arising in imaging applications. However, for simplicity, we chose to work out the theoretical results with associated algorithms and numerical examples only for the specific case of planar limited-angle tomography.

A central part of the proposed algorithm consists of the theoretical analysis of how the Learned Primal-Dual architecture affects the extracted wavefront set. At this point, the reader may justifiably wonder if it would not instead be possible to estimate the wavefront set of the output of this architecture directly via the wavefront set extractor DeNSE. This is theoretically possible but turned out to be practically infeasible. While the wavefront sets of the data set can be precomputed, this cannot be done for the outputs of the Learned Primal-Dual network as the Learned Primal-Dual network varies during the training. This means that during training a full application of DeNSE would have to be performed in every training step. On a modern machine, an application of DeNSE to a single image in the data set takes approximately 20 seconds, which shows that this approach would slow down the training process dramatically compared to our method where the propagation of the wavefront set through the reconstruction architecture is established analytically.

As a small appetiser, we conclude this overview of the main contributions by illustrating the superior performance of our approach. Fig. 1 compares reconstruction performed by a DNN, here the learned primal-dual network, against our joint approach that combines this Learned Primal-Dual network with a DNN for inpainting the wavefront set. In this figure, the tomographic data is given in the form of highly noisy samples of the Radon transform with 40° missing angular wedge. Reconstructions shown in the top row are from two supervised deep learning approaches: namely, Learned Primal-Dual (middle) and our joint approach (right). The latter is essentially the Learned Primal-Dual network combined with a learned wavefront set inpainting that complements missing data in a microlocally consistent manner. One can see that the qualities of both image reconstruction and wavefront set inpainting increase when they are trained jointly.

Both the Learned Primal-Dual network and the joint DNN for reconstruction and wavefront set inpainting were trained against the same training data with the same total number of training steps. The joint approach clearly shows the benefit of complementing missing data as it is able to recover essential features that are lost using the Learned Primal-Dual network. Moreover, the joint approach also seems to recover reasonably well singularities (edges) in parts of the image that are not in the visible wavefront set. This is remarkable as these singularities leave, according to microlocal theory, no measurable footprint in data. Furthermore, applying wavefront set inpainting to data separately as a pre-processing step prior to reconstruction yields a significantly worse estimator for the wavefront set.
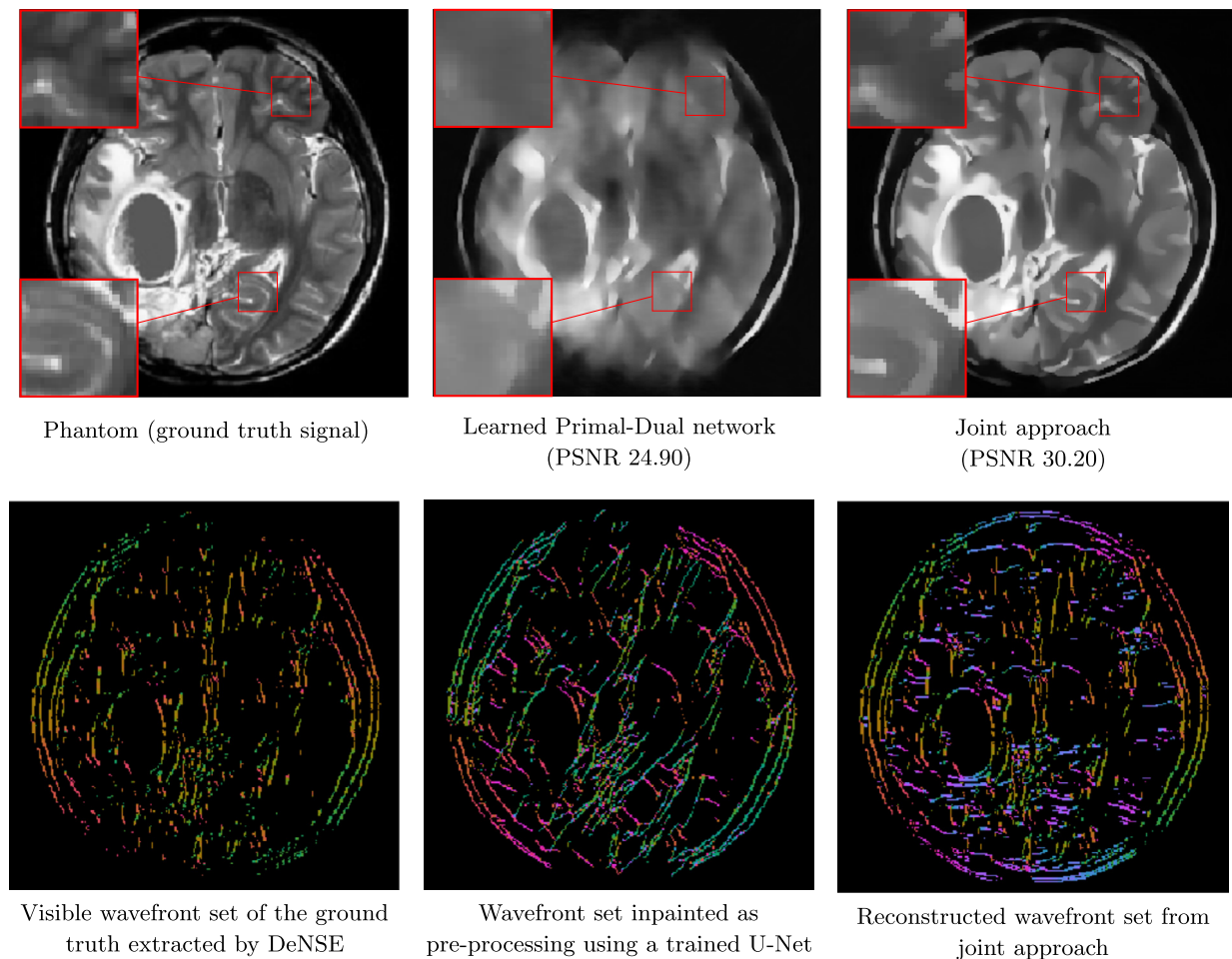
Phantom (ground truth signal)

Learned Primal-Dual network
(PSNR 24.90)

Joint approach
(PSNR 30.20)

Visible wavefront set of the ground
truth extracted by DeNSE

Wavefront set inpainted as
pre-processing using a trained U-Net

Reconstructed wavefront set from
joint approach

Fig. 1. Limited-angle parallel beam tomography with a realistic brain phantom.

This shows the benefit of performing these steps simultaneously as in the joint approach. Figs. 8 and 7 present a more extensive comparison that also includes sparse-view tomography and other model based and data-driven methods.

1.2. Outline of the paper

Sections 1, 3 and 2 – along with Appendix A – primarily serve as background material, whereas the main scientific contribution is contained in Sections 4 and 5. Finally, Section 6 provides numerical examples that showcase the performance of the suggested approach and compares it to other methods for image reconstruction in limited-angle tomography. We present a more detailed outline below:

Section 1 provides background motivation from applications along with an overview of the main scientific contributions (Subsection 1.1) and a survey of current state-of-the-art for limited-angle tomography (Subsection 1.3). This is followed by Section 2, that mathematically formalises an inverse problem, both as an operator equation as in (2.1) as well as a statistical inference problem as in (2.2). This section also introduces the notions of ill-posedness and regularisation (Subsection 2.1). Emphasis is next on the Learned Primal-Dual network introduced in Subsection 2.2. This is a DNN with an architecture that incorporates the forward operator which is to be inverted.

Background on the mathematics of tomographic imaging is provided in Section 3. There, we define the Radon transform arising in planar tomography (Definition 3.1) and recall some of its properties
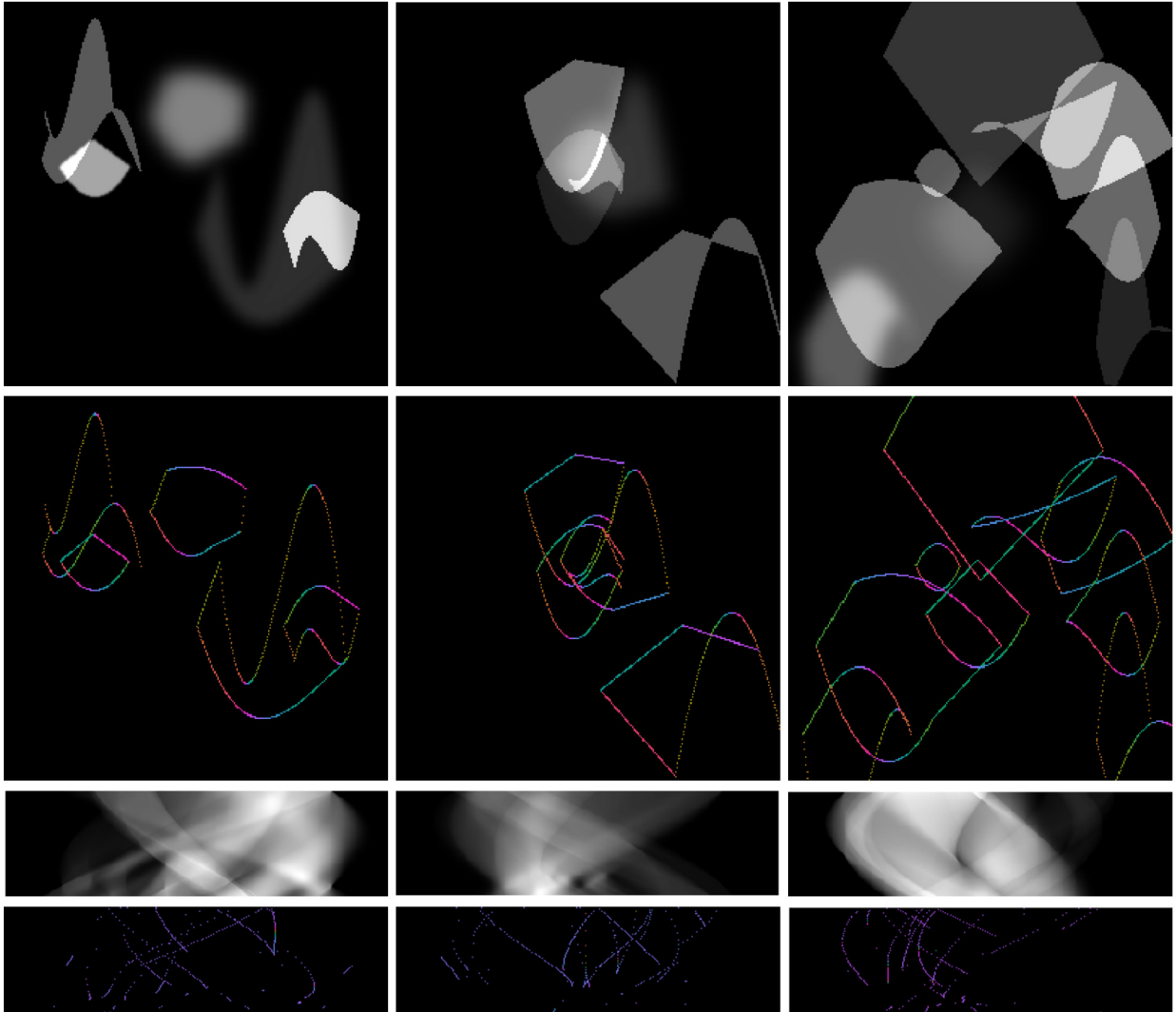
Fig. 2. Examples of the piecewise smooth data set comprised of functions which are piecewise polynomial and where the interfaces are given by splines of degree at most four. In the second row, we depict the associated analytically found wavefront sets. The third and fourth rows show, respectively, the corresponding sinograms and their wavefront set, both in the limited-angle setting with 40° wedge.

(Subsection 3.1). Focus here is on extending the Radon transform and its corresponding back-projection (Definition 3.5) to tempered distributions. This uses basic notions from distribution theory provided in Appendix A (some of the material in this appendix is also used later in Section 4). Subsection 3.1 also introduces the restricted Radon transform arising in limited-angle tomography along with its corresponding restricted back-projection defined in (3.5). Finally, Subsection 3.3 formally defines the wavefront set (Definition 3.6) and then states the canonical relation for the Radon transform given in (3.7). In (3.8) we also provide a characterisation from microlocal analysis of the visible (and invisible) parts of the wavefront set for the Radon transform.

The main scientific contributions are contained in Sections 4 and 5. More precisely, Section 4 provides a theoretical analysis of the propagation of the wavefront set through the distinct layers in a continuum version of the Learned Primal-Dual architecture with the Radon transform as forward operator. Section 5 then presents, in the digital setting, how the digital wavefront set is propagated through the layers of the Learned Primal-Dual architecture. This propagated wavefront set, along with the characterisation of the

visible wavefront set, is used for setting up the DNN for wavefront set inpainting, which recovers the invisible part of the wavefront set of a signal from the visible part. This DNN for wavefront set inpainting is then combined with the Learned Primal-Dual network for reconstruction, and both DNNs are trained jointly following the task-adapted reconstruction paradigm outlined in [1], see Section 5.6 for further details.

Finally, Section 6 provides numerical evidence and benchmarks related to image reconstruction in limited-angle tomography.

1.3. Related work

The severe ill-posedness associated with limited data in inverse problems has attracted much interest within the research community. This survey will mainly focus on the development of theory and methods for inversion of the Radon transform from limited-angle data, which is an archetype of a severely ill-posed inverse problem.

1.3.1. Analytic approaches

Lambda tomography is one of the first examples of specifically designing a reconstruction method for limited data. It is an analytic approach that was initially developed for 'inverting' the Radon transform from region of interest data [85,28,27,46]. The idea is to replace the standard filter in filtered back-projection [10,78,36] (that has infinite support) with a filter that takes a second derivative in the detector variable. This numerical derivative filter has small support, just near the line being evaluated, hence the reconstruction method is local and thus it applies to region of interest data. Lambda tomography was later applied to limited-angle data [47], where it recovers the visible wavefront set [65,68,31,67]. The recovery is only mildly ill-posed [68,50], which is in stark contrast to the severe ill-posedness that is associated with attempts at reconstructing a function from limited-angle Radon data.

Lambda tomography is computationally efficient and offers an improvement over the traditional filtered back-projection method. Its robustness and usefulness was successfully demonstrated when Lambda tomography was applied for solving the limited-angle problem with highly noisy data in electron tomography [69]. The analytic nature of the method also means the method is feasible for time-critical and/or large-scale problems. The drawbacks of Lambda tomography are similar to those of filtered back-projection, namely that its filter needs to be specifically designed for the acquisition geometry and that the prior implicitly contained for regularising the problem has limited power.

1.3.2. Variational models for reconstruction

Much effort has been spent on adapting variational models of the form in (2.3) to limited-angle tomography. Methods cited here aim to design regularisers that are specifically adapted for limited-angle tomography, e.g., by accounting for the anisotropic resolution due to limited-angle data.

Many variational models for limited-angle tomography build on modifying the total variation (TV) regulariser, like various anisotropic versions that were introduced in [25,8], see also [15, Section 2.4]. These can account for the scanning configuration bearing in mind the missing angular region in limited-angle data, which was also done in [19,95]. This idea was further elaborated in [91], where an iteratively re-weighted anisotropic TV regularisation method was introduced to approximate the sparsity of ℓ_0 quasi-norm. A further development came in [93] which set-up an alternating (directional) edge-preserving diffusion based on the one-dimensional ℓ_0 quasi-norm and a (directional) smoothing model based on a one-dimensional Dirichlet energy. The aforementioned directions are dictated by the missing angular region, and the regularisation strategy takes into consideration the fact that the ℓ_0 norm is better at preserving edges, while the ℓ_1 norm is better at suppressing noise. A further refinement to better balance the need for edge-preservation against noise-suppression was presented in [24]. Based on the theory of visible and invisible wavefront sets developed in [65], the authors of [24] propose a reconstruction method that encodes the visible singularities as priors

to recover the invisible ones. The model utilises generalised p-shrinkage operators introduced in [17] as regularisers to perform edge-preserving smoothing while using visible edges as anchors to recover piecewise constant or piecewise smooth reconstructions, while noises and artefacts are suppressed or removed. Another similar approach minimises an L_1/L_2 term on the gradient with additional box constraint that is reasonable for imaging applications [89].

Other attempts at sparsity-promoting variational models for limited-angle tomography rely on dictionary learning [80] or use a regulariser that promotes sparse solutions with respect to wavelets [90,94] or curvelets/shearlets [30,71]. One can further constrain a sparse solution against a given prior image as shown in [18,90,94,34].

Variational models tend to preserve edges better and reduce streak artefacts that are common in limited-angle tomography. However, the improvements are most notable for the recovery of simplistic images, like piecewise constant or piecewise smooth images. Furthermore, this improvement is only realised when (multiple) hyper-parameters are properly tuned. Finally, variational models are computationally demanding regarding run time and memory footprint, which in turn limits their usefulness in time-critical and/or large-scale imaging studies. In summary, the above cited variational models offer surprisingly little, if any, benefit in many cases where images have more complex features that are essential for their interpretation.

1.3.3. Deep-learning-based methods for reconstruction

There are several attempts at performing tomographic reconstruction by deep neural networks that are extensively surveyed in [5]. Most can be viewed as using techniques from deep learning to approximate different estimators for the statistical formulation of the reconstruction problem in (2.2).

A popular approach for using deep learning in tomographic image reconstruction is to use a trained DNN as a post-processing step for improving an initial imperfect reconstruction. Such an approach was used for limited-angle tomography in [42], which trains a U-Net against synthetic ellipsoid data and multi-category data to reduce artefacts from images obtained by filtered back-projection.

A more domain-adapted approach that outperforms post-processing is to consider DNN architectures for reconstruction that are obtained from unrolling a suitable iterative scheme. One example is the Learned Primal-Dual network in [2] outlined in Subsection 2.2 that incorporates a handcrafted forward operator and the adjoint of its derivative along with the acquisition geometry. This DNN was recently further adapted to limited-angle tomography in breast tomosynthesis [81] by incorporating additional prior information about the geometry in the form of the thickness measurement of the breast under compression.

A different deep-learning-based approach for limited-angle tomography is presented in [13]. Here, one learns the invisible part of the image using the visible part of its shearlet coefficients. This amounts to learning an anisotropic regulariser in a variational model.

1.3.4. Sinogram inpainting

An entirely different approach to limited-angle tomography is to fill in the missing angular data by some extrapolation scheme (sinogram inpainting or sinogram-recovery). This pre-processing step needs to be done in a stable manner and [22,23] uses a regularised iterative scheme for this purpose. Another approach uses projection onto convex sets to ensure the extrapolated sinogram is indeed valid by making sure it lies in the range of the Radon transform (Helgason-Ludwig consistency condition) [51].

As to be expected, there have also been attempts at using deep learning for sinogram inpainting. Here, much work has been inspired by the success that generative adversarial networks (GANs) have had in restoring missing parts of an image (image inpainting). In particular, [62] develops a deep-learning-based image inpainting where one jointly trains the DNN that inpaints the edges and the image, using the philosophy 'lines first, colour next'. The DNN for wavefront set inpainting (Subsection 5.6) that we use in this paper is strongly inspired by [62]. In our case, the missing parts are in the sinogram domain. Consequently, we need to use the reconstruction method to map the singularities to the image domain.

Most approaches for *deep-learning-based* sinogram inpainting are based on setting up and training a GAN to generate the missing sinogram data in order to suppress the streak artefacts from the truncated sinogram in limited-angle data. An example of such an approach is [56] that uses a U-Net generator and patch-design discriminator in the GAN to make the network suitable for standard medical tomography images. The GAN is trained against paired limited-angle and full sinogram data using a joint projection domain and image domain loss function where the weighted image domain loss can be added by back-projection. In this regime, we also refer to [92,64] for similar approach based on GANs.

1.3.5. Joint sinogram inpainting and reconstruction

The final series of methods jointly perform the two tasks of sinogram inpainting and reconstruction. A variational model for doing this is presented in [83] where the resulting non-convex and non-smooth minimisation is solved using an alternating (block) descent approach.

A deep-learning-based approach is developed in [96]. This approach combines a sinogram inpainting network and an image processing network. A key step is to use layers that correspond to Radon transform and its inverse inserted into existing convolutional network architectures. These allow one to go between sinogram and image domains and one can use the image processing network to reduce the artefacts caused by inconsistencies in the inpainted sinogram generated by the sinogram inpainting network. The three parts form an end-to-end network from sinogram domain to image domain, with benefits of taking both image error and sinogram error into account in the sync process in supervised training, i.e., with limited-angle/full-view sinogram pairs. To tackle this training data bottleneck, we develop an unsupervised train method with only limited-angle projection on our proposed network. Inspired by the observation that the reconstruction and projection can form a closed loop, we can derive a calculated projection from the reconstructed image, and the disparity between the calculated projection and real projection provides feedback signals to train the proposed network unsupervised.

Finally, we also mention [97] that develops an approach with a transformer-based DNN architecture instead of convolutional DNNs. Streak artefacts in limited-angle tomography are non-local. Hence removing these with convolutional DNNs is challenging. One approach to encode such long-range dependencies is to use unrolling based DNN architectures, like the Learned Primal-Dual network, that couple many convolutional neural networks with the forward operator and the derivative of its adjoint. Another approach is to consider transformer-based architectures that are better suited than convolutional DNNs due to their non-local nature. As shown in [97], such transformer-based DNNs have excellent performance for reconstruction in limited angle tomography. A downside of transformer DNNs is that they are very demanding to train and require massive amounts of training data, so this approach will scale poorly. Some of the issues related to memory footprint could perhaps be addressed by using more domain adapted transformer architectures, like Fourier Image Transformer [14]. Furthermore, there are still no guarantees that [97] extrapolates the missing data in a microlocally consistent manner.

2. Inverse problems, ill-posedness and regularisation

To mathematically formalise the notion of an abstract inverse problem, we introduce the separable Banach spaces X (*reconstruction space*) and Y (*data space*) whose elements represent possible signals and data, respectively. In many applications, the reconstruction and data spaces are also Hilbert spaces. Next, the mapping $\mathcal{A}\colon X \to Y$ (*forward operator*) represents a model for how a signal generates noise free data.

The classical functional analytic/frequentist formalisation views an inverse problem as an operator equation. More precisely, an inverse problem is the task of recovering an unknown signal $f_{\text{true}} \in X$ from data $g \in Y$ that is a single sample of the Y-valued random variable $\mathbb{g}$ defined as

$$\mathbb{g} := \mathcal{A}(f_{\text{true}}) + \mathbb{e} \quad \text{where } \mathbb{e} \text{ is a } Y\text{-valued random variable representing observation noise.} \qquad (2.1)$$

The statistical formulation further generalises the above by first assuming one can equip X and Y with Borel σ-algebrae. One furthermore assumes that the signal and corresponding data are generated by some $(X \times Y)$-valued random variable $(\mathbb{f}, \mathbb{g})$ [79]. The inverse problem is now to recover a suitable estimator of the conditional random variable $(\mathbb{f} \mid \mathbb{g} = g)$ from data $g \in Y$ that is a single sample of the Y-valued conditional random variable $(\mathbb{g} \mid \mathbb{f} = f_{\text{true}})$, where $f_{\text{true}} \in X$ is the true unknown signal and

$$\mathbb{g} = \mathcal{A}(\mathbb{f}) + \mathbb{e} \quad \text{where } \mathbb{e} \text{ is a } Y\text{-valued random variable representing observation noise.} \quad (2.2)$$

2.1. Ill-posedness and regularisation

A reconstruction method is formally a mapping $\mathcal{R} \colon Y \to X$ that approximates the inverse of the forward operator. An inverse problem is said to be (intrinsically) unstable, i.e., ill-posed, whenever the forward operator $\mathcal{A}$ is not continuously invertible with respect to the topologies of X and Y. As already indicated, such inverse problems cannot be reliably solved by merely enforcing consistency against data. Regularisation theory addresses this by balancing the need for data consistency against consistency with respect to a *prior model*. Defining an appropriate prior model and how to balance this against data consistency are key topics in regularisation theory.

Most regularisation methods are based on the functional analytic/frequentist formalisation in (2.1) of an inverse problem. Variational models offer a powerful framework for regularisation. The reconstruction method $\widehat{\mathcal{R}} \colon Y \to X$ is here defined as solving a variational problem:

$$\widehat{\mathcal{R}}(g) \in \underset{f \in X}{\arg\min} \Big\{ \mathcal{L}\big(\mathcal{A}(f), g\big) + \mathcal{S}_\theta(f) \Big\} \quad \text{for given data } g \in Y. \quad (2.3)$$

In the above, $\mathcal{L} \colon Y \times Y \to \mathbb{R}$ is the data discrepancy functional that quantifies the data consistency. It is usually taken as a suitable affine transformation of the data log-likelihood [9]. For an ill-posed problem, merely minimising $f \mapsto \mathcal{L}\big(\mathcal{A}(f), g\big)$, which translates to seeking the solution that is maximally consistent with data, is an unstable procedure. The *regularisation functional* $\mathcal{S}_\theta \colon X \to \mathbb{R}$ stabilises the reconstruction by penalising those candidate solutions that are not consistent with respect to some prior model. The latter is typically given in terms of a-priori information about the (unknown) ground truth signal $f_{\text{true}} \in X$ such as sparsity or some type of regularity, see [75,15,7] for an extensive survey of various options. The parameter θ (*regularisation parameter*) governs the balance between data consistency and having a solution consistent against the prior. It is low-dimensional, in many cases a scalar, and its choice depends on the magnitude of the noise in data.

The statistical formalisation of the inverse problem in (2.2) contains many of the variational models. More precisely, if the regularisation functional is proportional to the negative log of a density on X, then one can often interpret (2.3) as computing a maximum a posteriori estimator. An advantage with the statistical formalisation is that it opens up for other reconstruction methods (estimators) such as the posterior mean estimator that is known to be stable in most cases [55]. This estimator is given as $\mathbb{E}[\mathbb{f} \mid \mathbb{g} = g]$, a formulation that requires access to the (posterior) distribution of $(\mathbb{f} \mid \mathbb{g} = g)$. Using Bayes' theorem, one can in principle express this posterior distribution in terms of the data likelihood $(\mathbb{g} \mid \mathbb{f} = f)$, which is given by the physics of data acquisition, and the true (prior) distribution of $\mathbb{f}$, which among others generates the true unknown signal $f_{\text{true}} \in X$. An alternative formulation is to phrase the posterior mean as a Bayes estimator with respect to the squared L^2-risk. Stated more precisely, given a parametrised family $\{\mathcal{R}_\theta\}_{\theta \in \Theta}$ of admissible estimators $\mathcal{R}_\theta \colon Y \to X$, we consider the reconstruction operator

$$\mathcal{R}_{\hat{\theta}} \colon Y \to X \quad \text{where} \quad \hat{\theta} \in \underset{\theta \in \Theta}{\arg\min} \, \mathbb{E}_{(\mathbb{f}, \mathbb{g})} \Big[\big\| \mathcal{R}_\theta(\mathbb{f}) - \mathbb{g} \big\|_2^2 \Big]. \quad (2.4)$$

The above approximates the posterior mean, i.e., $\mathcal{R}_{\hat{\theta}}(g) = \mathbb{E}[\mathbb{f} \mid \mathbb{g} = g]$ whenever $\{\mathcal{R}_\theta\}_{\theta \in \Theta}$ has sufficient expressive power. Note also that the expectation in (2.4) is w.r.t. the joint law for $(\mathbb{f}, \mathbb{g})$ in $X \times Y$. Many supervised learning approaches for reconstruction are based on replacing the expectation in (2.4) with its empirical counterpart given by supervised training data, see [5, Section 5] for an extensive survey. In particular, it is common to use a deep neural network (DNN) with an appropriate architecture to parametrise the family of estimators $\{\mathcal{R}_\theta\}_{\theta \in \Theta}$. One can use unrolling to define such an architecture as outlined in Subsection 2.2.

Finally, besides unrolling, there exist also DNN architectures that are specifically tailored for approximating Fourier integral operators. Examples are [26,29] which develop a DNN architecture based on operator splitting techniques derived from multiscale numerical analysis. Another example is [49], which builds an interpretable DNN inspired by Fourier integral operators that approximate the wave physics. Its main focus is on using a loss based on optimal transport to learning the geometry of wave propagation captured by Fourier integral operators, which is implicit in the training data.

2.2. The Learned Primal-Dual network

A particular challenge that arises in deep-learning-based approaches for solving inverse problems in imaging is to handle the large-scale nature of the problem as both images and data typically result in high dimensional arrays once digitised. Many of these applications also lack a sufficient amount of training data, hence using a generic DNN architecture will result in a learned reconstruction operator that does not perform well in the given application. For such applications, it is preferable to use a DNN architecture that is domain-adapted.

If the trained DNN $\mathcal{R}_{\hat{\theta}} \colon Y \to X$ should correspond to a (learned) reconstruction operator for an inverse problem of the form in (2.2), then a natural domain adaptation is to account for the requirement that $\mathcal{R}_{\hat{\theta}} \approx \mathcal{A}^{-1}$ where $\mathcal{A} \colon X \to Y$ (forward operator) is handcrafted (not learned). Such domain adaptation can be achieved by choosing a DNN architecture that is given by unrolling a suitable iterative scheme, see [5, Section 4]. The highly successful Learned Primal-Dual network introduced in [2] provides state-of-the-art results for tomographic imaging. A similar unrolling technique is used in [12] to construct a novel convolutional DNN architecture, called ΨDONet, for learning pseudodifferential operators. This is applied to reconstruction in limited-angle tomography.

It is based on unrolling the non-linear primal-dual hybrid gradient method [16]. Stated in a general setting, the Learned Primal-Dual network is a DNN with an architecture specifically adapted for solving an inverse problem of the form in (2.2). The Learned Primal-Dual network is here a reconstruction operator $\mathcal{R}_\theta \colon Y \to X$ with $\theta = (\theta_0^d, \theta_0^p, \dots, \theta_N^d, \theta_N^p)$ that is defined as $\mathcal{R}_\theta(g) := f_N$ where f_N is given by the following finite recursive scheme initialised by $(f_0, h_0) \in X \times Y$:

$$
\begin{cases}
h_i := \Lambda_{\theta_i^d}^{\mathrm{dual}}(h_{i-1}, \mathcal{A}(f_{i-1}), g), \\
f_i := \Lambda_{\theta_i^p}^{\mathrm{primal}}\big(f_{i-1}, [\partial \mathcal{A}(f_{i-1})]^*(h_i)\big),
\end{cases}
\qquad \text{for } i = 1, \dots, N.
\tag{2.5}
$$

A typical initialisation $(f_0, h_0) \in X \times Y$ is $f_0 = 0$ and $h_0 = g$. Next, $\mathcal{A} \colon X \to Y$ is the (handcrafted) forward operator in (2.2) and $\partial \mathcal{A} \colon X \to L(X, Y)$ is its Fréchet derivative with $*$ denoting the dual. Finally, the operators

$$
\Lambda_{\theta_i^d}^{\mathrm{dual}} \colon Y \times Y \times Y \to Y \quad \text{and} \quad \Lambda_{\theta_i^p}^{\mathrm{primal}} \colon X \times X \to X
\tag{2.6}
$$

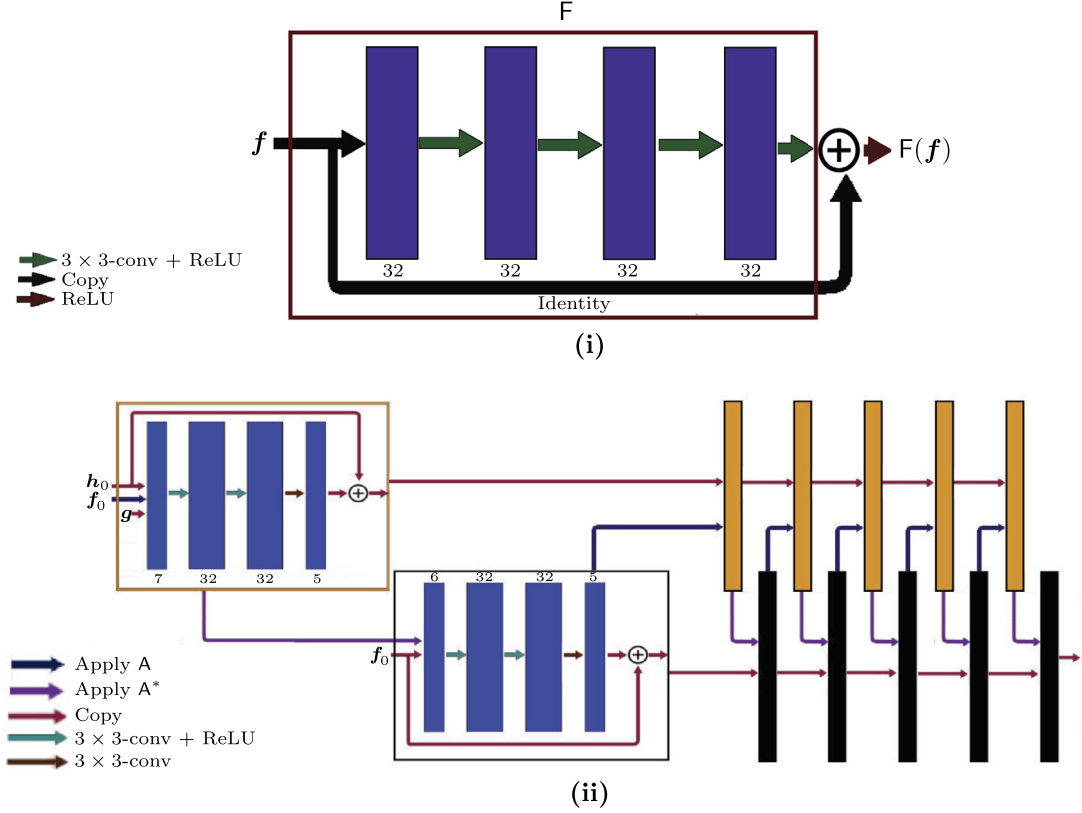are neural networks with suitable architectures.

Fig. 3. (i) The architecture for the two-dimensional convolutional ResNet in Definition 2.1. **(ii)** The full Learned Primal-Dual architecture, every yellow and black block corresponds to a two-dimensional convolutional ResNet. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

The Learned Primal-Dual network introduced in [2] for tomographic image reconstruction operates directly on arrays, which here play the role of functions evaluated on Cartesian grids. As outlined in Subsection 3.2, discretised images in tomography are represented by arrays in $\mathbb{R}^{n_1 \times n_2}$, whereas arrays in $\mathbb{R}^{m_1 \times m_2}$ are sinograms, which are discretisations of a real-valued function on $\Xi \subset \mathbb{R} \times (0, \pi)$. The Learned Primal-Dual network is then a mapping $\mathsf{R}_\theta \colon \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^{n_1 \times n_2}$ given as in (2.5), where the operators in (2.6) are convolutional residual neural networks (ResNets) [38]. The precise architecture used in [2] for these ResNets is given next, see also Fig. 3 for an illustration of these ResNets alongside the particular architecture for the corresponding Learned Primal-Dual network. We will in Subsection 4.2 extend these ResNets, and consequently the Learned Primal-Dual network, to operators that act on functions.

Definition 2.1 *(Discrete two-dimensional convolutional ResNet).* Assume arrays in $\mathbb{R}^{n_1 \times n_2}$ represent real-valued functions on $\mathbb{R}^2$ that are discretised at $n_1 \times n_2$ grid points. Let $\mathsf{W}_{\boldsymbol{\theta}_j, b_j} \colon (\mathbb{R}^{n_1 \times n_2})^{k_{j-1}} \to (\mathbb{R}^{n_1 \times n_2})^{k_j}$ for $j = 1, \dots, 4$ denote the following (discretised) convolutional affine operator:

$$\mathsf{W}_{\boldsymbol{\theta}_j, b_j}(\boldsymbol{f})(i_1, i_2, \tau) := b_j^\tau(i_1, i_2) + \sum_{l=1}^{k_{j-1}} (\boldsymbol{\theta}_j^{l,\tau} * \boldsymbol{f})(i_1, i_2) \quad \text{for } \tau \in \{1, \dots, k_j\} \text{ and } \boldsymbol{f} \in (\mathbb{R}^{n_1 \times n_2})^{k_{j-1}}. \quad (2.7)$$

In the above, $k_j \in \mathbb{N}$ for $j = 0, \dots, 4$ is the *numbers of channels in the j:th layer* with $k_4 = 1$, i.e., the final layer is always a single channel. Next,

$$\boldsymbol{\theta}_j := (\boldsymbol{\theta}_j^{l,\tau})_{l=1, \tau=1}^{k_{j-1}, k_j} \in (\mathbb{R}^{3 \times 3})^{k_{j-1} \times k_{j-1}} \quad \text{for } j = 1, \dots, 4$$

are the channel-wise 3×3 convolutional filters with $b_j \in (\mathbb{R}^{n_1 \times n_2})^{k_j}$ as corresponding channel-wise bias terms. The *residual convolutional network (ResNet)* operator is a mapping $\mathsf{ResNet} \colon (\mathbb{R}^{n_1 \times n_2})^{n_0} \to \mathbb{R}^{n_1 \times n_2}$ defined as

$$\mathsf{ResNet}(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_{n_0}) := \boldsymbol{f}_1 + \mathsf{F}(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_{n_0}) \quad \text{for } \boldsymbol{f}_1, \ldots, \boldsymbol{f}_{n_0} \in \mathbb{R}^{n_1 \times n_2},$$

where $\mathsf{F} \colon (\mathbb{R}^{n_1 \times n_2})^{n_0} \to \mathbb{R}^{n_1 \times n_2}$ is the operator

$$\mathsf{F}(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_{n_0}) := \big(\mathsf{W}_{\boldsymbol{\theta}_4, b_4} \circ \mathsf{ReLU} \circ \mathsf{W}_{\boldsymbol{\theta}_3, b_3} \circ \mathsf{ReLU} \circ \mathsf{W}_{\boldsymbol{\theta}_2, b_2} \circ \mathsf{ReLU} \circ \mathsf{W}_{\boldsymbol{\theta}_1, b_1} \big)(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_{n_0}).$$

In the above, $\mathsf{ReLU}(x) := \max\{x, 0\}$ is applied component-wise.

3. Tomographic imaging and the Radon transform

Planar tomographic imaging aims to recover a 2D image (signal) of the interior of an object from corresponding tomographic data. Thus, elements in X are real-valued functions on a fixed domain $\Omega \subset \mathbb{R}^2$ that represent possible 2D images, and X itself is some suitable function space, e.g., $X \subset L^2(\mathbb{R}^2)$.

Next, many tomographic imaging modalities rely on probing the object with x-rays. In the planar setting, these rays are all contained in a cross sectional hyperplane through the object. After adopting a simplified physics model for the interaction between x-rays and the object, measured data (after appropriate preprocessing) can be viewed as noisy digitised samples of the Radon transform of the aforementioned signal.

Definition 3.1 *(Radon transform).* The *(planar) Radon transform* of $f \colon \mathbb{R}^2 \to \mathbb{R}$ is defined as

$$\mathcal{A}(f)(s, \theta) := \int\limits_{-\infty}^{\infty} f\big(s\omega(\theta) + t\omega(\theta)^{\perp}\big) dt, \quad \text{for } (s, \theta) \in \mathbb{R} \times (0, \pi). \tag{3.1}$$

In the above, $\omega(\theta) := (\cos\theta, \sin\theta)$ is the unitary vector with orientation described by the angle θ with respect to the x_1-axis and $\omega(\theta)^{\perp} := (-\sin\theta, \cos\theta)$.

Remark 3.2. There do exist integrable functions f for which there is a direction $\omega(\theta) \in \mathbb{S}^1$ and a point $s \in \mathbb{R}$ such that $\mathcal{A}(f)(s, \theta)$ does not exist. Nonetheless, it is possible to show that $\mathcal{A}(f)(s, \theta)$ exists almost everywhere when f is an integrable function [61, Proposition 2.38].

Note that (s, θ) represents the line $t \mapsto s\omega(\theta) + t\omega(\theta)^{\perp}$ that is orthogonal to $\omega(\theta) \in \mathbb{S}^1$ with (signed) distance $s \in \mathbb{R}$ to the origin, so $\mathcal{A}(f)$ is a function on lines in $\mathbb{R}^2$. Since data in tomographic imaging can be seen as noisy samples of $\mathcal{A}(f)$, the data space Y is an appropriate space of real-valued functions on lines in $\mathbb{R}^2$ representing non-digitised data. The angle $\theta \in (0, \pi)$ governs the direction of the x-ray that probes the object. It typically varies during the tomographic data acquisition and limited-angle tomography is the case when θ is restricted to some interval $I \subset (0, \pi)$.

Definition 3.3 *(Tomographic reconstruction).* *Tomographic (image) reconstruction* is the inverse problem of recovering a *ground truth* image $f_{\text{true}} \in L^2(\mathbb{R}^2)$ from noisy measurements $g \in L^2(\Xi)$ for some open set $\Xi \subset \mathbb{R} \times [0, \pi]$. Here, g is a single sample of the $L^2(\Xi)$-valued random variable g in (2.1) (or (2.2) for the statistical formulation) with $\mathcal{A}$ denoting the Radon transform (Definition 3.1).

3.1. Basic properties of the Radon transform

From a functional analytic viewpoint, the Radon transform is a linear operator that maps functions on $\mathbb{R}^2$ to functions on the open set $\mathbb{R} \times (0, \pi)$ (representing lines in $\mathbb{R}^2$). Continuity of the Radon transform depends on the domain of the functions chosen. As an example, the Radon transform is a bounded map on $L^1(\mathbb{R}^2)$, implying that it is a continuous linear operator on $L^1(\mathbb{R}^2)$ [61, Corollary 3.25], it is however unbounded on $L^2(\mathbb{R}^2)$. The Radon transform is also invertible but its inverse is not continuous on $L^1(\mathbb{R} \times (0, \pi))$ [40], yielding that the tomographic image reconstruction problem in Definition 3.3 is an ill-posed inverse problem.

We will consider the Radon transform on Schwartz functions $\mathcal{S}(\mathbb{R}^2)$, i.e., rapidly decreasing and smooth functions with the typical locally convex topology. It is well-known that the Fourier transform maps $\mathcal{S}(\mathbb{R}^2)$ onto itself. An analogous result holds for the Radon transform, namely that $\mathcal{A} \colon \mathcal{S}(\mathbb{R}^2) \to \mathcal{S}(\mathbb{R} \times (0, \pi))$ is a linear one-to-one mapping [39, Theorem 2.4].

Remark 3.4. $\mathcal{S}(\Omega)$ for some open set $\Omega \subset \mathbb{R}^2$ is defined as the set of functions on Ω such that their extension by 0 to all of $\mathbb{R}^2$ is a Schwartz function. In particular, this defines the space of Schwartz functions on any open set $\Xi \subset \mathbb{R} \times (0, \pi)$.

Next, we introduce the back-projection as the dual to the Radon transform in a sense analogous to the way the adjoint of a linear transformation on Euclidean space is dual to the original transformation.

Definition 3.5 *(Back-projection).* The *back-projection* of $g \colon \mathbb{R} \times (0, \pi) \to \mathbb{R}$ is the function $\mathcal{A}^*(g) \colon \mathbb{R}^2 \to \mathbb{R}$ defined as

$$\mathcal{A}^*(g)(x) := \int_0^\pi g(x \cdot \omega(\theta), \theta) \, d\theta \quad \text{for } x \in \mathbb{R}^2. \tag{3.2}$$

The back-projection maps a function g on lines in $\mathbb{R}^2$ to a function on points in $x \in \mathbb{R}^2$ by simply averaging g over all lines that pass through x. A simple calculation shows that the back-projection is the dual to the Radon transform [61, Theorem 2.75]:

$$\langle \mathcal{A}(f), g \rangle = \langle f, \mathcal{A}^*(g) \rangle. \tag{3.3}$$

The inner product on the right-hand side refers to the natural inner product on $L^2(\mathbb{R}^2)$, whereas the inner product on the left-hand side is the natural inner product on $L^2(\mathbb{R} \times (0, \pi))$. This product is well defined in tempered distributions since $\mathcal{S}(\mathbb{R}^2) \subset L^2(\mathbb{R}^2)$ (respectively $\mathcal{S}(\mathbb{R} \times (0, \pi)) \subset L^2(\mathbb{R} \times (0, \pi)))$, [76].

The duality in (3.3) can be used to extend the Radon transform to various classes of distributions, like compactly supported distributions [39] and tempered distributions [32,59].

In particular, one can define the Radon transform on a tempered distribution $f \in \mathcal{S}'(\mathbb{R}^2)$ as

$$\mathcal{A}(f)(\phi) := f\big(\mathcal{A}^*(\phi)\big) \quad \text{for all } \phi \in \mathcal{S}(\mathbb{R} \times (0, \pi)).$$

The extension of $\mathcal{A}^*$ to $\mathcal{S}(\mathbb{R} \times (0, \pi))$ is defined analogously and (following [61, Section 2.9.3.2 and 4.3.1]) one can in addition show that,

$$\begin{aligned} \mathcal{A} \colon \mathcal{S}'(\mathbb{R}^2) \to \mathcal{S}'(\mathbb{R} \times (0, \pi)) \quad &\text{is a topological isomorphism,} \\ \mathcal{A}^* \colon \mathcal{S}'(\mathbb{R} \times (0, \pi)) \to \mathcal{S}'(\mathbb{R}^2) \quad &\text{is surjective.} \end{aligned} \tag{3.4}$$

In limited-angle tomography, data is given on lines contained in some open subset $\Xi \subset \mathbb{R} \times [0, \pi)$. The *partial* Radon transform is then defined as

$$\mathcal{A}_{\Xi} := P_{\mathcal{S}'(\Xi)} \circ \mathcal{A},$$

where $P_{\mathcal{S}'(\Xi)}$ is the restriction of $\mathcal{S}'(\mathbb{R}^2)$ to $\mathcal{S}'(\Xi)$. Note that the restriction of a tempered distribution $\mathcal{S}'(\mathbb{R}^2)$ to $\mathcal{S}(\Xi)$, where $\Xi \subset \mathbb{R}^2$ is open, is well defined and it corresponds to a tempered distribution in $\mathcal{S}'(\Xi)$. The corresponding (restricted) back-projection is simply defined as in (3.2) but by setting g to 0 on lines not contained in Ξ. In particular, if $\Xi := \mathbb{R} \times I \subset \mathbb{R} \times (0, \pi)$ for some open interval $I \subset (0, \pi)$, then

$$\mathcal{A}_{\Xi}^*(g)(x) = \int_I g(x \cdot \omega(\theta), \theta) \, d\theta \quad \text{for } x \in \mathbb{R}^2. \tag{3.5}$$

3.2. Discretisation

An image (signal) $f \colon \Omega \to \mathbb{R}$ defined on a domain $\Omega \subset \mathbb{R}^2$ is in tomographic imaging typically digitised as $\boldsymbol{f} = (f_{i,j})_{i,j=1}^{n_1,n_2}$ where $f_{i,j} := f(x_{i,j})$ for $x_{i,j} \in \Omega$. Similarly, a sinogram $g \in Y$ is digitised by an array $\boldsymbol{g} \in \mathbb{R}^{m_1 \times m_2}$ by evaluating g on $m_1 \times m_2$ grid points in $\Xi \subset \mathbb{R} \times (0, \pi)$ that are given by the data acquisition protocol used in the actual scanner. In the sequel, we will use boldface notation for the discrete signals. The m_1 grid points in $\mathbb{R}$ would simply correspond to detector elements in the scanner, whereas the m_2 grid points in $(0, \pi)$ correspond to directions used in the actual scanning.

The (discretised) Radon transform and back-projection are then mappings

$$\mathsf{A} \colon \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{m_1 \times m_2} \quad \text{and} \quad \mathsf{A}^* \colon \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^{n_1 \times n_2}. \tag{3.6}$$

We wish to remark that there exist sophisticated numerical schemes for evaluating these mappings in an accurate, yet computationally feasible manner, see, e.g., [48,57,87,45,86].

It is now possible to formulate the tomographic reconstruction problem entirely in the discrete setting. This is essentially a digitised version of the tomographic reconstruction problem in Definition 3.3, i.e., to recover the digital image $\boldsymbol{f}_{\mathrm{true}} \in \mathbb{R}^{n_1 \times n_2}$ from noisy measurements $\boldsymbol{g} \in \mathbb{R}^{m_1 \times m_2}$ where

$$\boldsymbol{g} = \mathsf{A}(\boldsymbol{f}_{\mathrm{true}}) + \boldsymbol{e} \quad \text{for some (unknown) observation error } \boldsymbol{e} \in \mathbb{R}^{m_1 \times m_2}.$$

One could then proceed to develop regularised reconstruction methods for the above finite dimensional inverse problem. This 'discretise first then reconstruct' approach was common in the 1970's with the advent of iterative schemes with early stopping, like algebraic reconstruction techniques (ART) [35], simultaneous ART (SART) [3] or simultaneous iterative reconstruction technique (SIRT) [33]. The success of FBP and subsequent variational models did however point to the advantage of developing reconstruction methods for the continuum formulation, which then are discretised. This is also the approach we take in our work, since the microlocal analysis that we will rely on is naturally formulated in the continuum setting.

3.3. Microlocal analysis in tomography

A central part of microlocal analysis is the study of how singularities are transformed by specific operators. This is not possible given only the singular support that describes the location of the singularities. Instead, an appropriate notion of singularity needs to include information about directions in the Fourier domain that causes the singularity. This leads to the wavefront set that we define below.

Definition 3.6 *(Wavefront set).* Let $u \in \mathcal{S}'(\mathbb{R}^n)$ and $N \in \mathbb{N}$. A point $(x, \xi) \in \mathbb{R}^n \times \mathbb{S}^{n-1}$ is an *N-regular directed point* of u if there exist an open neighbourhood U_x of x, a conical neighbourhood V_ξ of ξ, and a smooth cut-off function $\psi \in \mathcal{S}(\mathbb{R}^n)$ with $\operatorname{supp}\psi \subset U_x$ and $\psi(x) = 1$ such that the following holds for some $C_N > 0$:

$$\left| \widehat{\psi u}(\eta) \right| \leq C_N \left(1 + |\eta| \right)^{-N} \quad \text{for all } \eta \in \mathbb{R}^n \text{ with } \eta/|\eta| \in V_\xi.$$

The *N-wavefront set* $\mathrm{WF}_N(u)$ is the complement of the set of all *N*-regular directed points. Finally, the *wavefront set* $\mathrm{WF}(u)$ of u is defined as

$$\mathrm{WF}(u) := \bigcup_{N \in \mathbb{N}} \mathrm{WF}_N(u).$$

From the above, we see that the wavefront set $\mathrm{WF}(u)$ is the set of all position-orientation pairs in $\Omega \times \mathbb{S}^1$ along which u is non-smooth.

Remark 3.7. The general setting in [41, Section 8.1], where functions/distributions are defined on manifolds, views the wavefront set as a conical subset of the cotangent bundle. This is equivalent to Definition 3.6 when the manifold is an open domain in $\mathbb{R}^2$.

3.3.1. The canonical relation for the Radon transform and its back-projection

For many operators arising in applications, it is possible to describe how they transform the wavefront set. This description is called *(microlocal) canonical relation*. It plays an important role in inverse problems, since it allows one to relate singularities in data to those in the unknown signal.

The canonical relation for the Radon transform in Definition 3.1 at tempered distribution $f \in \mathcal{S}'(\mathbb{R}^2)$ is a precise relationship between $\mathrm{WF}(\mathcal{A}(f))$ and $\mathrm{WF}(f)$. If $\mathcal{P}$ denotes taking the power set, then this can be expressed as a map

$$K \colon \mathcal{P}\Big(\big(\mathbb{R} \times (0, \pi) \big) \times \mathbb{S}^1 \Big) \to \mathcal{P}(\mathbb{R}^2 \times \mathbb{S}^1), \quad \text{where} \quad K\Big(\mathrm{WF}\big(\mathcal{A}(f) \big) \Big) = \mathrm{WF}(f) \quad \text{for } f \in \mathcal{S}'(\mathbb{R}^2). \quad (3.7)$$

In limited-angle tomography we only have access to the Radon transform on an open subset $\Xi \subset \mathbb{R} \times (0, \pi)$. The canonical relation then holds for the so-called *visible* wavefront set of a function/distribution f that is given by

$$\mathrm{WF}^{\mathrm{vis}}(f) := \mathrm{WF}(f) \cap K(\Xi). \quad (3.8)$$

Following [50], we next provide a more precise characterisation of the canonical relation in terms of a mapping between wavefront sets in image and sinogram, respectively.

Theorem 3.8. *The canonical relation for the Radon transform* $\mathcal{A} \colon \mathcal{S}'(\mathbb{R}^2) \to \mathcal{S}'\big(\mathbb{R} \times (0, \pi) \big)$ *at* $f \in \mathcal{S}'(\mathbb{R}^2)$ *can be represented by the mapping*

$$\mathrm{Can}_{\mathcal{A}(f)} \colon \mathbb{R}^2 \times \mathbb{S}^1 \to \big(\mathbb{R} \times (0, \pi) \big) \times \mathbb{S}^1$$

defined as

$$\mathrm{Can}_{\mathcal{A}(f)}\big(x; \omega(\theta) \big) := \Big(\big(x \cdot \omega^\perp(\theta), \theta + \pi/2 \big); \omega\Big(\arctan\big(-x \cdot \omega(\theta) \big) \Big) \Big) \quad \textit{for } \big(x; \omega(\theta) \big) \in \mathrm{WF}(f), \quad (3.9)$$

with $\omega(\theta) := (\cos\theta, \sin\theta)$ *and* $\omega(\theta)^\perp := (-\sin\theta, \cos\theta)$. *This means,*

$$\big(x; \omega(\theta) \big) \in \mathrm{WF}(f) \iff \mathrm{Can}_{\mathcal{A}(f)}\big(x; \omega(\theta) \big) \in \mathrm{WF}\big(\mathcal{A}(f) \big).$$

Proof. Following [50, Definition 7] a Fourier integral operator, $\mathcal{P} \colon \mathcal{S}'(\mathbb{R}^2) \to \mathcal{S}(\mathbb{R} \times (0, \pi))$, has the form

$$\mathcal{P}f(y) = \int\limits_{\xi \in \mathbb{R}^2} \int\limits_{x \in \mathbb{R}^2} e^{i\phi(y, x, \xi)} p(y, x, \xi) dx d\xi \quad \text{for } y \in \mathcal{S}(\mathbb{R} \times (0, \pi)), \quad (3.10)$$

where $p, \phi \in C^\infty(\mathbb{R}^2 \times \mathbb{R} \times (0, \pi) \times \mathbb{R}^2)$ are the amplitude and the phase functions, respectively (see [50, Definition 4 and Definition 6]). Using [50, Definition 7] and duality gives the following expression for the canonical relation of $\mathcal{P} \colon \mathcal{S}'(\mathbb{R}^2) \to \mathcal{S}'(\mathbb{R} \times (0, \pi))$, given by

$$\mathrm{WF}\big(\mathcal{A}(f)\big) \subset C_\phi \circ \mathrm{WF}(f) \quad \text{for } f \in \mathcal{S}'(\mathbb{R}^2), \tag{3.11}$$

where

$$C_\phi := \Big\{ \Big((y; \partial_y \phi(y, x, \xi)), (x, -\partial_x \phi(y, x, \xi)) \Big) : (y, x, \xi) \in \Sigma_\phi \Big\}$$

with

$$\Sigma_\phi := \Big\{ ((s, \theta), x, \xi) \in (\mathbb{R} \times [0, 2\pi)) \times \mathbb{R}^2 \times \mathbb{R} \setminus \{0\} : \partial_\xi \phi\big((s, \theta), x, \xi\big) = 0 \Big\}.$$

We now aim to write the Radon transform $\mathcal{A}$ in the form (3.10). By the Fourier slice theorem [50, Theorem 2], the Radon transform $\mathcal{A} \colon \mathcal{S}'(\mathbb{R}^2) \to \mathcal{S}'(\mathbb{R} \times (0, \pi))$ can be written in the integral form as

$$\mathcal{A}(f)(s, \theta) = \frac{1}{2\pi} \int\limits_{\xi \in \mathbb{R}} \int\limits_{x \in \mathbb{R}^2} e^{i(s - (x \cdot \omega(\theta)))\xi} f(x) dx d\xi \quad \text{for } s \in \mathbb{R} \text{ and } \theta \in (0, \pi), \tag{3.12}$$

where $\omega(\theta) := (\cos \theta, \sin \theta)$. Hence, the Radon transform is a Fourier integral operator with phase function $\phi((s, \theta), x, \xi) := (s - (x \cdot \omega(\theta)))\xi$ and amplitude $p(y, x, \xi) := 1/(2\pi)$. Thus the canonical relation of $\mathcal{A} \colon \mathcal{S}'(\mathbb{R}^2) \to \mathcal{S}'(\mathbb{R} \times (0, \pi))$ is given by (3.11).

We then derive the exact form of C_ϕ by computing the derivatives of ϕ, which are

$$\begin{aligned}
\partial_x \phi\big((s, \theta), x, \xi\big) &= -\xi \omega(\theta), \\
\partial_{(s, \theta)} \phi\big((s, \theta), x, \xi\big) &= \xi\big(1, -x \cdot \omega^\perp(\theta)\big), \\
\partial_\xi \phi\big((s, \theta), x, \xi\big) &= \big(s - x \cdot \omega(\theta), 0\big).
\end{aligned} \tag{3.13}$$

Notice that $\partial_x \phi$ and $\partial_{(s, \theta)} \phi$ are not zero for $\xi \neq 0$, which means that the phase function is non-degenerate. This confirms that $\mathcal{A}$ is a Fourier integral operator.

Next, $\mathcal{A}$ is of order $m = -1/2$ with an amplitude function $p((s, \theta), x, \xi) = 1/(2\pi)$ that is homogeneous of degree zero, implying that $\mathcal{A}$ is elliptic. Using the derivatives (3.13), Σ_ϕ is given by

$$\Sigma_\phi = \Big\{ ((s, \theta), x, \xi) \in (\mathbb{R} \times [0, 2\pi)) \times \mathbb{R}^2 \times \mathbb{R} \setminus \{0\} \colon s - x \cdot \omega(\theta) = 0 \Big\}.$$

Therefore, the canonical relation can be represented by the coordinate mapping

$$\begin{aligned}
((s, \theta), x, \xi) &\mapsto \Big(\big((x \cdot \omega(\theta), \theta); \partial_{(s, \theta)\phi} \big), (x, -\partial_x \theta) \Big) \\
&= \Big(\big((x \cdot \omega(\theta), \theta); \xi(1, -x \cdot \omega(\theta)^\perp) \big), (x, \xi \omega(\theta)) \Big).
\end{aligned} \tag{3.14}$$

Now, let $(x; \omega(\theta)) \in \mathrm{WF}(f)$ be an oriented singular point of f. By (3.14) we obtain that

$$(x; \omega(\theta)) \in \mathrm{WF}(f) \to \Big((x \cdot \omega(\theta)^\perp, \theta + \pi/2); \omega\big(\arctan(-x \cdot \omega(\theta))\big) \Big) \in \mathrm{WF}\big(\mathcal{A}(f)\big).$$

Finally, [68, Theorem 6.3] gives

$$\big(x; \omega(\theta)\big) \in \mathrm{WF}(f) \iff \Big(\big(x \cdot \omega(\theta)^{\perp}, \theta + \pi/2\big); \omega\Big(\arctan\big(-x \cdot \omega(\theta)\big)\Big)\Big) \in \mathrm{WF}\big(\mathcal{A}(f)\big). \tag{3.15}$$

This concludes the proof. $\quad\square$

We next focus on the propagation of singularities performed by the adjoint Fréchet derivative, of the Radon transform, which is the back-projections operator $\mathcal{A}^*$ in Definition 3.5. Using [50, Theorem 13] we know that $\mathcal{A}^*$ is also a Fourier integral operator and

$$\mathcal{A}^* \mathcal{A}(f)(x) = \int_{\mathbb{R}^2} e^{ix\cdot\xi} \frac{2}{\|\xi\|} \widehat{f}(\xi) d\xi = \frac{1}{\pi} \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} e^{i(x-y)\cdot\xi} \frac{1}{\|\xi\|} f(y) dy d\xi, \quad \text{for } f \in \mathcal{S}(\mathbb{R}^2) \text{ and } x \in \mathbb{R}^2. \tag{3.16}$$

In the next proposition, we use (3.16) to introduce the corresponding mapping associated with the canonical relation for $\mathcal{A}^*$ in a similar fashion to Theorem 3.8.

Proposition 3.9. *The canonical relation for the back-projection operator* $\mathcal{A}^* \colon \mathcal{S}'\big(\mathbb{R} \times (0, \pi)\big) \to \mathcal{S}'(\mathbb{R}^2)$ *in* (3.2) *at* $g \in \mathcal{S}'\big(\mathbb{R} \times (0, \pi)\big)$ *can be represented by the mapping*

$$\mathrm{Can}_{\mathcal{A}^*(g)} \colon \mathrm{WF}(g) \to \mathrm{WF}\big(\mathcal{A}^*(g)\big),$$

which is defined at $\big((s, \theta); \omega(\vartheta)\big) \in \mathrm{WF}(g)$ *as*

$$\mathrm{Can}_{\mathcal{A}^*(g)}\big((s, \theta); \omega(\vartheta)\big) := \big((s \cos\theta - \tan\vartheta \sin\theta, s \sin\theta + \tan\vartheta \cos\theta); \theta - \pi/2\big), \tag{3.17}$$

where $\omega(\theta) := (\cos\theta, \sin\theta)$. *This means,*

$$\big((s, \theta); \omega(\vartheta)\big) \in \mathrm{WF}(g) \iff \mathrm{Can}_{\mathcal{A}^*(g)}\big((s, \theta); \omega(\vartheta)\big) \in \mathrm{WF}\big(\mathcal{A}(f)\big).$$

Proof. Note first that (3.16) implies that the operator $\mathcal{A}^* \mathcal{A} \colon \mathcal{S}(\mathbb{R}^2) \to \mathcal{S}(\mathbb{R}^2)$ is an elliptic pseudodifferential operator with amplitude function $p(y, x, \xi) := 1/\|\xi\|$. By duality we can extend this to a mapping $\mathcal{A}^* \mathcal{A} \colon \mathcal{S}'(\mathbb{R}^2) \to \mathcal{S}'(\mathbb{R}^2)$. In addition, the pseudolocal property of pseudodifferential operators (see Theorem 4.6) implies that $\mathcal{A}^* \mathcal{A}$ will preserve the wavefront set of functions in $\mathcal{S}(\mathbb{R}^2)$, i.e., $\mathrm{WF}(\mathcal{A}^* \mathcal{A}(f)) = \mathrm{WF}(f)$. This allows us to represent the canonical relation for the inverse mapping in terms of the canonical relation mapping for $\mathcal{A}$. Finally, by inverting the mapping implicit in (3.15), for $g \in \mathcal{S}'\big(\mathbb{R} \times (0, \pi)\big)$ we obtain that

$$\big((s, \theta); \omega(\vartheta)\big) \in \mathrm{WF}(g) \iff \big((s \cos\theta - \tan\vartheta \sin\theta, s \sin\theta + \tan\vartheta \cos\theta); \theta - \pi/2\big) \in \mathrm{WF}\big(\mathcal{A}^*(g)\big). \quad\square \tag{3.18}$$

3.4. Computational microlocal analysis

Our aim is to develop a computational counterpart to microlocal analysis that is based on defining a notion of a 'digital wavefront set' and also to provide computational means for extracting such an object from an array that represents a discretised function. This turns out to be theoretically and computationally challenging. The definition of a digital wavefront set we will exploit in our work is based on 'discretising' Definition 3.6. More precisely, the *digital wavefront set* of an array is defined as follows.

Definition 3.10 *(Digital wavefront set).* Let the array $\boldsymbol{u} \in \mathbb{R}^N$ represent $u \colon \Omega \to \mathbb{R}$ at sample points $\Omega = \{x_1, \ldots, x_N\} \subset U$ in some fixed domain $U \subset \mathbb{R}^n$, i.e., $\boldsymbol{u} = (u_1, \ldots, u_N)$, where $u_j := u(x_j)$ for $x_j \in \Omega$. Next, let $\Sigma = \{\omega_1, \ldots, \omega_M\} \subset \mathbb{S}^{n-1}$ denote a fixed set of M points and $\{U_{j,k}\}_{j=1,\ldots,N, k=1,\ldots,M}$ is a set of neighbourhoods in $\Omega \times \mathbb{S}^{n-1}$ where $(x_j, \omega_k) \in U_{j,k}$ and $U \times \mathbb{S}^{n-1} \subset \bigcup_{j,k} U_{j,k}$. Then, the *digital wavefront set* of the array $\boldsymbol{u}$ is defined as $\mathsf{DWF}(\boldsymbol{u}) \subset \Omega \times \Sigma$ where

$$(x_j, \omega_k) \in \mathsf{DWF}(\boldsymbol{u}) \quad \text{if } U_{j,k} \text{ intersects } \mathrm{WF}(u) \subset U \times \mathbb{S}^{n-1} \text{ non-trivially.} \tag{3.19}$$

The *visible digital wavefront set* is defined in a similar manner.

Remark 3.11. Note that digital wavefront set of Definition 3.10, depends on the choice of the spatial sampling points Ω, as well as the sampled directions Σ and their associated neighbourhoods $U_{j,k}$. In the two dimensional applications that we consider below, Ω will contain points on a Cartesian grid in the unit cube, i.e., $\Omega = \delta \times \mathbb{Z}^2 \cap [0,1]^2$ for an appropriate value of $\delta > 0$, Σ will correspond to 180 equidistant points on the 1-sphere. Moreover, the neighbourhoods will be the natural choice $U_{j,k} = U_j \times O_k$, where the U_j are cubes of sidelength δ centred at x_j and O_k are connected disjoint subsets of $\mathbb{S}^1$ of equal diameter centred at ω_k. The reason for the generality in the definition above is that for higher dimensional versions of digital wavefront sets, such a natural choice may not exist.

The definition also applies to arrays $\boldsymbol{f} \in \mathbb{R}^{n_1 \times n_2}$ and $\boldsymbol{g} \in \mathbb{R}^{m_1 \times m_2}$ representing images and sinograms in tomographic imaging (Subsection 3.2). Note finally that Definition 3.10 suggests a natural way to represent a digital wavefront set as a multi-channel 'image'. Simply assign M binary channels to each sample point in $x_j \in \Omega$ and set the k:th channel at that point to 1 if $(x_j, \omega_k) \in \mathsf{DWF}(\boldsymbol{u})$, otherwise set the value of that channel to 0.

Having defined a notion of a digital wavefront set of an array that represents a discretised function, a natural task that follows is to computationally extract such an object. This is however not possible to do in a mathematically consistent way as there is *no correspondence between the digitisation of a function and its digital wavefront set* [4, Theorem 3.3]. The approach taken in [4] is therefore to view the task of extracting the digital wavefront as a statistical estimation problem. More precisely, extracting the digital wavefront set is phrased as computing the probability distribution of possible digital wavefront sets. A *single* digital wavefront set can then be computed by choosing at each point of the digitised function the wavefront set orientation with the highest probability. Much of [4] is devoted to developing a deep-learning-based approach for this task. A key part is the development of DeNSE, which is a DNN with an architecture specifically suited for computing probability distributions of digital wavefront sets of functions represented by their discrete shearlet transforms. DeNSE was in [4] successfully applied to extract digital wavefront sets of functions discretised by shearlets in both image and sinogram space in tomographic imaging.

4. Microlocal analysis of the Learned Primal-Dual network

This section aims to derive the canonical relation for the non-linear operator given by the Learned Primal-Dual network defined in (2.5) with $\mathcal{A}$ as the Radon transform (Definition 3.1). The relation allows to describe how the Learned Primal-Dual network transforms the digital wavefront set of an input array that is a discretisation of a function/distribution representing data, which is a key step in our approach to tomographic image reconstruction, outlined in Subsection 1.1.

4.1. Overview of approach

The starting point is the discretised tomographic inverse problem given in Subsection 3.2, where data and images are arrays in $\mathbb{R}^{m_1 \times m_2}$ and $\mathbb{R}^{m_1 \times m_2}$, respectively. Assume next that the digital wavefront set is known for an input array $\boldsymbol{g} \in \mathbb{R}^{m_1 \times m_2}$, representing measured data. Our aim is to compute the visible digital wavefront set of the array $\mathsf{R}_\theta(\boldsymbol{g}) \in \mathbb{R}^{n_1 \times n_2}$ representing the reconstructed image, i.e., to compute $\mathsf{DWF}^{\mathrm{vis}}(\mathsf{R}_\theta(\boldsymbol{g}))$ where $\mathsf{R}_\theta \colon \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^{n_1 \times n_2}$ is the non-linear Learned Primal-Dual network in (2.5) for tomographic reconstruction that was first introduced in [2]. This is a DNN that is made up of stacked convolutional residual neural networks of the form (2.6) that in our setting are given as in Definition 2.1 with $\mathcal{A} = \mathsf{A} \colon \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{m_1 \times m_2}$ denoting the discretised Radon transform in (3.6).

One approach to derive the mapping between $\boldsymbol{g}$ and $\mathsf{DWF}^{\mathrm{vis}}(\mathsf{R}_\theta(\boldsymbol{g}))$ is to work entirely within the discrete setting. Such an approach would require us to describe how the discrete Radon transform along with its adjoint transforms the digital wavefront set of an array. An alternative approach is to utilise the rich and well-developed microlocal theory for the (continuum) Radon transform outlined in Subsection 3.3. In particular, this theory describes how the Radon transform, and hence also its adjoint (back-projection), modifies visible wavefront sets in limited-angle tomography. However, such an attempt at leveraging on the continuum theory requires us to formulate a continuum version of the Learned Primal-Dual network.

In Subsection 4.2, we derive the non-linear operator, which is a continuum version of the Learned Primal-Dual network. This results from replacing every step used in the construction of $\mathsf{R}_\theta \colon \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^{n_1 \times n_2}$ with a natural corresponding continuum version, resulting in an operator $\mathcal{R}_\theta \colon Y \to X$, where Y and X are not necessarily finite dimensional vector spaces. The key part is to assemble appropriate continuum versions of the convolutional residual neural networks in (2.6). While a continuous convolution seems to be the most natural correspondence to a discrete convolution, this begs the question, which continuous filter to choose. Alternatively, following [74] one can view each discrete convolution as a discretisation of a corresponding fourth-order differential operator with coefficients that relate precisely to the discrete filter. Naturally, the coordinate-wise application of an activation function to a discrete input corresponds to the composition with that activation function in the continuous realm. Since the differential operators do not necessarily yield L^2 functions but only distributions, this concept needs to be extended to tempered distributions (see Definition 4.1). Finally, the residual block consists only of summation, which naturally translates to summation of continuous inputs.

Subsection 4.3 shows how each of the above operations affect the wavefront set of a function or tempered distribution. A key result is Theorem 4.11 that analyses the action of ReLU. When combined, these provide a precise theoretical description of the way a continuum version of the Learned Primal-Dual network transforms the wavefront set. Its canonical relation can then be used to derive the associated digital canonical relation for the Learned Primal-Dual network in [2].

4.2. Continuum Learned Primal-Dual network

Towards the end of Subsection 2.2, we specified the architecture for the Learned Primal-Dual network that was introduced in [2] for tomographic reconstruction. This is a mapping

$$\mathsf{R}_\theta \colon \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^{n_1 \times n_2}$$

given by a DNN that acts on arrays in finite dimensional vector spaces. In inverse problems these spaces typically represent discretised functions as outlined in Subsection 3.4. The aim here is to formulate a natural continuum version of $\mathcal{R}_\theta \colon Y \to X$ that acts on functions or distributions that are not necessarily discretised.

As outlined in the overview in Subsection 4.1, a key step lies in appropriately extending the residual convolutional networks in (2.5). To this end, we replace every discrete operation of the Learned Primal-Dual by a continuum analogue, i.e., an operator that takes as an input a distribution. In principle, only four types of operations happen in the definition of the Learned Primal-Dual. These are the convolutions, the application of a ReLU, the application of a discretised Radon transform or the adjoint of the Fréchet derivative of the Radon transform, and taking sums of functions, either between channels in the convolution or due to residual connections. The discretised Radon transform as well as the back-projection, which is the adjoint of its Fréchet derivative, already stem from continuum counterparts which are the natural replacement. Furthermore, a sum of discrete images naturally corresponds to sums of continuum images. However, it is not immediately clear how to extend a discrete convolution to a continuum convolution. For example, replacing discrete convolutions by convolutions defined over $\mathbb{R}^2$ begs the question how the continuum convolution kernel should be chosen. We will see in the subsection below, that there exists a

natural replacement of the convolution operator, which however, necessitates that we work with distributions instead of L^2 functions. Because of this, also the application of the ReLU will need to be generalised to distributions. In the following two subsections we discuss the continuum counterparts of convolution and ReLU. Thereafter, we present the continuum Learned Primal-Dual network.

4.2.1. From discrete convolutions to differential operators

To find an appropriate continuum counterpart to the discrete convolution step in (2.7), we interpret the discrete convolution as a discretisation of a differential operator. This approach is inspired by [74].

Consider a continuum image $f \in L^2(\mathbb{R}^2)$ with discretisation $\boldsymbol{f} \in \mathbb{R}^{n_1 \times n_2}$. Concretely, let

$$
\boldsymbol{f} = \begin{pmatrix} f_{11} & \cdots & f_{1n_1} \\ \vdots & \ddots & \vdots \\ f_{n_2 1} & \cdots & f_{n_1 n_2} \end{pmatrix}.
$$

Furthermore, let $\mathsf{K}_{\boldsymbol{\theta}}$ be a 3×3 (discretised) convolutional operator parametrised by the filter:

$$
\boldsymbol{\theta} = \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ \theta_{31} & \theta_{32} & \theta_{33} \end{pmatrix}, \quad \text{where } \theta_{ij} \in \mathbb{R}. \tag{4.1}
$$

Note that $(\Delta_{ij})_{i,j=1}^3 \subset \mathbb{R}^{3 \times 3}$ such that the following forms a basis for $\mathbb{R}^{3 \times 3}$:

$$
\boldsymbol{\Delta}_{11} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \boldsymbol{\Delta}_{12} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix}, \qquad \boldsymbol{\Delta}_{13} = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 2 & 0 \\ 0 & -1 & 0 \end{pmatrix},
$$

$$
\boldsymbol{\Delta}_{21} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \boldsymbol{\Delta}_{22} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}, \qquad \boldsymbol{\Delta}_{23} = \begin{pmatrix} 1 & -2 & 1 \\ 0 & 0 & 0 \\ -1 & 2 & -1 \end{pmatrix}, \tag{4.2}
$$

$$
\boldsymbol{\Delta}_{31} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & -2 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \boldsymbol{\Delta}_{32} = \begin{pmatrix} 1 & 0 & -1 \\ -2 & 0 & 2 \\ 1 & 0 & -1 \end{pmatrix}, \qquad \boldsymbol{\Delta}_{33} = \begin{pmatrix} -1 & 2 & -1 \\ 2 & -4 & 2 \\ -1 & 2 & -1 \end{pmatrix}.
$$

Hence, we can, for a given $h > 0$, express $\boldsymbol{\theta}$ as

$$
\begin{aligned}
\boldsymbol{\theta} = {} & \beta_{11} \boldsymbol{\Delta}_{11} + \frac{\beta_{12}}{2h} \boldsymbol{\Delta}_{12} + \frac{\beta_{21}}{2h} \boldsymbol{\Delta}_{21} + \frac{\beta_{22}}{4h^2} \boldsymbol{\Delta}_{22} + \frac{\beta_{13}}{h^2} \boldsymbol{\Delta}_{13} \\
& + \frac{\beta_{31}}{h^2} \boldsymbol{\Delta}_{31} + \frac{\beta_{32}}{2h^3} \boldsymbol{\Delta}_{32} + \frac{\beta_{23}}{2h^3} \boldsymbol{\Delta}_{23} + \frac{\beta_{33}}{h^4} \boldsymbol{\Delta}_{33}.
\end{aligned} \tag{4.3}
$$

Note that the 3×3 matrices Δ_{ij} can be seen as the finite difference discretisations of the partial derivatives of f if h corresponds to the distance between the sampling density underlying the discretisation $\boldsymbol{f}$. For a smooth function f, we therefore observe that if the discretisation h goes to zero, then

$$
\begin{aligned}
\boldsymbol{\theta} * \boldsymbol{f}(i,j) \to \big(& \beta_{11} f + \beta_{12} \partial_2 f + \beta_{21} \partial_1 f + \beta_{22} \partial_1 \partial_2 f + \beta_{13} \partial_2^2 f \\
& + \beta_{31} \partial_1^2 f + \beta_{23} \partial_2^2 \partial_1 f + \beta_{32} \partial_1^2 \partial_2 f + \beta_{33} \partial_1^2 \partial_2^2 f \big) \left((x_i, y_j) \right),
\end{aligned}
$$

where $(x_i, y_j)_{i,j=1}^N$ are the discretisation points. For an open set $\Omega \subset \mathbb{R}^2$, this yields the following operator $\mathcal{K}_{\boldsymbol{\theta}}$ defined on $\mathcal{S}(\Omega)$:

$$
\begin{aligned}
\mathcal{K}_{\boldsymbol{\theta}}(f) := {} & \beta_{11} f + \beta_{12} \partial_2 f + \beta_{21} \partial_1 f + \beta_{22} \partial_1 \partial_2 f + \beta_{13} \partial_2^2 f \\
& + \beta_{31} \partial_1^2 f + \beta_{23} \partial_2^2 \partial_1 f + \beta_{32} \partial_1^2 \partial_2 f + \beta_{33} \partial_1^2 \partial_2^2 f.
\end{aligned} \tag{4.4}
$$

By duality, we can extend $\mathcal{K}_{\boldsymbol{\theta}}$ to $\mathcal{S}'(\Omega)$. Note also that $\mathcal{K}_{\boldsymbol{\theta}}$ is a linear second-order differential operator. In particular, it is a pseudodifferential operator with its symbol given by

$$
\begin{aligned}
p_{\boldsymbol{\theta}}(\xi) = {} & \beta_{11} + \beta_{12}\xi_2 + \beta_{21}\xi_1 + \beta_{22}\xi_1\xi_2 \\
& + \beta_{13}\xi_2^2 + \beta_{31}\xi_1^2 + \beta_{23}\xi_2^2\xi_1 + \beta_{32}\xi_1^2\xi_2 + \beta_{33}\xi_1^2\xi_2^2 \text{ for } \xi \in \Omega.
\end{aligned}
\tag{4.5}
$$

The interpretation above of a discrete convolutional operator that takes non-smooth images as inputs necessitates to have a continuum definition that acts on distributions. Consequently, also all further operations will need to be applicable to distributions.

4.2.2. Pointwise application of ReLU to distributions

The rectified linear unit (ReLU) is an activation function used in many neural network architectures. It is defined as the positive part of its argument, i.e., $\mathsf{ReLU}\colon \mathbb{R} \to \mathbb{R}$ is given as $\mathsf{ReLU}(x) := \max\{x, 0\}$. Our aim is to extend the ReLU to an operator that acts on tempered distributions on $\Omega \subset \mathbb{R}^n$, denoted as ReLU.

We start by rewriting the ReLU function in terms of the Heaviside function $\mathsf{H}\colon \mathbb{R} \to \mathbb{R}$:

$$
\mathsf{ReLU}(x) = \mathsf{H}(x)x, \quad \text{where} \quad \mathsf{H}(x) := \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{if } x \leq 0. \end{cases}
\tag{4.6}
$$

The above can be used to extend ReLU in a straightforward manner to $f \in \mathcal{C}^{\infty}(\Omega)$, by simply defining

$$
\mathrm{ReLU}(f)(x) := \mathsf{ReLU}\big(f(x)\big) = \mathsf{H}\big(f(x)\big)f(x) = \begin{cases} f(x), & \text{if } f(x) > 0, \\ 0, & \text{if } f(x) \leq 0. \end{cases}
\tag{4.7}
$$

We only know that $\mathrm{ReLU}\colon \mathcal{S}(\Omega) \to L^{\infty}(\Omega)$ and in fact $\mathrm{ReLU}(f)$ may not be smooth for $f \in \mathcal{S}(\Omega)$. Hence, ReLU does not map $\mathcal{S}(\Omega)$ to $\mathcal{S}(\Omega)$, i.e., we cannot use duality to define ReLU on distributions. Using the characterisation in (4.7) to extend ReLU to distributions involves extending the Heaviside function to tempered distributions and also ensuring the subsequent multiplication is well-defined.

We start by defining the Heaviside function of a distribution $f \in \mathcal{S}'(\Omega)$ as the characteristic function of its positive support, i.e., we define the Heaviside operator $\mathcal{H}\colon \mathcal{S}'(\Omega) \to L^{\infty}(\Omega)$ as

$$
\mathcal{H}(f) := \mathbb{1}_{\mathrm{supp}_+(f)}, \quad \text{for } f \in \mathcal{S}'(\Omega),
\tag{4.8}
$$

where $\mathrm{supp}_+(f) \subset \Omega$ is the positive support of f (Definition A.2) and $\mathbb{1}_{\mathrm{supp}_+(f)}$ denotes the characteristic function of $\mathrm{supp}_+(f)$. Before proceeding, we list desirable properties for an extension of ReLU to distributions. More precisely, $\mathrm{ReLU}\colon \mathcal{S}'(\Omega) \to \mathcal{S}'(\Omega)$ should preferably come with the following properties:

1. $\mathrm{esssup}\,\mathrm{ReLU}(f) \subset \mathrm{supp}_+(f)$ (the essential support esssup is defined in Definition A.2),
2. $\mathrm{ReLU}(f)(\phi) = f(\phi)$ for all test functions $\phi \in \mathcal{S}(\Omega)$ supported in $\mathrm{supp}_+(h)$,
3. $\mathrm{ReLU}(f) = \mathcal{H}(f)\,f$ whenever $f \in \mathcal{S}'(\Omega)$.

Having extended the Heaviside function to distributions as in (4.8), our main concern is to ensure that the multiplication between the distribution $f \in \mathcal{S}'(\Omega)$ and $\mathcal{H}(f) \in L^{\infty}(\Omega)$ is well-defined. By Definition A.3 and Theorem A.4, this is indeed the case whenever $(x, -\lambda) \notin \mathrm{WF}(f)$ for all $(x, \lambda) \in \mathrm{WF}\big(\mathcal{H}(f)\big)$, i.e., we can define $\mathrm{ReLU}(f)$ by (4.7) for any $f \in \mathcal{S}'(\Omega)$ that satisfies these criteria. However, the multiplication is not necessarily well-defined whenever there exists $(x, \lambda) \in \mathrm{WF}\big(\mathcal{H}(f)\big)$ such that $(x, -\lambda) \in \mathrm{WF}(f)$. Thus, all we know is that the multiplication of $\mathcal{H}(f)$ and f is always defined if $f \in L^2_{\mathrm{loc}}(\Omega)$ (Remark A.5).

One idea is therefore to locally dampen f close to points where we cannot define the multiplication of f with $\mathcal{H}(f)$. This leads to the following definition of ReLU on distributions.

Definition 4.1. Let $\Omega \subset \mathbb{R}^2$ be open, $\kappa > 0$, and $\phi_\kappa \in \mathcal{S}(\mathbb{R}^2)$ be a function that integrates to 1, is positive and is supported on a compact subset of $B_\kappa(0)$. Then define

$$\mathrm{ReLU}_{\kappa,\phi_\kappa}(f) := \mathcal{H}(f)f^s, \quad \text{for } f \in \mathcal{S}'(\Omega), \tag{4.9}$$

where $f^s := (1 - \theta_\kappa)f$. Here $\theta_\kappa := \mathbb{1}_X * \phi_\kappa$ with

$$X := \left\{ x \in \mathbb{R}^2 \setminus \mathrm{supp}_{L^2}(f) \colon (x,\lambda) \in \mathrm{WF}(H(h)), (x,-\lambda) \in \mathrm{WF}(h) \text{ for a } \lambda \in \mathbb{S}^1 \right\} + B_\kappa(0).$$

In the above, $\mathrm{supp}_{L^2}(f) \subset \Omega$ denotes the L^2-support of f (Definition A.6).

We next show that Definition 4.1 can be used to extend the ReLU function to distributions.

Proposition 4.2. *Consider* $\mathrm{ReLU}_{\kappa,\phi_\kappa}$ *in* (4.9) *for some* $\kappa > 0$ *and let* $\phi_\kappa \in \mathcal{S}(\mathbb{R}^2)$ *that integrates to 1, is positive, and is supported on a compact subset of* $B_\kappa(0)$. *Then* $\mathrm{ReLU}_{\kappa,\phi_\kappa} \colon \mathcal{S}'(\Omega) \to \mathcal{S}'(\Omega)$ *for an open domain* $\Omega \subset \mathbb{R}^2$.

Proof. We need to show that $\mathrm{ReLU}_{\kappa,\phi_\kappa}(f) \in \mathcal{S}'(\Omega)$, whenever $f \in \mathcal{S}'(\Omega)$. To see this, note first that $1 - \theta_\kappa$ is smooth and vanishes on a neighbourhood of every $x \in \Omega \setminus \mathrm{supp}_{L^2}(f)$, where

$$(x,\lambda) \in \mathrm{WF}\big(\mathcal{H}(f)\big) \quad \text{and} \quad (x,-\lambda) \in \mathrm{WF}(f) \quad \text{for some } \lambda \in \mathbb{S}^1.$$

Hence, the product $(1 - \theta_\kappa)f$ is well-defined and by Theorem A.4, there does not exist an $x \in \Omega \setminus \mathrm{supp}_{L^2}(f)$ such that

$$(x,\lambda) \in \mathrm{WF}\big(\mathcal{H}(f)\big) \quad \text{and} \quad (x,-\lambda) \in \mathrm{WF}((1-\theta_\kappa)f).$$

Theorem A.4 and Remark A.5 now imply that $\mathrm{ReLU}_{\kappa,\phi_\kappa}(f) \in \mathcal{S}'(\Omega)$ whenever $f \in \mathcal{S}'(\Omega)$, which concludes the proof. $\square$

Remark 4.3. The set X in Definition 4.1 is a neighbourhood of the set on which the definition of $\mathsf{ReLU}(h)$ via the multiplication $\mathcal{H}(h)h$ is not well defined. To understand the nature of this set, we consider three examples:

1. $f \in \mathcal{S}'(\Omega)$. Then $X = \varnothing$, and hence $\mathrm{ReLU}_{\kappa,\phi_\kappa}(f) = \mathsf{ReLU}(f)$. In particular, if $f = \mathcal{H}(h)$ for some $h \in \mathcal{S}'(\Omega)$, then $\mathrm{ReLU}_{\kappa,\phi_\kappa}(f) = \mathsf{ReLU}(f) = f$.
2. $f = P(\mathbb{1}_B)$ for some domain $B \subset \Omega$ and P is an elliptic linear differential operator of order at least one. Then $\mathrm{esssup}(f) \subset \partial B$, so $\mathcal{H}(f) = 0$ which in turn implies that $X = \varnothing$ and $\mathsf{ReLU}_{\kappa,\phi_\kappa}(f) = 0$.
3. $f = P(\mathbb{1}_B + h)$ for some domain $B \subset \Omega$ and P is an elliptic linear differential operator. Assume furthermore that $h \in \mathcal{C}^\infty(\Omega)$ is such that $P(h)$ is positive on B. Then $X = \partial B + B_\kappa(0)$, since f is not a function at ∂B and $\mathcal{H}(f) = \mathbb{1}_B$. Thus $\mathrm{WF}\big(\mathsf{H}(f)\big) = \mathrm{WF}(f)$.

We conclude by pointing out that $\mathrm{ReLU}_{\kappa,\phi_\kappa} \colon \mathcal{S}'(\Omega) \to \mathcal{S}'(\Omega)$ in Definition 4.1 fulfils almost all of the criteria stipulated earlier for an extension of ReLU to distributions. It satisfies the first and third criteria. Moreover, $\mathrm{ReLU}_{\kappa,\phi_\kappa}(f)(\phi) = f(\phi)$ holds for all $\phi \in \mathcal{S}(\Omega)$ with a support that has a distance of more than 2κ from $\mathrm{WF}\big(\mathcal{H}(f)\big) \subset \partial \mathrm{supp}_+(f)$.

4.2.3. The continuum Learned Primal-Dual network

To define the continuum Learned Primal-Dual network, we start by introducing a continuum ResNet.

Definition 4.4 *(Continuum two-dimensional convolutional ResNet).* Let $\Omega \subset \mathbb{R}^2$ be open and let $N \in \mathbb{N}$, $j \in \{1, 2, 3, 4\}$, where $n_4 = 1$ be the *numbers of channels per layer*. Further, for $j = 1, \ldots, 4$, let $\boldsymbol{\theta}_j := (\boldsymbol{\theta}_j^{l,k})_{l=1,k=1}^{n_{j-1},n_j} \subset (\mathbb{R}^{3\times3})^{n_{j-1}\times n_{j-1}}$ be a *set of coefficients*. Let $\kappa > 0$ and let $\phi_\kappa \in \mathcal{S}(\Omega)$ be a function that integrates to 1, is positive and is supported on a compact subset of $B_\kappa(0)$.

We define the *continuum convolutional affine operator* $W_{\boldsymbol{\theta}_j}^c : (\mathcal{S}'(\Omega))^{n_{j-1}} \to (\mathcal{S}'(\Omega))^{n_j}$ as

$$W_{\boldsymbol{\theta}_j}(f)_k = \sum_{l=1}^{n_{j-1}} \mathcal{K}_{\boldsymbol{\theta}_j^{l,k}}(f) \quad \text{for } k \in \{1, \ldots, n_j\} \text{ and } f \in (\mathcal{S}'(\Omega))^{n_{j-1}}. \tag{4.10}$$

The *continuum ResNet operator* $\mathrm{ResNet}_{\kappa,\phi_\kappa} : (\mathcal{S}'(\Omega))^{n_0} \to \mathcal{S}'(\Omega)$ is then given by

$$\mathrm{ResNet}_{\kappa,\phi_\kappa}(f_1, \ldots, f_{n_0}) = f_1 + \mathcal{G}(f_1, \ldots, f_{n_0}) \quad \text{for } f_1, \ldots, f_{n_0} \in \mathcal{S}'(\Omega),$$

where $\mathcal{G} : (\mathcal{S}'(\Omega))^{n_0} \to \mathcal{S}'(\Omega)$ is the operator

$$\mathcal{G}(f_1, \ldots, f_{n_0}) = \left(W_{\boldsymbol{\theta}_4} \circ \mathrm{ReLU}_{\kappa,\phi_\kappa} \circ \mathcal{W}_{\boldsymbol{\theta}_3} \circ \mathrm{ReLU}_{\kappa,\phi_\kappa} \circ \mathcal{W}_{\boldsymbol{\theta}_2} \circ \mathrm{ReLU}_{\kappa,\phi_\kappa} \circ \mathcal{W}_{\boldsymbol{\theta}_1}\right)(f_1, \ldots, f_{n_0})$$

for $f_1, \ldots, f_{n_0} \in \mathcal{S}'(\Omega)$.

Remark 4.5.

1. Slightly deviating from Definition 2.1, we do not include a bias term in the definition of the continuum ResNet above, since no such term will appear in our implementation in Subsection 6.
2. Note that besides the previously defined operators $\mathcal{K}_{\boldsymbol{\theta}_j^{l,k}}$ and $\mathrm{ReLU}_{\kappa,\phi_\kappa}$, only addition is applied in the continuum ResNet. Since the set of distributions is a linear space, we conclude that $\mathrm{ResNet}_{\kappa,\phi_\kappa}$ is a well-defined operator from $\mathcal{S}'(\Omega)$ to $\mathcal{S}'(\Omega)$.

Based on the definition above for the continuum ResNet, we can now define the continuum Learned Primal-Dual network as in (2.5) with operators in (2.6) given as continuum ResNets. Hence, a continuum Learned Primal-Dual network is a mapping

$$\mathcal{R}_\theta : \mathcal{S}'(\Xi) \to \mathcal{S}'(\Omega) \quad \text{where} \quad \mathcal{R}_\theta(g) := f_N$$

with $f_N \in \mathcal{S}'(\Omega)$ given by the N-step iterative scheme in Algorithm 1 in which Λ_i and Γ_i are continuum two-dimensional convolutional ResNets as in Definition 4.4.

Algorithm 1: Continuum Learned Primal-Dual network.

Input: $f_0 \in \mathcal{S}'(\mathbb{R}^2)$, $h_0 \in \mathcal{S}'(\Xi)$ and $g \in \mathcal{S}'(\Xi)$.
Output: Primal solution $f_N \in \mathcal{S}'(\mathbb{R}^2)$ and dual solution $h_N \in \mathcal{S}'(\Xi)$.
for $i = 1, \ldots, N-1$ **do**
  $h_{i+1} \longleftarrow \Gamma_i(h_i, \mathcal{A}(f_i), g)$;
  $f_{i+1} \longleftarrow \Lambda_i(f_i, [\partial\mathcal{A}(f_i)]^*(h_{i+1}))$;
end

4.3. Canonical relation for the continuum Learned Primal-Dual network

It is clear that we can describe the canonical relation for the continuum Learned Primal-Dual network, if we can identify such a relation for continuum ResNets Λ_i and Γ_i for $i = 1, \ldots, I$. In addition, we also need

to combine these canonical relations with the canonical relations for the Radon transform, and the relations for the Fréchet derivative of the adjoint of the Radon transform.

4.3.1. Differential operator

The canonical relation for a differential operator is typically very straightforward to compute, if this operator is an (elliptic) pseudodifferential operator. In this respect, we recall the following result.

Theorem 4.6 *(Pseudolocal property, [50, Theorem 14]). A pseudodifferential operator $\mathcal{P}$ satisfies the pseudolocal property, i.e.,*

$$\text{sing supp}(\mathcal{P}f) \subset \text{sing supp}(f) \;\; and \;\; \text{WF}(\mathcal{P}f) \subset \text{WF}(f) \quad for \; all \; f \in \mathcal{S}'(\mathbb{R}^2).$$

If $\mathcal{P}$ is elliptic, then we have equality instead of inclusion, i.e.,

$$\text{sing supp}(\mathcal{P}f) = \text{sing supp}(f) \;\; and \;\; \text{WF}(\mathcal{P}f) = \text{WF}(f) \quad for \; all \; f \in \mathcal{S}'(\mathbb{R}^2).$$

Since $\mathcal{K}_{\boldsymbol{\theta}}$ defined in (4.4) is a pseudodifferential operator, we obtain that

$$\text{WF}(\mathcal{K}_{\boldsymbol{\theta}}f) \subset \text{WF}(f). \tag{4.11}$$

This means that $\mathcal{K}_{\boldsymbol{\theta}}$ does not introduce new singularities to f, and might even delete some of them; in the case that the coefficients β_{ij} are such that

$$0 < |p_{\boldsymbol{\theta}}(\xi)| \text{ for all } \|\xi\| \neq 0,$$

the operator $\mathcal{K}_{\boldsymbol{\theta}}$ is elliptic and preserves the singularities, i.e., $\text{WF}(\mathcal{K}_{\boldsymbol{\theta}}f) = \text{WF}(f)$. Here $p_{\boldsymbol{\theta}}$ is the symbol defined in (4.5).

4.3.2. ReLU application

Since for $h \in \mathcal{S}'(\mathbb{R}^2)$ the distribution $\text{ReLU}_{\kappa,\phi_\kappa}(h)$ is defined in most parts of the domain as $H(h)h$, we can study its wavefront set using Theorem A.4. We now first study the wavefront set of $H(h)$. Afterward, we estimate the wavefront set of $\text{ReLU}_{\kappa,\phi_\kappa}(h)$ in Subsection 4.3.4.

4.3.3. The wavefront set of $\mathcal{H}(f)$

For a function $g \in L^2(\Omega)$, the wavefront set of $\mathcal{H}(g)$ is determined through the following factors: A point $x \in \Omega$, such that on a neighbourhood thereof g is almost always positive will be mapped to something constant by the Heaviside function. Since constant functions are smooth, this operation erases the wavefront set associated with a neighbourhood of x. The same argument can be made on neighbourhoods where g is almost everywhere negative. Points $x' \in \Omega$ in which g vanishes have the potential to create new discontinuities, since the Heaviside function has a jump at 0. If g is smooth in x' and also has non-vanishing gradient, then the implicit function theorem tells us the form of the discontinuity of $\mathcal{H}(g)$. We will see below in Proposition 4.8 that the same argument can be made for tempered distributions. A crucial ingredient for that result will be the following estimation of the wavefront set of indicator functions.

Proposition 4.7 *([11, Proposition 20]). Let $n \in \mathbb{N}$ and $B \subset \mathbb{R}^n$ be a domain with smooth boundary. Then,*
$$\text{WF}(\mathbb{1}_B) = \{(x, \lambda) \in \mathbb{R}^n \times \mathbb{S}^1 : x \in \partial B, \lambda \text{ normal to } \partial B \text{ in } x\}.$$

We can now state the following result describing the wavefront set of $\mathcal{H}(g)$.

Proposition 4.8. *Let $g \in \mathcal{S}'(\Omega)$ for $\Omega \subset \mathbb{R}^2$ an open domain. Let further*

$$R_g := \{x \in \mathbb{R}^2 : x \in \partial(\mathrm{supp}_+(g)), x \notin \mathrm{sing\,supp}(g), \nabla g(x) \neq 0\},$$
$$C_g := \{x \in \mathbb{R}^2 : x \in \partial(\mathrm{supp}_+(g)), x \notin \mathrm{sing\,supp}(g), \nabla g(x) = 0\},$$
$$S_g := \{x \in \mathbb{R}^2 : x \in \partial(\mathrm{supp}_+(g)), x \in \mathrm{sing\,supp}(g)\}.$$

Then, $(x, \lambda) \in \mathrm{WF}\big(\mathcal{H}(g)\big)$, if, for an $\alpha \neq 0$,

$$x \in R_g \text{ and } \lambda = \pm \nabla_x(g)/\|\nabla_x(g)\|. \tag{4.12}$$

Moreover, $(x, \lambda) \in \mathrm{WF}\big(\mathcal{H}(g)\big)$ only if (4.12) holds or $x \in C_g \cup S_g$.

Proof. We start with the "only if" part. The statement is clear if $\mathrm{WF}\big(\mathcal{H}(g)\big) = \varnothing$. Otherwise, let $(x, \lambda) \in \mathrm{WF}\big(\mathcal{H}(g)\big)$.

Assume first that $x \in \partial(\mathrm{supp}_+(g))^c$. Then either $x \in \mathrm{supp}_{-,0}(g)$ or $x \in \mathrm{supp}_+(f)^\circ$. Since both $\mathrm{supp}_{-,0}(g)$ and $\mathrm{supp}_+(g)^\circ$ are open sets, we have that there exists an open neighbourhood U of x such that $U \subset \mathrm{supp}_{-,0}(g)$ or $U \subset \mathrm{supp}_+(g)^\circ$. As a result, $\mathcal{H}(g)$ is constant on U. Therefore, (x, λ) cannot be in $\mathrm{WF}\big(\mathcal{H}(g)\big)$, which produces a contradiction.

Hence, we can assume that $(x, \lambda) \in \mathrm{WF}\big(\mathcal{H}(g)\big)$ and $x \in \partial(\mathrm{supp}_+(g))$. In addition, we assume that $x \notin C_g \cup S_g$. Then, $x \notin \mathrm{sing\,supp}(g)$. Therefore, there exists a neighbourhood U' of x, on which g is smooth and ∇g does not vanish.

We wish to show now that on U' the set $\{g = 0\}$ is a smooth curve with normal $\nabla_x g$ at x. For this, we invoke a *smooth version of the implicit function theorem* [54, Theorem 2.1]. In this form, the theorem considers a smooth function $\tilde{g} \colon \Omega \to \mathbb{R}$ such that

$$0 = \tilde{g}(x_1^*, x_2^*), \quad \text{for } (x_1^*, x_2^*) \in \Omega.$$

Assuming that $\frac{\partial \tilde{g}}{\partial x_2} \neq 0$, then there exists a smooth κ defined on a neighbourhood of x_1^* such that locally, i.e., for x_1 in an open neighbourhood of x_1^*,

$$\tilde{g}(x_1, \kappa(x_1)) = 0 \quad \text{and} \quad \kappa'(x_1) = \frac{\partial \tilde{g}}{\partial x_1}(x_1) \bigg/ \frac{\partial \tilde{g}}{\partial x_2}(x_1).$$

Moreover, in an open neighbourhood of x_1^*, x_2^* every (x_1, x_2) such that $\tilde{g}(x_1, x_2) = 0$ is of the form $(x_1, \kappa(x_1))$. Applying the implicit function theorem to g if $\frac{\partial g}{\partial x_2} \neq 0$ yields that $\eta_x = \nabla g(x)/\|\nabla g\|$ is a normal at the zero level set of g at x. By swapping variables, the same argument can be made if $\frac{\partial g}{\partial x_1} \neq 0$. We obtain that locally $\mathcal{H}(g) = \mathbb{1}_B$ with ∂B being a smooth curve that has normal η_x at x. By Proposition 4.7, this implies that $(x, \lambda) \in \mathrm{WF}(\mathbb{1}_B)$ if $\lambda = \pm \eta_x$ for an $\alpha \neq 0$. This concludes the proof of the "only if" part.

For the "if" part, we notice again that if $x \in R_g$, then $x \in \partial(\mathrm{supp}_+(g))$. Thus, $x \notin S_g$ which implies that $x \notin \mathrm{sing\,supp}(g)$. Therefore, and since $x \notin C_g$, the implicit function theorem is applicable. The same argument as before yields (4.12). $\square$

Remark 4.9. We may ask ourselves whether or not the statement in Proposition 4.8 is tight. To improve our intuition in this regard, we provide an example for each of the cases of Proposition 4.8 that may lead to a wavefront set.

1. *Creation of wavefront set according to* (4.12):

Let $g(x) = 1 - \|x\|^2$. The squared Euclidean norm is a smooth function and $\{g = 0\} = \{x\colon \|x\| = 1\}$ is the unit circle. Moreover, $\mathcal{H}(g) = \mathbb{1}_{B_1}$ is the indicator of the unit ball. It is not hard to see that the wavefront set of this function is $\{(x,x)\colon x \in \mathbb{S}^1\}$. Also

$$\nabla_x(1 - \|x\|^2) = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix} = 2x.$$

2. $x \in C_g$ and $x \in \operatorname{sing\,supp}\big(\mathcal{H}(g)\big)$:
   Let g_1 be a positive $\mathcal{C}^\infty$ function, supported on a set D_1 that contains $(0,0)$. Let g_2 be another such function, however, with $D_1 \cap D_2 = \{(0,0)\}$. If $D_1 \cup D_2$ is not an open neighbourhood of $(0,0)$, which is possible, then $H(g_1 + g_2)$ is discontinuous at $(0,0)$ implying that $(0,0)$ is a singular point of $H(g_1 + g_2)$. One concrete example, would be given by $D_1 = [-1,0]^2$, $D_2 = [0,1]^2$. In this case, $\partial(\operatorname{supp}_+(g_1 + g_2))$ is not given by a single curve in the neighbourhood of $(0,0)$. Note that, necessarily by the smoothness of g_1 and g_2 it holds that $\nabla_x(g_1 + g_2) = 0$ for $x = (0,0)$.

3. $x \in C_g$ and $x \notin \operatorname{sing\,supp}\big(\mathcal{H}(g)\big)$: Let g be a smooth compactly supported positive function. Then every $x \in \partial \operatorname{supp}(g)$ satisfies that $x \in C_g$ and $x \notin \operatorname{sing\,supp}\big(\mathcal{H}(g)\big) = \operatorname{sing\,supp}(g)$.

4. $x \in S_g$ and $x \in \operatorname{sing\,supp}\big(\mathcal{H}(g)\big)$:
   Let $g = \mathbb{1}_{\mathbb{R}^+ \times \mathbb{R}}$, then $x \in S_g$ and $g = \mathcal{H}(g)$.

5. $x \in S_g$ and $x \notin \operatorname{sing\,supp}\big(\mathcal{H}(g)\big)$:
   Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a $\mathcal{C}^\infty$ function with compact support on $\mathbb{R}^+$. The function $g(x) = \phi(x_1) + \mathbb{1}_{\mathbb{R}^-}(x_1)x_1^3$ is not smooth since it has a jump in its third derivative at the $x_1 = 0$ axis. At the same time $\{x_1 = 0\} = \partial(\operatorname{supp}_+(g))$. Finally, we observe that $\mathcal{H}(g)(x) = \phi(x)$ and hence the wavefront set of $\mathcal{H}(g)$ is empty.

Notice that Proposition 4.8 stays short of a precise characterisation of the wavefront set of $\mathcal{H}(f)$. It implies that all singularities must be in one of the sets R_g, C_g, or S_g. However, there is a closed-form of the orientations of the singularities only if $x \in R_g$.

4.3.4. Wavefront set of $\mathsf{ReLU}(f) = \mathcal{H}(f)f$

In this subsection, we chose a fixed $\kappa > 0$ and $\phi_\kappa \in \mathcal{S}(\mathbb{R}^2)$ that integrates to 1, is positive, and is supported on a compact subset of $B_\kappa(0)$. To reduce the computation of $\mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f))$ to that of $\mathrm{WF}(\mathcal{H}(f))$, we will make use of the following version of the product theorem:

Theorem 4.10 ([11, Theorem 13]). *Let u and v be distributions in $\mathcal{S}'(U)$ for an open domain U. Assume that for no point (x, λ) in $\mathrm{WF}(u)$ we have $(x, -\lambda) \in \mathrm{WF}(v)$. Then, $uv \in \mathcal{S}'(U)$ and*

$$\mathrm{WF}(uv) = S_+ \cup S_u \cup S_v,$$

where $S_+ := \{(x, \lambda + \mu)\colon (x, \lambda) \in \mathrm{WF}(u), (x, \mu) \in \mathrm{WF}(v)\}$, $S_u := \{(x, \lambda)\colon (x, \lambda) \in \mathrm{WF}(u), x \in \operatorname{ess\,supp}(v)\}$, and $S_v := \{(x, \lambda)\colon (x, \lambda) \in \mathrm{WF}(v), x \in \operatorname{ess\,supp}(u)\}$.

In particular, for $g \in \mathcal{S}'(\mathbb{R}^2)$ and $f \in \mathcal{C}^\infty(\mathbb{R}^2)$ where $\operatorname{supp}(f)$ is compact, we have that $\mathrm{WF}(fg) \subset \mathrm{WF}(g) \cap (\operatorname{supp}(f) \times \mathbb{R}^2)$.

Let $\Omega \subset \mathbb{R}^2$, $f \in \mathcal{S}'(\Omega)$, and let us assume that $\mathrm{WF}(f)$ is known. In addition, using the results of Subsection 4.3.3 we have also access to $\mathrm{WF}(\mathcal{H}(f))$. We denote as in Definition 4.1:

$$X_h := \{x \in \mathbb{R}^2 \setminus \operatorname{supp}_{L^2}(h)\colon (x, \lambda) \in \mathrm{WF}(H(h)), (x, -\lambda) \in \mathrm{WF}(h) \text{ for a } \lambda \in \mathbb{S}^1\} + B_\kappa(0) \qquad (4.13)$$

and

$$X_h^{3\kappa} := \{x \in \mathbb{R}^2 \setminus \operatorname{supp}_{L^2}(h) \colon (x, \lambda) \in \operatorname{WF}(H(h)), (x, -\lambda) \in \operatorname{WF}(h) \text{ for a } \lambda \in \mathbb{S}^1\} + B_{3\kappa}(0). \quad (4.14)$$

Note that per Definition 4.1 we have that $\theta_\kappa = 0$ on $(X_h^{3\kappa})^c$ and hence $h^s = h$ on $(X_h^{3\kappa})^c$.

Now we have collected all necessary ingredients to be able to compute the wavefront set of $\operatorname{ReLU}_{\kappa,\phi_\kappa}(f) = \mathcal{H}(f)f$, which is the goal of the following result.

Theorem 4.11. *Let $\Omega \subset \mathbb{R}^2$ be open and let $f \in \mathcal{S}'(\Omega)$. In addition, let*

$$\mathcal{A}_f := \operatorname{WF}(f) \cap (\operatorname{supp}_+(f)^\circ \times \mathbb{S}^1), \qquad (4.15)$$

$$\mathcal{R}_f := \{(x, \lambda) \in R_f \times \mathbb{S}^1 : (x, \lambda) \text{ follows } (4.12)\}, \qquad (4.16)$$

where R_f is defined as in Proposition 4.8. In addition, $\mathcal{CS}_f$ is given by

$$\mathcal{CS}_f := \{(x, \xi) \in (S_f \cup C_f) \times \mathbb{S}^1 : (x, \xi) \in \operatorname{WF}(\operatorname{ReLU}_{\kappa,\phi_\kappa}(f))\}, \qquad (4.17)$$

where C_f and S_f are defined as in Proposition 4.8. Then $\operatorname{WF}(\operatorname{ReLU}_{\kappa,\phi_\kappa}(f))$ is given by:

$$\operatorname{WF}(\operatorname{ReLU}_{\kappa,\phi_\kappa}(f)) \cap (X_h^{3\kappa} \times \mathbb{S}^1)^c = (\mathcal{A}_f \cup \mathcal{R}_f \cup \mathcal{CS}_f) \cap (X_h^{3\kappa} \times \mathbb{S}^1)^c, \qquad (4.18)$$

$$\operatorname{WF}(\operatorname{ReLU}_{\kappa,\phi_\kappa}(f)) \cap (X_h^{3\kappa} \times \mathbb{S}^1) \subset (\mathcal{A}_f \cup \mathcal{R}_f \cup \mathcal{CS}_f) \cap (X_h^{3\kappa} \times \mathbb{S}^1). \qquad (4.19)$$

In particular, we have that

$$\operatorname{WF}(\operatorname{ReLU}_{\kappa,\phi_\kappa}(f)) \subset \mathcal{A}_f \cup \mathcal{R}_f \cup \left((C_f \cup S_f) \times \mathbb{S}^1\right). \qquad (4.20)$$

Proof. Since $\mathbb{R}^2$ can be decomposed as

$$\mathbb{R}^2 = \operatorname{supp}_+(f)^\circ \cup \partial(\operatorname{supp}_+(f)) \cup \operatorname{supp}_{-,0}(f),$$

we have that $\operatorname{WF}(\operatorname{ReLU}_{\kappa,\phi_\kappa}(f))$ can be decomposed as

$$\operatorname{WF}(\operatorname{ReLU}_{\kappa,\phi_\kappa}(f)) = \mathrm{A}_{f,\kappa} \cup \mathrm{B}_{f,\kappa} \cup \mathrm{D}_{f,\kappa}, \qquad (4.21)$$

where

$$\mathrm{A}_{f,\kappa} := \operatorname{WF}(\operatorname{ReLU}_{\kappa,\phi_\kappa}(f)) \cap (\operatorname{supp}_+(f)^\circ \times \mathbb{S}^1),$$
$$\mathrm{B}_{f,\kappa} := \operatorname{WF}(\operatorname{ReLU}_{\kappa,\phi_\kappa}(f)) \cap (\operatorname{supp}_{-,0}(f) \times \mathbb{S}^1),$$
$$\mathrm{D}_{f,\kappa} := \operatorname{WF}(\operatorname{ReLU}_{\kappa,\phi_\kappa}(f)) \cap (\partial(\operatorname{supp}_+(f)) \times \mathbb{S}^1).$$

Notice in addition that $\mathrm{A}_{f,\kappa}$, $\mathrm{B}_{f,\kappa}$ and $\mathrm{D}_{f,\kappa}$ are disjoint. Now, since $\operatorname{supp}_+(f)^\circ$ is open, we find for every $x \in \operatorname{supp}_+(f)^\circ$ with $x \notin X_h^{3\kappa}$ an open neighbourhood U of x such that $\operatorname{ReLU}_{\kappa,\phi_\kappa}(f)|_U = \mathcal{H}(f)|_U f^s|_U = f|_U$, since $\mathcal{H}(f)(x) = 1$ for every $x \in U$ and $h^s = h$ on $(X_h^{3\kappa})^c$. Thus

$$\mathrm{A}_{f,\kappa} \cap (X_h^{3\kappa} \times \mathbb{S}^1)^c = \operatorname{WF}(f) \cap (\operatorname{supp}_+(f)^\circ \times \mathbb{S}^1) \cap (X_h^{3\kappa} \times \mathbb{S}^1)^c = \mathcal{A}_f \cap (X_h^{3\kappa} \times \mathbb{S}^1)^c, \qquad (4.22)$$

where $\mathcal{A}_f$ is as in the statement of the proposition. Moreover, for every $x \in \operatorname{supp}_+(f)^\circ$ with $x \in X_h^{3\kappa}$ there is an open neighbourhood U' of x such that $\operatorname{ReLU}_{\kappa,\phi_\kappa}(f)|_{U'} = \mathcal{H}(f)|_{U'} f^s|_{U'} = f^s|_{U'}$. Therefore,

$$\mathrm{A}_{f,\kappa} \cap (X_h^{3\kappa} \times \mathbb{S}^1) = \operatorname{WF}(f^s) \cap (\operatorname{supp}_+(f)^\circ \times \mathbb{S}^1) \cap (X_h^{3\kappa} \times \mathbb{S}^1)$$
$$\subset \operatorname{WF}(f) \cap (X_h^{3\kappa} \times \mathbb{S}^1) = \mathcal{A}_f \cap (X_h^{3\kappa} \times \mathbb{S}^1). \qquad (4.23)$$

On the other hand, since $\operatorname{ess\,supp}(\mathcal{H}(f)) = \operatorname{supp}_+(f)$, by Theorem 4.10, we can conclude that

$$\operatorname{supp}_{-,0}(f) \subset (\operatorname{sing\,supp}(\mathcal{H}(f)f^s))^c = (\operatorname{sing\,supp}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)))^c.$$

Then, we have

$$\mathrm{B}_{f,\kappa} = \mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)) \cap (\operatorname{supp}_{-,0}(f) \times \mathbb{S}^1) = \varnothing. \tag{4.24}$$

Let us now study the set $\mathrm{D}_{f,\kappa} := \mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)) \cap (\partial(\operatorname{supp}_+(f)) \times \mathbb{S}^1)$. Following the notation of Proposition 4.8, we can decompose the set $\partial(\operatorname{supp}_+(f))$ as

$$\partial(\operatorname{supp}_+(f)) = R_f \cup C_f \cup S_f. \tag{4.25}$$

Using this decomposition, we can write $\mathrm{D}_{f,\kappa}$ as

$$\mathrm{D}_{f,\kappa} = \mathrm{R}_{f,\kappa} \cup \mathrm{CS}_{f,\kappa},$$

where

$$\mathrm{R}_{f,\kappa} := \mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)) \cap (R_f \times \mathbb{S}^1),$$
$$\mathrm{CS}_{f,\kappa} := \mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)) \cap ((C_f \cup S_f) \times \mathbb{S}^1).$$

Next, we would like to show that

$$\mathrm{R}_{f,\kappa} \cap (X_h^{3\kappa} \times \mathbb{S}^1)^c = \{(x,\lambda) \in R_f \times \mathbb{S}^1 : (x,\lambda) \text{ follows } (4.12)\} \cap (X_h^{3\kappa} \times \mathbb{S}^1)^c = \mathcal{R}_f \cap (X_h^{3\kappa} \times \mathbb{S}^1)^c, \tag{4.26}$$

$$\mathrm{R}_{f,\kappa} \cap (X_h^{3\kappa} \times \mathbb{S}^1) \subset \{(x,\lambda) \in R_f \times \mathbb{S}^1 : (x,\lambda) \text{ follows } (4.12)\} \cap (X_h^{3\kappa} \times \mathbb{S}^1) = \mathcal{R}_f \cap (X_h^{3\kappa} \times \mathbb{S}^1). \tag{4.27}$$

Let us start with (4.26). Consider first $(x,\lambda) \in \mathcal{R}_f$, $x \notin X_h^{3\kappa}$. Then, $x \in R_f$ and thus $x \notin \operatorname{sing\,supp} f$. In particular, $x \notin \operatorname{sing\,supp} f^s$. Moreover, since $\nabla_x f \neq 0$, we conclude that $x \in \operatorname{ess\,supp}(f)$ and therefore $x \in \operatorname{ess\,supp}(f^s)$. Using Theorem 4.10, we conclude that $(x,\lambda) \in \mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)) = \mathrm{WF}(f^s \mathcal{H}(f))$ if and only if $(x,\lambda) \in \mathrm{WF}(\mathcal{H}(f))$. Since $x \notin C_f \cup S_f$, we conclude from Proposition 4.8 that (x,λ) satisfies (4.12). To show the converse embedding, assume that (x,λ) is such that $x \in R_f$ and (x,λ) satisfies (4.12). By Proposition 4.8, we have that $(x,\lambda) \in \mathrm{WF}(\mathcal{H}(f))$. Furthermore, $x \notin \operatorname{sing\,supp} f$ and $\nabla_x f \neq 0$, which implies that $x \in \operatorname{ess\,supp} f$. We conclude by Theorem 4.10 that $(x,\lambda) \in \mathrm{WF}(f^s \mathcal{H}(f)) = \mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f))$. This shows (4.26).

To show (4.27) it suffices to observe that $x \in R_f$ implies that $x \notin \operatorname{sing\,supp} f$ and hence $x \notin \operatorname{sing\,supp}(1 - \theta_\kappa)f$. Therefore, we conclude that

$$\mathrm{R}_{f,\kappa} \cap (X_h^{3\kappa} \times \mathbb{S}^1) \subset \mathrm{WF}(\mathcal{H}(f)) \cap (R_f \times \mathbb{S}^1) \cap (X_h^{3\kappa} \times \mathbb{S}^1) = \mathcal{R}_f \cap (X_h^{3\kappa} \times \mathbb{S}^1),$$

where the last equality follows from Proposition 4.8. This yields (4.27).

The full result now follows by considering the decomposition (4.21). The part associated with $\mathrm{A}_{f,\kappa}$ is estimated via (4.22) and (4.23). The part associated with $\mathrm{B}_{f,\kappa}$ vanishes due to (4.24). Finally, the part associated with $\mathrm{D}_{\kappa,\phi_\kappa}$ is estimated via the decomposition (4.25), where the $\mathrm{R}_{f,\kappa}$ part is estimated via (4.26) and (4.27) and $\mathcal{CS}_f = \mathrm{CS}_{f,\kappa}$ holds per definition. $\quad\square$

Remark 4.12. Theorem 4.11 only yields an estimate for the wavefront set associated with the set $(X_h^\kappa)^c$. In the sequel, since we are using the continuum relations to infer certain properties of digital relations, we will assume that κ is chosen very small and X_h^κ is not seen by the discretisation. In other words, in practice, we compute the wavefront set only via (4.18).

Even on $(X_h^\kappa)^c$, Theorem 4.11 does not entirely save us from computing $\mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f))$, but restricts the necessity for such a computation to the cases where $x \in S_f \cup C_f$. The set $\mathcal{CS}_f$ can also be further split up: For $x \in \mathbb{R}^2 \setminus X_h^\kappa$, we denote by $\mathrm{WF}(f)_x$ the x-slice of the wavefront set of f defined as $\Lambda \subset \mathbb{R}^2$ such that $(x, \lambda) \in \{x\} \times \Lambda$ for all $(x, \lambda) \in \mathrm{WF}(f)$.

Proposition 4.13. *Let $\Omega \in \mathbb{R}^2$ be open, $f \in \mathcal{S}'(\Omega)$ be a distribution and $\mathcal{CS}_f$ be as in Theorem 4.11. Then*

$$\mathcal{CS}_f \bigcap (X_h^\kappa \times \mathbb{S}^1)^c = \Big\{ (x, \lambda) : x \in C_g \setminus X_h^\kappa \text{ and } (x, \lambda) \in \mathrm{WF}(\mathcal{H}(f)) \Big\}$$

$$\bigcup \Big\{ (x, \lambda) : x \in S_g \setminus X_h^\kappa, \mathrm{WF}(f)_x \cap -\mathrm{WF}(\mathcal{H}(f))_x = \varnothing, \lambda \in (\mathrm{WF}(f)_x + \mathrm{WF}(\mathcal{H}(f))_x) \setminus \{0\} \Big\} \quad (4.28)$$

$$\bigcup \Big\{ (x, \lambda) : x \in S_g \setminus X_h^\kappa, \mathrm{WF}(f)_x \cap -\mathrm{WF}(\mathcal{H}(f))_x \neq \varnothing, (x, \lambda) \in \mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)) \Big\}.$$

Proof. The result follows immediately from Theorem 4.10. □

Theorem 4.11 yields two ways to estimate the wavefront set of $\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)$ on $(X_h^\kappa)^c$. First, it can be precisely computed by (4.18). This, however, may require us to compute $\mathcal{CS}_f$ via Proposition 4.13. This computation could be performed according to (4.28), by using a method such as DeNSE to find $\mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f))$ if required.

Second, the wavefront set on $(X_h^\kappa)^c$ can be estimated using (4.20). Since we expect that it is not problematic to overestimate the wavefront set slightly, we opt for the second option and cast this wavefront set extraction algorithm as Algorithm 2.

Algorithm 2: Wavefront set classifier of $\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)$.

Input: Distribution $f \in \mathcal{S}'(\Omega)$, $\mathrm{WF}(f)$, $x \in \Omega$.
Output: Estimate $\mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f))_x \subset \Omega$.
initialisation;
if $x \in \mathrm{supp}_+(f)^\circ$ **then**
  | return $\Lambda_x = \mathrm{WF}(f)_x$;
end
if $x \in R_f$ **then**
  | return $\Lambda_x = \{\pm \nabla_x(f)/\|\nabla_x(f)\|\}$;
end
if $x \in C_f \cup S_f$ **then**
  | return $\Lambda_x = \Omega$;
end

4.3.5. Microlocal analysis of the residual layer and sum-taking

In the continuous setting, residual neural networks are operators $\mathcal{P} \colon \mathcal{S}'(\Omega) \to \mathcal{S}'(\Omega)$ of the form

$$\mathcal{P}(f) = f + \mathcal{C}(f),$$

where $\mathcal{C} \colon \mathcal{S}'(\Omega) \to \mathcal{S}'(\Omega)$ is continuum two-dimensional convolutional ResNet according to Definition 4.4. In addition, in Subsection 4.2.3, we allow summing over the channels of a convolutional block.

Because of this, we would like to identify $\mathrm{WF}(f+g)$ for two distributions of which we know the wavefront set. The following two results yield a complete description thereof.

Theorem 4.14 *([70, Page 93]). Let $\Omega \subset \mathbb{R}^2$ be open, let $f, g \in \mathcal{S}'(\Omega)$ and $(x; \lambda)$ be a regular directed point of f and g, then $(x; \lambda)$ is a regular directed point of $f + g$.*

In particular, if $(x, \lambda) \in \mathrm{WF}(f)$ and $(x; \lambda)$ is a regular directed point of g, then $(x, \lambda) \in \mathrm{WF}(f + g)$.

Corollary 4.15. *Let $\Omega \subset \mathbb{R}^2$ be open and let $f, g \in \mathcal{S}'(\Omega)$, then $\mathrm{WF}(f + g)$ is given by*

$$\mathrm{WF}(f + g) = \mathcal{A}_f \cup \mathcal{A}_g \cup \mathcal{A}_{f+g}, \tag{4.29}$$

where

$$\mathcal{A}_f := \{(x; \lambda) \in \mathrm{WF}(f) : x \notin \mathrm{WF}(g)\}, \qquad \mathcal{A}_{f+g} := ((\mathrm{WF}(f) \cap \mathrm{WF}(g))) \cap \mathrm{WF}(f + g),$$
$$\mathcal{A}_g := \{(x; \lambda) \in \mathrm{WF}(g) : x \notin \mathrm{WF}(f)\}$$

In particular, we have that

$$\mathcal{A}_{f+g} \subset \mathrm{WF}(f) \cap \mathrm{WF}(g). \tag{4.30}$$

Proof. The result follows immediately from Theorem 4.14. □

In a similar fashion as we did for the wavefront set of $\mathsf{ReLU}(f)$, using Corollary 4.15 we can find two ways to extract the wavefront set of $f + g$ via Corollary 4.15. We cast the one that yields a superset of the wavefront set of $f + g$ via (4.30) as Algorithm 3.

Algorithm 3: Wavefront set classifier of $f + g$.

Input: Distribution $f, g \in \mathcal{S}'(\Omega)$, $\mathrm{WF}(f), \mathrm{WF}(g)$, $x \in \Omega$.
Output: Estimate $\mathrm{WF}(f + g)_x \subset \Omega$.
initialisation;
if $x \in \mathrm{WF}(f) \cap \mathrm{WF}(g)^c$ **then**
  | return $\Lambda_x = \mathrm{WF}(f)_x$;
end
if $x \in \mathrm{WF}(f)^c \cap \mathrm{WF}(g)$ **then**
  | return $\Lambda_x = \mathrm{WF}(g)_x$;
end
if $x \in \mathrm{WF}(f) \cap \mathrm{WF}(g)$ **then**
  | return $\Lambda_x = \mathrm{WF}(g)_x \cup \mathrm{WF}(f)_x$;
end

4.3.6. Microlocal analysis of continuous Learned Primal-Dual networks

Let us first notice that the continuum residual neural network operator introduced in Definition 4.4 has four basic components, the differential layers, summation over channels, application of the ReLU, and the residual connection. We have seen in Subsection 4.3.5 that the effect on the wavefront set through summation over channels and the residual connection can be described as the output of Algorithm 3. In addition, the effect of the differential layers is described by (4.11) and the wavefront set after an application of the ReLU can be found through an application of Algorithm 2. Overall, there is an algorithm that produces for every continuum convolutional ResNet and every input function of which the wavefront set is known, an estimate of the wavefront set of the output. The microlocal behaviour of the full continuum Learned Primal-Dual network (Algorithm 1) can now be found by iteratively applying the canonical relation for the continuum residual neural network operator as well as the canonical relation for the Radon transform (3.11) and the adjoint of its Fréchet derivative (back-projection) in (3.17).
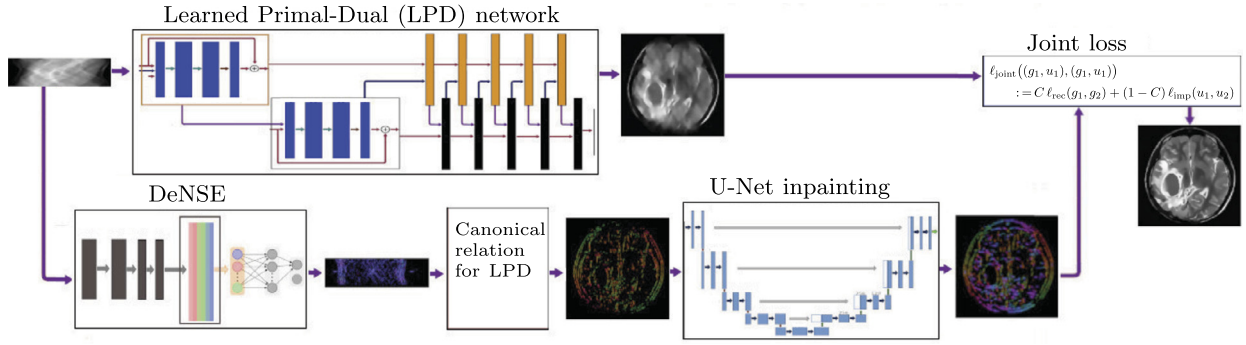
Learned Primal-Dual (LPD) network

Joint loss

$\ell_{\text{joint}}\big((g_1, u_1), (g_1, u_1)\big)$
$:= C\, \ell_{\text{rec}}(g_1, g_2) + (1 - C)\, \ell_{\text{imp}}(u_1, u_2)$

DeNSE

Canonical relation for LPD

U-Net inpainting

Fig. 4. Graphical representation of the workflow for the joint reconstruction and wavefront set inpainting algorithm. The input is partial sinogram data. In the top row first a Learned Primal-Dual architecture is applied. In the bottom row we first apply DeNSE to extract the wavefront set, then we apply the canonical relation for the Learned Primal-Dual network of Subsection 4.3.6. To the output thereof we apply the U-Net for inpainting. This together with the output from the Learned Primal-Dual is then input into the joint loss function.

5. The joint reconstruction and wavefront set extraction algorithm

In this section, we present our numerical algorithm for the digital reconstruction problem associated with the inverse tomography problem. We first introduce the algorithm in the framework of statistical learning theory as an empirical risk minimisation problem over a special set of deep neural networks with a specific loss term. Then, we evaluate our approach on a test set and compare its results with various benchmarks.

5.1. Outline of our algorithm

The setting for our data driven approach is to have a supervised tomographic training dataset of the form $(\boldsymbol{f}_i, \boldsymbol{g}_i)_{i=1}^N \in \mathbb{R}^n \times \mathbb{R}^m$. Here, $\boldsymbol{f}_i \in \mathbb{R}^n$ is an array that is a discretisation of a function $f \in X = L^2(\mathbb{R}^2)$ that represents the true image (signal). Likewise, $\boldsymbol{g}_i \in \mathbb{R}^m$ is an array corresponding to a discretisation of continuum data, which here is a function $g_i \in Y = L^2(\Xi)$ representing a noisy limited-angle sinogram with Ξ denoting the corresponding manifold of lines. In addition, we also know the complete wavefront set for data g_i and signal $\boldsymbol{f}_i$. Subsection 5.2 explores this statistical framework approach in more detail.

From the above, in absence of observation noise we have $g_i = \mathcal{A}(f_i)$ where $\mathcal{A}: X \to Y$ is the Radon transform (forward operator) restricted to Ξ (limited-angle data). The elements $(f_i, g_i) \in X \times Y$ are assumed to be independent samples generated by some $(X \times Y)$-valued random variable $(\mathbb{f}, \mathbb{g})$. The reconstruction operator $\mathcal{R}_\theta: \mathbb{R}^m \to \mathbb{R}^n$ is now a DNN given by the Learned Primal-Dual network [2] and its parameter $\theta = \hat{\theta}$ is set by training this DNN to achieve two tasks simultaneously:

1. Applying $\mathcal{R}_{\hat{\theta}}: \mathbb{R}^m \to \mathbb{R}^n$ (trained reconstruction operator) to an input sinogram $\boldsymbol{g} \in \mathbb{R}^m$ that is a discretisation of $g \in Y$ drawn from $\mathbb{g}$, should produce $\boldsymbol{f} \in \mathbb{R}^n$ that is a discretisation of a function $f \in X$ such that $\mathcal{A}(f) = g$.
2. The trained reconstruction operator $\mathcal{R}_{\hat{\theta}}: \mathbb{R}^m \to \mathbb{R}^n$ should output a discretisation of the wavefront set of f, which we denote by $\mathsf{DWF}(f)$, i.e., $\mathsf{DWF}\big(\mathcal{R}_{\hat{\theta}}(\boldsymbol{g})\big) = \mathsf{DWF}(f)$.

While we are, in the end, only interested in an accurate reconstruction of f (see Subsection 5.3), we believe that requiring a neural network to perform both tasks at the same time constitutes a strong prior. Indeed, we will see below that this joint approach yields vastly superior reconstruction results to a neural network that only reconstructs f. As a key ingredient towards the solution of the aforementioned task, we can rely on a digital wavefront set extraction operator DeNSE [4] which allows us to extract the digital wavefront set of a digital image, this task is briefly introduced in Subsection 5.4. Furthermore, it is reasonable to

construct an appropriate loss function $\mathcal{L}$ that measures the discrepancy between $\Phi(\boldsymbol{g})$ and $\boldsymbol{f}$ and between $\mathsf{DWF}(\Phi(\boldsymbol{g}))$ and $\mathsf{DWF}(f)$.

However, it turns out that for the training of the neural network it is computationally prohibitive to compute $\mathsf{DWF}(\Phi(\boldsymbol{g}))$ in every training iteration with the shearlet-based wavefront set extractor DeNSE [4]. Because of this, we introduced a heuristic based on the mapping properties of a continuum operator, the continuum Learned Primal-Dual network, to estimate $\mathsf{DWF}(\Phi(\boldsymbol{g}))$ in Subsection 4.3.

Through the canonical relation for the continuous Learned Primal-Dual operator, we obtain an estimate of a discretisation of the visible wavefront set $\mathrm{WF}^{\mathrm{vis}}(\Phi(g))$ based on the weights of the neural network Φ. To relate this estimate to $\mathsf{DWF}(f)$, we apply wavefront set inpainting with a neural network Ψ of U-Net architecture [72], this is explained in detail in Subsection 5.5. Overall, this leads to an operator $\mathcal{Q}$ that takes as an input $\boldsymbol{g}$ and the neural networks Φ and Ψ and outputs an estimate of $\boldsymbol{f}$ and $\mathsf{DWF}(f)$. We therefore use the loss function

$$\mathcal{L}(\Phi, \Psi, \boldsymbol{g}, \boldsymbol{f}, \mathsf{DWF}(f)) := C\,\ell_1(\boldsymbol{f}, \Phi(\boldsymbol{g})) + (1 - C)\,\ell_2(\mathsf{DWF}(f), \Psi(\mathcal{Q}(\boldsymbol{g}, \Phi))), \tag{5.1}$$

where $C \in (0, 1]$ and ℓ_1, ℓ_2 are appropriate distance measures on discretised images and discretised wavefront sets that will be discussed in Subsection 3.4 below. This strategy that jointly trains a reconstruction and a task (wavefront set inpainting) falls in the framework of *task-adapted reconstruction* was introduced in [1]. Based on the loss function of (5.1), we now train the neural networks Φ and Ψ to minimise the objective

$$\frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\Phi, \Psi, \boldsymbol{g}_i, \boldsymbol{f}_i, \mathsf{DWF}(f_i)). \tag{5.2}$$

To find a minimiser, we use stochastic gradient descent-based optimisation over the neural network's weights as is standard in deep learning. Finally, Subsection 5.6 explores in detail the joint reconstruction in the digital setting.

5.2. Statistical learning framework

We would like to frame the problem of reconstruction and joint wavefront set extraction as a statistical learning problem [77,20]. For this, we first introduce some relevant notions:

Definition 5.1 *(Setting of statistical learning theory).* Let X, Y be two sets. We call X the *sample space*, and Y the *label space*. Further, let $\mathcal{D}$ be a distribution on $X \times Y$, and let $\ell \colon Y \times Y \to \mathbb{R}$ be a measurable function. We refer to ℓ as *loss function*. Then the *risk* R of a measurable function $h : X \to Y$ is defined as

$$\mathrm{R}(h) := \mathbb{E}_{(x,y) \sim \mathcal{D}}(\ell(h(x), y)).$$

For $m \in \mathbb{N}$, a set of samples $(x_i, y_i)_{i=1}^{m} \sim \mathcal{D}^m$ and for a hypothesis class $\mathcal{H}$, we define the *empirical risk minimiser* h^* as

$$h^* := \arg\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i), y_i).$$

Under certain assumptions on the complexity of the hypothesis set $\mathcal{H}$, it can be shown that for sufficiently large m, the function h^* will also achieve a small risk [77,20]. This means that the empirical risk minimiser h^* resolves the unknown relation between input and output described by $\mathcal{D}$. Because of this, we treat the empirical risk minimisation problem in this work as the central problem of interest.

Next we would like to phrase the problems of digital tomographic reconstruction, of digital visible wavefront set extraction, of full digital wavefront set extraction via inpainting, and of the joint reconstruction and digital wavefront set extraction and inpainting in the framework of statistical learning theory.

5.3. Digital tomographic reconstruction

A joint probability distribution on function spaces as described through the statistical formulation of the tomographic reconstruction problem in Definition 3.3 implies, via the discretisation procedure of Subsection 3.4, a joint probability distribution $\mathcal{D}_{\mathrm{rec}}$ on discretised image-sinogram data. In the discrete problem, the sample space is $\mathbb{R}^{I_s} =: X$ for $I_s \subset [m_1] \times [m_2]$, where $[M] := \{1, \ldots, M\}$, with the label space being defined by $\mathbb{R}^{n_1 \times n_2} =: Y$. As loss function we choose ℓ_{rec} to be the squared Euclidean distance between two elements in Y, i.e.,

$$\ell_{\mathrm{rec}}(y_1, y_2) := \sum_{i,j=1}^{n_1, n_2} |(y_1)_{i,j} - (y_2)_{i,j}|^2. \tag{5.3}$$

Moreover, we choose as a hypothesis set $\mathcal{H}_{\mathrm{rec}} := (h_\vartheta)_{\vartheta \in \Theta_{\mathrm{res}}}$ the set of functions which perform the mapping $f_0 \mapsto f_I$ according to Algorithm 1 parametrised via the weights of the discrete two-dimensional convolutional ResNets introduced in Definition 2.1.

5.4. Digital visible wavefront set extraction

Similarly to the previous Subsection 5.3, the discretisation procedure of Subsection 3.4 implies a joint distribution between the digital wavefront set of a measured sinogram and the digital visible wavefront set of the image, precisely defined below. Here we use the shorthand notation $A^B := \{h \colon B \to A\}$ for two sets A, B. If B is a finite set then this can be interpreted as a set of A-valued vectors indexed by B, i.e., $A^B = \{(a_b)_{b \in B} \colon a_b \in A \text{ for all } b \in B\}$.

Definition 5.2. Let $I_s \subset [m_1] \times [m_2]$ be a digital grid in the sinogram domain and $I_d \subset [180]$. The *digital visible wavefront set of the sinogram data* is given by an element of the set $X_{\mathrm{vis}} := \mathbb{R}^{I_s} \times \{0, 1\}^{I_d}$, where I_d represents the set of *visible angles*. Moreover, let $J_s \subset [n_1] \times [n_2]$ be a digital grid in the image domain and $J_d \subset [180]$, determined through the canonical relation for the Radon transform. The *digital visible wavefront set of the image data* is then described by an element of the set $Y_{\mathrm{vis}} := \mathbb{R}^{J_s} \times \{0, 1\}^{J_d}$.

The underlying loss function over Y_{vis} is the *multi-class cross-entropy loss given by*:

$$\ell_{\mathrm{vis}}(y, y') := - \sum_{(i_1, i_2) \in J_s} \sum_{i_3 \in J_d} y_{i_1, i_2, i_3} \log\big(y'_{i_1, i_2, i_3}\big) \quad \text{for } y, y' \in Y_{\mathrm{vis}}.$$

As a hypothesis set we choose $\mathcal{H}_{\mathrm{vis}} := (\hat{h}_\vartheta)_{\vartheta \in \Theta_{\mathrm{res}}}$ as the set of maps resulting from the digital canonical relation described in Subsection 4.3.6. These maps are parametrised via the weights of the discrete two-dimensional convolutional ResNets introduced in Definition 2.1. The important point to notice here is that $\mathcal{H}_{\mathrm{vis}}$ and $\mathcal{H}_{\mathrm{rec}}$ are parametrised by the same parameter set.

5.5. Digital wavefront set extraction and inpainting

To reconstruct the full digital wavefront set of an image from its sinogram data, we combine the visible wavefront set extraction of the previous subsection with a wavefront set inpainting step which will be

performed using a U-Net. We have chosen this architecture since it has shown state-of-the-art performance in image inpainting and segmentation, [72]. In this case, the sample set is $X_{\text{inp}} := \mathbb{R}^{I_s} \times \{0,1\}^{I_d}$, where $I_s \subset [m_1] \times [m_2]$ and $I_d \subset [180]$, and label space has consequently to be chosen as $Y_{\text{inp}} := \mathbb{R}^{n_1 \times n_2} \times \{0,1\}^{180}$. The loss function is then

$$\ell_{\text{inp}}(y, y') := - \sum_{i_1, i_2 = 1}^{n_1, n_2} \sum_{i_3 = 1}^{180} y_{i_1, i_2, i_3} \log\big(y'_{i_1, i_2, i_3}\big) \quad \text{for } y, y' \in Y_{\text{inp}}.$$

As a hypothesis set $\mathcal{H}_{\text{inp}}$, we use the set of maps resulting from the digital canonical relation described in Subsection 4.3.6 composed with a U-Net. These maps are parametrised via the weights of the discrete two-dimensional convolutional ResNets introduced in Definition 2.1 as well as by the weights of the U-Net. Because of this, we can write $\mathcal{H}_{\text{inp}} := (\boldsymbol{h}_{\vartheta, \theta})_{\vartheta \in \Theta_{\text{res}}, \theta \in \Theta_U}$.

5.6. Joint reconstruction and digital wavefront set extraction and inpainting

We will now describe a joint approach in which the reconstruction problem as well as the wavefront set reconstruction and inpainting problem are solved simultaneously. This task certainly requires that the hypotheses classes underlying the risk minimisation problem are, at least partially, parametrised by the same parameter set.

This joint approach leads to the sample space being chosen as $X_{\text{joint}} := \mathbb{R}^{I_s} \times \mathbb{R}^{I_s} \times \{0,1\}^{I_d}$ for $I_s \subset [m_1] \times [m_2]$ and $I_d \subset [180]$, and the label space being defined as $Y_{\text{inp}} := \mathbb{R}^{n_1 \times n_2} \times (\mathbb{R}^{n_1 \times n_2} \times \{0,1\}^{180})$. The loss function needs then to be chosen ℓ_{joint} as a weighted average of ℓ_{rec} and ℓ_{inp} in the following sense: For a $\lambda \in (0,1]$ we set, for $g_1, g_2 \in \mathbb{R}^{n_1 \times n_2}, (h_1, h_2) \in \mathbb{R}^{n_1 \times n_2} \times \{0,1\}^{180}$

$$\ell_{\text{joint}}((g_1, h_1), (g_2, h_2)) := \lambda \ell_{\text{rec}}(g_1, g_2) + (1 - \lambda)\ell_{\text{inp}}(h_1, h_2). \tag{5.4}$$

The task-adapted reconstruction architecture involving the joint loss function (5.4) is depicted in Fig. 4.

Finally, the hypothesis set for the joint approach is given by $\mathcal{H}_{\text{joint}} := ((h_\vartheta, \boldsymbol{h}_{\vartheta, \theta}))_{\vartheta \in \Theta_{\text{res}}, \theta \in \Theta_U}$.

6. Numerical results

We now provide numerical results for the joint reconstruction and wavefront set extraction algorithm of Section 5. We first present the set-up, followed by a description of our results, and finally an interpretation of the outcome.

6.1. Set-up

The set-up follows our viewpoint taken in Subsection 5.6, namely regarding the joint reconstruction problem from an empirical risk minimisation standpoint. Therefore, as a training set we use an artificial data set comprised of piecewise smooth functions, where the singularity curves are given by random splines of degree at most four and the smooth regions are polynomials of degree at most two. We call this data set the *random cartoon-functions data set*. Some examples of functions in this data set were shown in Fig. 2. Notice that for the images in this data set the digital wavefront set is known. We then study two types of partial data, namely limited-angle data, where we assume that a wedge of $40°$ is missing (a total angle interval of $80°$), and sparse-view, where we only measure 40 angles. This is followed by evaluating the trained task-adapted architecture (see Fig. 4) on the OASIS dataset [53] formed by real brain scans (see https://www.oasis-brains.org/).

The results are then compared with classical approaches for the limited-angle and sparse-view tomography, including filtered back-projection [50], Tikhonov [82], and total variation [88]. In addition, we have also

Table 1

Realistic dataset performance for general benchmarks for sparse-view CT on the left and limited-angle CT on the right.

Sparse-view CT:			Limited-angle CT:		
Method	SSIM	PSNR	Method	SSIM	PSNR
FBP	0.51	19.90	FBP	0.44	14.53
Tikhonov	0.73	24.77	Tikhonov	0.73	22.62
TV	0.88	26.59	TV	0.83	23.09
Shearlet-L2 sparse	0.73	24.69	Shearlet-L2 sparse	0.70	22.20
Shearlet-L1 sparse	0.78	25.42	Shearlet-L1 sparse	0.76	22.29
Learned Primal-Dual	0.89	27.55	PhantomNet	0.87	25.50
Joint approach	0.92	28.46	Learned Primal-Dual	0.86	25.55
			Joint approach	0.95	29.80

performed variational regularisation using as regulariser the L2 and L1-norm of the shearlet coefficients of the reconstruction [37,52]. We named these methods *Shearlet-L2 sparse* and *Shearlet-L1 sparse*. The deep-learning based benchmarks, namely the Learned Primal-Dual [2] and the Phantom-Net architectures [13], were similarly trained using the random cartoon functions data set. The code to reproduce our experiments can be found in http://shearlab.math.lmu.de/applications.

6.2. Results

We now present some exemplary results for our algorithm in the case of limited-angle tomography (80° wedge) and sparse-view (40 measured angles). The visual results are depicted in Figs. 5 and 6. The comparison with the benchmark approaches from Subsection 6.1 is shown in Table 1.

In addition, we show in Figs. 7 and 8 reconstruction results of the limited-angle and sparse-view tomography for the benchmarks already mentioned in Subsection 6.1. In all cases, the joint approach outperforms all classical approaches. Table 1 presents the performance measure of the benchmarks in terms of the average self similarity (SSIM) and peak signal-to-noise ratio (PSNR).

6.3. Interpretation

In Table 1, we see that the joint approach significantly outperforms all competing approaches. The same observation can be made by observing Figs. 8 and 7 that demonstrate a strong improvement in the reconstruction accuracy in comparison with the other methods and in particular to the Learned Primal-Dual architecture. In this context, it is noteworthy that the performance gap in this application is much more pronounced than in the sparse-view application. Since the Learned Primal-Dual corresponds to one half of the workflow of the joint approach, we observe that the additional wavefront set information is especially helpful in the limited-angle case. Since the gaps in the visible wavefront set are significantly smaller in the sparse-view set-up, we expect that, in this set-up, considerably more information on the wavefront set is already included in the observed data, whereas in the limited-angle set-up more prior knowledge needs to be invoked. It appears as if this necessary prior knowledge was very successfully extracted and incorporated in the training phase of the joint approach.

In Figs. 5 and 6 we performed a type of ablation study by looking at the prowess of the inpainting algorithm alone. It transpires from those figures that the inpainting is less successful if it is not coupled with the reconstruction workflow. This clearly shows that the strength of the joint approach does not solely lie in wavefront set identification and inpainting but in the interplay of inpainting and reconstruction jointly.

We wish to remark that *our algorithm is not trained on real-world images* but instead on artificial phantoms from the random cartoon-functions data set. This shows that the joint prior that incorporates physical information is not based on memorisation or overfitting. From the standpoint of applications, this could be a very desirable feature, as it prevents overfitting on specific biases in real-world data sets.
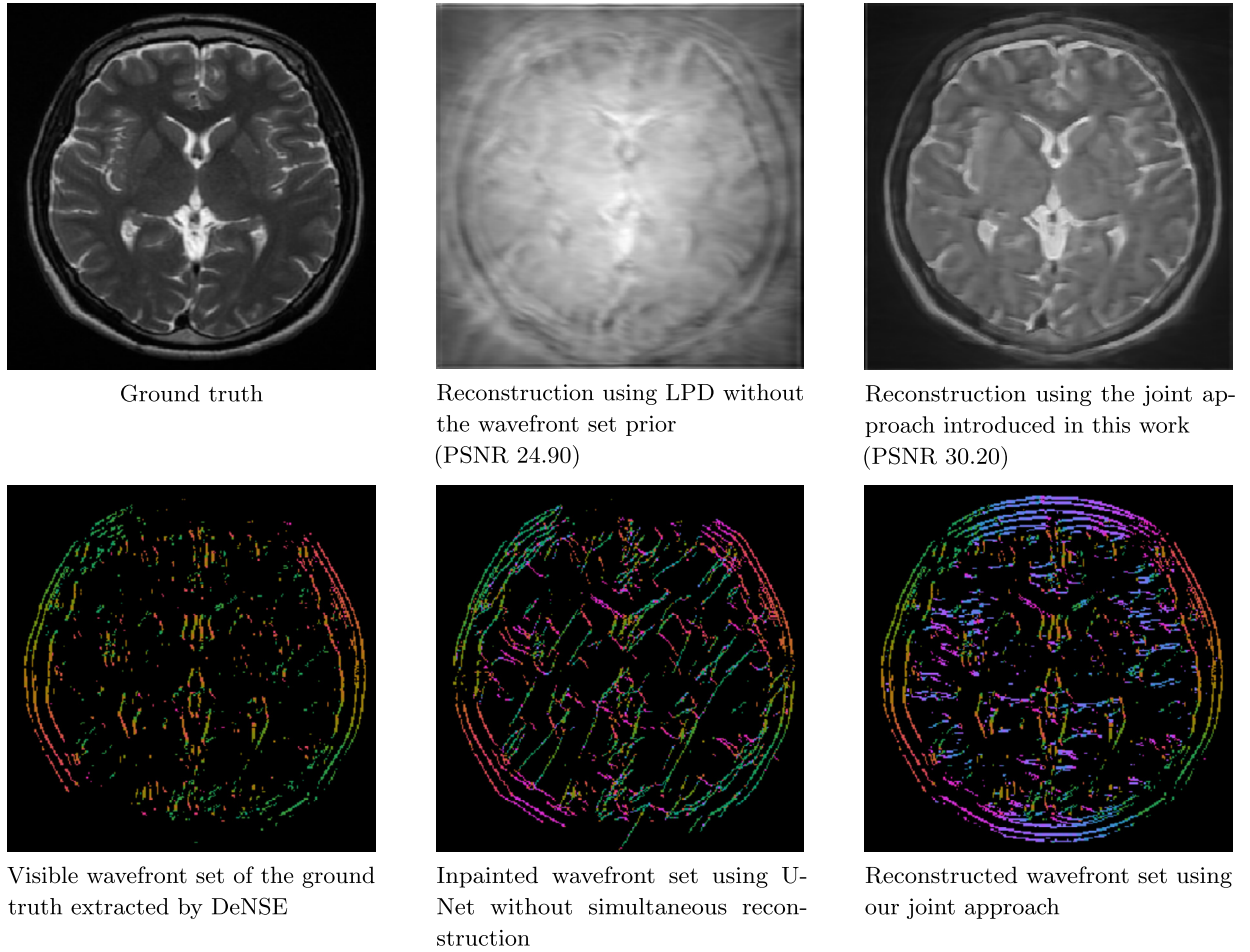
Ground truth	Reconstruction using LPD without the wavefront set prior (PSNR 24.90)	Reconstruction using the joint approach introduced in this work (PSNR 30.20)
Visible wavefront set of the ground truth extracted by DeNSE	Inpainted wavefront set using U-Net without simultaneous reconstruction	Reconstructed wavefront set using our joint approach

Fig. 5. Realistic data set results on the joint CT reconstruction and WFset inpainting for limited-angle case, missing wedge $= 40°$.

In Fig. 6, we see that the Learned Primal-Dual, as well as the joint approach, appear to produce reconstructions that exhibit almost no texture-like areas but instead are piecewise smooth. This effect could potentially be an effect of the underlying data set. In this context, it may be worthwhile to expand the training data set by artificial images with more texture-like features.

Acknowledgments

Appendix A. Notions and results from distribution theory

Below, we collect some essential notions and results from distribution theory and the analysis of the wavefront set. These were used extensively in Subsection 4.3.4.

Definition A.1 *(Essential support [73]).* Let $\Omega \subset \mathbb{R}^n$ be a domain. The *essential support* of $f \in \mathcal{S}'(\Omega)$ is the defined as the set
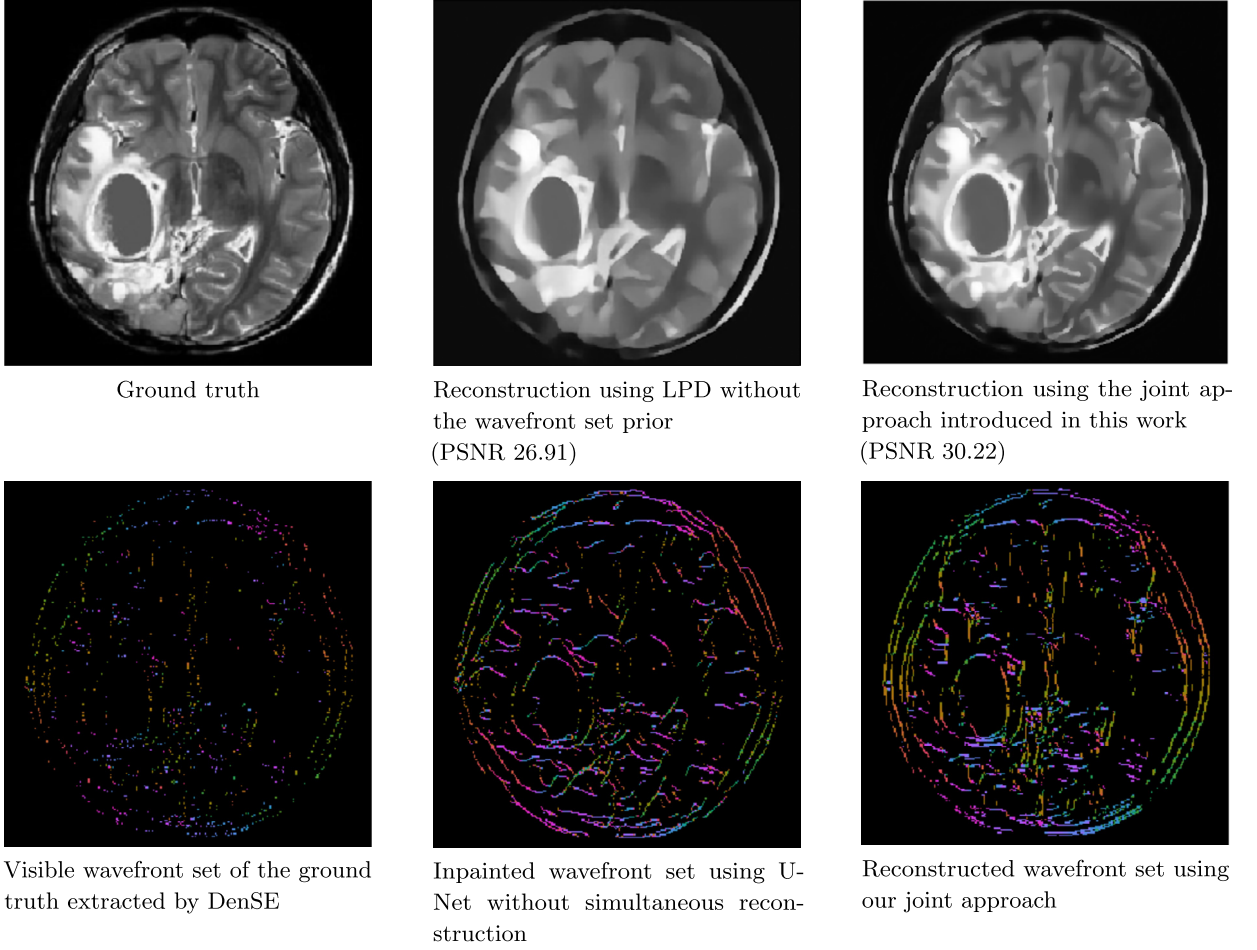
Ground truth

Reconstruction using LPD without the wavefront set prior (PSNR 26.91)

Reconstruction using the joint approach introduced in this work (PSNR 30.22)

Visible wavefront set of the ground truth extracted by DenSE

Inpainted wavefront set using U-Net without simultaneous reconstruction

Reconstructed wavefront set using our joint approach

Fig. 6. Realistic dataset results on the joint CT reconstruction and WFset inpainting for sparse-view case, number of angles = 40.

$$\mathrm{esssup}(f) := \mathbb{R}^n \setminus \bigcup_{U \in \mathfrak{U}} U,$$

where $\mathfrak{U} := \{U \subset \Omega : U \text{ is open and } f|_U = 0\}$ with $f|_U$ denoting the restriction of f to $U \subset \Omega$.

Note that if $f \in \mathcal{C}(\Omega)$, then $\mathrm{esssup}(f) = \mathrm{supp}(f)$, i.e., the essential support coincides with the usual support. We also introduce the notion of positive and negative supports of a distribution.

Definition A.2 *(Positive support, [73]).* Let $\Omega \subset \mathbb{R}^n$ be a domain. The *positive support* of $f \in \mathcal{S}'(\Omega)$ is the defined as the set

$$\mathrm{supp}_+(f) := \overline{\bigcup_{U \in \mathfrak{U}} U},$$

where $\mathfrak{U} := \{U \subset \Omega : f(\phi) > 0 \text{ for all } \phi \in \mathcal{S}(\Omega) \setminus \{0\}, \mathrm{supp}\,\phi \subset U, \phi \geq 0\}$. Finally, the *negative support* of f is defined as $\mathrm{supp}_-(f) := (\overline{\mathrm{supp}_+(f)})^c$.

From the definition, we see that

$$\mathrm{supp}_+(f) = \mathrm{supp}(|f| + f) = \overline{\{x \in \Omega : f(x) > 0\}} \quad \text{whenever } f \in \mathcal{C}^\infty(\Omega).$$
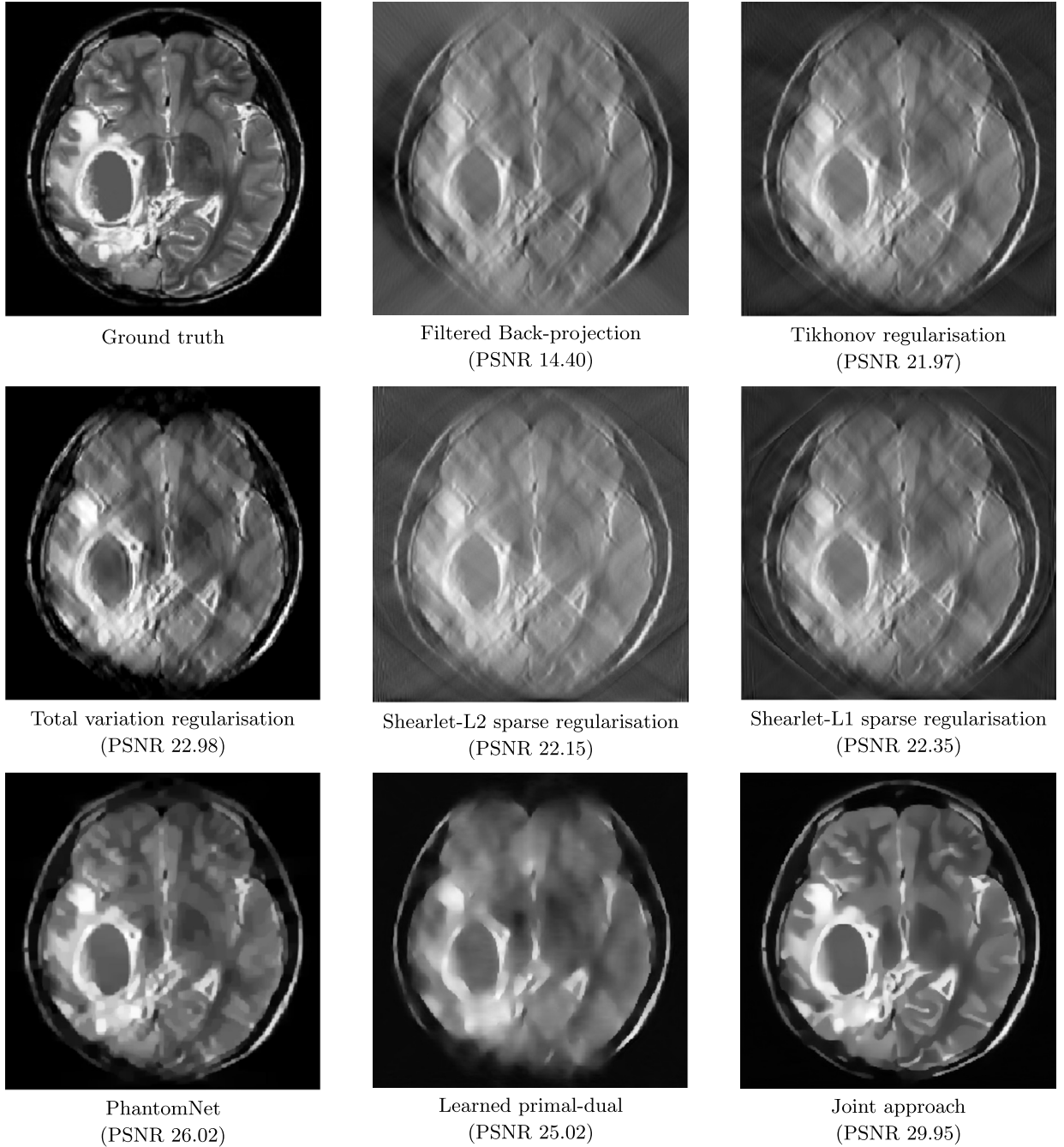
Fig. 7. Realistic dataset results for general benchmarks for limited-angle CT, wedge = $40°$.

We now turn our attention to defining the product of two distributions.

Definition A.3 *(Product of distributions, [11, Definition 2]).* Let $\Omega \subset \mathbb{R}^n$ be a fixed domain and consider the distributions $u, v \in \mathcal{S}'(\Omega)$. We say that $w \in \mathcal{S}'(\Omega)$ is the *product of u and v* if an only if, for each $x \in \mathbb{R}^n$, there exists a test function $\psi \in \mathcal{S}(\Omega)$, with $\psi = 1$ on a neighbourhood of x, so that

$$\widehat{\psi^2 w}(\xi) = \left(\widehat{\psi u} * \widehat{\psi v}\right)(\xi) = \int_{\mathbb{R}^n} \widehat{\psi u}(\xi)\, \widehat{\psi v}(\xi - \eta)\, d\eta \tag{A.1}$$
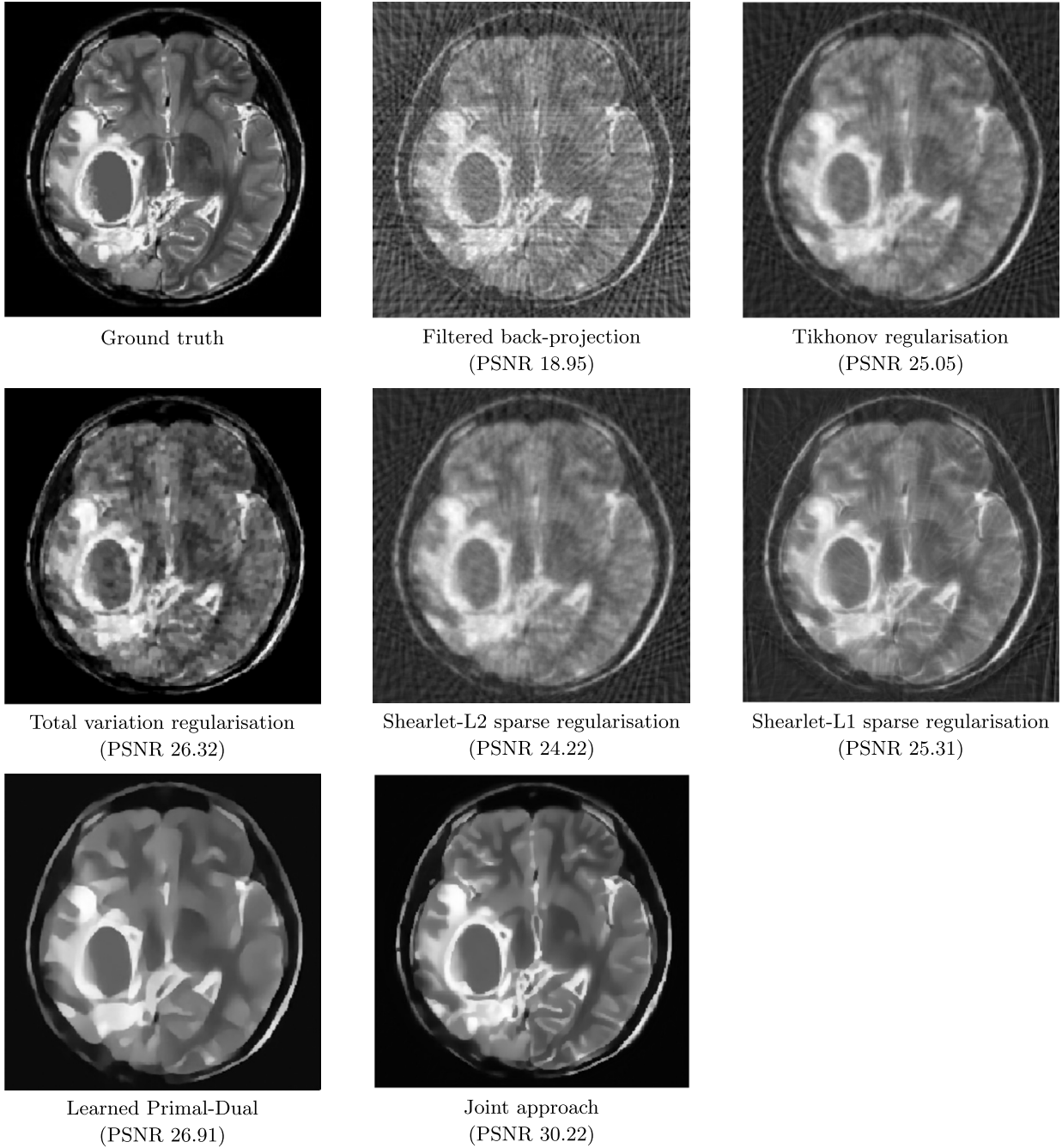
Fig. 8. Realistic data set results for general benchmarks for sparse-view CT using 40 angles only.

converges absolutely for each $\xi \in \mathbb{R}^n$. In addition, under these conditions, we can define w by (A.1).

Theorem A.4 *(Product theorem/Hörmander condition, [11, Section 3.2]). Let $\Omega \subset \mathbb{R}^n$ and $u, v \in \mathcal{S}'(\Omega)$. Assume there are no points $(x, \xi) \in \mathrm{WF}(u)$ such that $(x, -\xi) \in \mathrm{WF}(v)$. Then, w defined as in Definition A.3 is a well-defined unique distribution that is the product uv. Moreover, in this case we have*

$$\mathrm{WF}(uv) \subset S_+ \cup S_u \cup S_v,$$

where

$$S_+ := \left\{\left(x, (\xi + \omega)/\|\xi + \omega\|_2\right) : (x,\xi) \in \mathrm{WF}(u) \text{ and } (x,\omega) \in \mathrm{WF}(v)\right\},$$
$$S_u := \left\{(x,\xi) : (x,\xi) \in \mathrm{WF}(u) \text{ and } x \in \mathrm{supp}(v)\right\},$$
$$S_v := \left\{(x,\omega) : (x,\omega) \in \mathrm{WF}(v) \text{ and } x \in \mathrm{supp}(u)\right\}.$$

Remark A.5. If $u, v \in L^2_{\mathrm{loc}}(\mathbb{R})$ and we define $uv(x) = u(x)v(x)$ almost everywhere, then the multiplication of u and v defined in (A.1) coincides with uv almost everywhere. This holds even if there exist $(x, \lambda) \in \mathrm{WF}(u)$ such that $(x, -\lambda) \in \mathrm{WF}(v)$. To see this, let $x \in \mathbb{R}^2$ and ψ be as in Definition A.3. Then

$$\widehat{\psi^2(uv)}(\xi) = \int_{\mathbb{R}^2} \widehat{\psi u}(\xi)\widehat{\psi v}(\nu - \xi)d\xi$$

holds in an L^2 sense. Moreover, by Plancherel's identity, we have that $\widehat{\psi u}, \widehat{\psi v} \in L^2(\mathbb{R}^2)$, which yields with the Cauchy-Schwarz identity, that

$$\int_{\mathbb{R}^2} \left|\widehat{\psi u}(\xi)\widehat{\psi v}(\nu - \xi)\right|d\xi \leq \|\widehat{\psi u}\|_2 \left\|\widehat{\psi v}(\nu - \cdot)\right\|_2 = \|\widehat{\psi u}\|_2 \|\widehat{\psi v}\|_2 < \infty.$$

This yields absolute convergence in (A.1).

Definition A.6. The L^2-support of $f \in \mathcal{S}'(\Omega)$ is defined as the largest open set on Ω where f is given by an L^2-function:

$$\mathrm{supp}_{L^2}(h) := \bigcup\left\{U \subset \Omega \text{ open } : f\big|_U \in \mathcal{S}'(U)\right\}.$$

Thus, if $x \in \mathrm{supp}_{L^2}(h)$ then there is an open set $x \in U \subset \Omega$ and $f_U \in L^2(U)$ such that $f(\phi) = \int_U f_U(x)\phi(x)dx$ for all $\phi \in \mathcal{S}(U)$.

References

[1] J. Adler, S. Lunz, O. Verdier, C.-B. Schönlieb, O. Öktem, Task-adapted reconstruction for inverse problems, ArXiv preprint, arXiv:1809.00948.
[2] J. Adler, O. Öktem, Learned primal-dual reconstruction, IEEE Trans. Med. Imaging 37 (6) (2018) 1348–1357.
[3] A.H. Andersen, A.C. Kak, Simultaneous algebraic reconstruction technique (SART): a superior implementation of the ART algorithm, Ultrason. Imag. 6 (1984) 81–94.
[4] H. Andrade-Loarca, G. Kutyniok, O. Öktem, P. Petersen, Extraction of digital wavefront sets using applied harmonic analysis and deep neural networks, SIAM J. Imaging Sci. 12 (4) (2019) 1936–1966.
[5] S. Arridge, P. Maass, O. Öktem, C.-B. Schönlieb, Solving inverse problems using data-driven models, Acta Numer. 28 (2019) 1–174.
[6] J.A. Baker, J.Y. Lo, Breast tomosynthesis: state-of-the-art and review of the literature, Acad. Radiol. 18 (10) (2011) 1298–1310.
[7] M. Benning, M. Burger, Modern regularization methods for inverse problems, Acta Numer. 27 (2018) 1–111.
[8] B. Berkels, M. Burger, M. Droske, O. Nemitz, M. Rumpf, Cartoon extraction based on anisotropic image classification, in: L. Kobbelt, T. Kuhlen, T. Aach, R. Westerman (Eds.), Vision, Modeling, and Visualization, IOS Press, Aachen, 2006.
[9] M. Bertero, H. Lantéri, L. Zanni, Iterative image reconstruction: a point of view, in: Proceedings of the Interdisciplinary Workshop on Mathematical Methods in Biomedical Imaging and Intensity-Modulated Radiation, 2008.
[10] R.N. Bracewell, A.C. Riddle, Inversion of fan-beam scans in radio astronomy, Astrophys. J. 150 (1967) 427–434.
[11] C. Brouder, N.V. Dang, F. Hélein, A smooth introduction to the wavefront set, J. Phys. A, Math. Theor. 47 (44) (2014) 443001, 33 pp.
[12] T.A. Bubba, M. Galinier, M. Lassas, M. Prato, L. Ratti, S. Siltanen, Deep neural networks for inverse problems with pseudodifferential operators: an application to limited-angle tomography, SIAM J. Imaging Sci. 14 (2) (2021) 470–505.
[13] T.A. Bubba, G. Kutyniok, M. Lassas, M. März, W. Samek, S. Siltanen, V. Srinivasan, Learning the invisible: a hybrid deep learning-shearlet framework for limited angle computed tomography, Inverse Probl. 35 (6) (2019).

[14] T.-O. Buchholz, F. Jug, Fourier image transformer, ArXiv, arXiv:2104.02555, 2021.
[15] M. Burger, S. Osher, A guide to the TV zoo, in: M. Burger, S. Osher (Eds.), Level Set and PDE Based Reconstruction Methods in Imaging, in: Lecture Notes in Mathematics, vol. 2090, Springer Verlag, 2013.
[16] A. Chambolle, T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, J. Math. Imaging Vis. 40 (1) (2010) 120–145.
[17] R. Chartrand, Shrinkage mappings and their induced penalty functions, in: International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 1026–1029.
[18] G.-H. Chen, J. Tang, S. Leng, Prior image constrained compressed sensing (PICCS): a method to accurately reconstruct dynamic CT images from highly undersampled projection data sets, Med. Phys. 35 (2008) 660–663.
[19] Z. Chen, X. Jin, L. Li, G. Wang, A limited-angle CT reconstruction method based on anisotropic TV minimization, Phys. Med. Biol. 58 (2013) 2119.
[20] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Am. Math. Soc. 39 (2002) 1–49.
[21] M.E. Davison, The ill-conditioned nature of the limited angle tomography problem, SIAM J. Appl. Math. 43 (2) (1983) 428–448.
[22] M. Defrise, C. de Mol, A regularized iterative algorithm for limited-angle inverse Radon transform, Opt. Acta 30 (4) (1983) 403–408.
[23] M. Defrise, C. de Mol, Resolution limits in full – and limited-angle tomography, in: F. Deconinck (Ed.), Information Processing in Medical Imaging, Proceedings of the 8th Conference, Brussels, 29 August – 2 September 1983, Springer Verlag, 1984, pp. 94–105.
[24] X. Deng, X. Zhao, M. Li, H. Li, Limited-angle CT reconstruction with generalized shrinkage operators as regularizers, Inverse Probl. Imaging (2021).
[25] S. Esedoglu, S.J. Osher, Decomposition of images by the anisotropic Rudin-Osher-Fatemi model, Commun. Pure Appl. Math. 57 (2004) 1609–1626.
[26] Y. Fan, O.C. Bohorquez, L. Ying, BCR-Net: a neural network based on the nonstandard wavelet form, J. Comput. Phys. 384 (2019) 1–15.
[27] A. Faridani, D.V. Finch, E.L. Ritman, K.T. Smith, Local tomography II, SIAM J. Appl. Math. 57 (1997) 1095–1127.
[28] A. Faridani, E.L. Ritman, K.T. Smith, Local tomography, SIAM J. Appl. Math. 52 (1992) 459–484.
[29] J. Feliu-Faba, Y. Fan, L. Ying, Meta-learning pseudo-differential operators with deep neural networks, J. Comput. Phys. 408 (2020) 109309.
[30] J. Frikel, Sparse regularization in limited angle tomography, Appl. Comput. Harmon. Anal. 34 (2013) 117–141.
[31] J. Frikel, E.T. Quinto, Characterization and reduction of artifacts in limited angle tomography, Inverse Probl. 29 (2013) 125007.
[32] I.M. Gel'fand, M.I. Graev, N.Y. Vilenkin, Generalized Functions, Volume 5, Integral Geometry and Representation Theory, Academic Press, New York, 1966.
[33] P. Gilbert, Iterative methods for the three-dimensional reconstruction of an object from projections, J. Theor. Biol. 36 (1972) 105–117.
[34] C. Gong, L. Zeng, C. Wang, Image reconstruction model for limited-angle CT based on prior image induced relative total variation, Appl. Math. Model. 74 (2019) 586–606.
[35] R. Gordon, R. Bender, G.T. Herman, Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and x-ray photography, J. Theor. Biol. 29 (1970) 471–481.
[36] G.T. Gullberg, The reconstruction of fan-beam data by filtering the back-projection, Comput. Graph. Image Process. 10 (1979) 30–47.
[37] K. Guo, G. Kutyniok, D. Labate, Sparse multidimensional representations using anisotropic dilation and shear operators, in: Wavelets and Splines, Birkhäuser, Nashville, TN, 2005, pp. 189–201.
[38] K. He, X. Zhang, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), 2016.
[39] S. Helgason, The Radon Transform, 2nd edition, Birkhäuser, Boston, MA, 1999.
[40] A. Hertle, Continuity of the Radon transform and its inverse on Euclidean space, Math. Z. 184 (2) (1983) 165–192.
[41] L. Hörmander, The Analysis of Linear Partial Differential Operators, I, Distribution Theory and Fourier Analysis, Grundlehren der Mathematischen Wissenschaften, Springer Verlag, 1990.
[42] Y. Huang, S. Wang, Y. Guan, A. Maier, Limited angle tomography for transmission x-ray microscopy using deep learning, ArXiv preprint, arXiv:2001.02469.
[43] N. Hyvönen, M. Kalke, M. Lassas, H. Setälä, S. Siltanen, Three-dimensional dental x-ray imaging by combination of panoramic and projection data, Inverse Probl. Imaging 4 (2) (2010) 257–271.
[44] M. Kalke, Image processing methods for limited angle tomography and sparse angle tomography, PhD thesis, Faculty of Science, University of Helsinki, 2014.
[45] D. Karimi, R. Ward, On the computational implementation of forward and back-projection operations for cone-beam computed tomography, Med. Biol. Eng. Comput. 54 (8) (2016) 1193–1204.
[46] A. Katsevich, Local tomography for the generalized Radon transform, SIAM J. Appl. Math. 57 (4) (1997) 1128–1162.
[47] A. Katsevich, Local tomography for the limited-angle problem, J. Math. Anal. Appl. 213 (1) (1997) 160–182.
[48] K.J. Kim, J.A. Fessler, Z. Zhang, Forward-projection architecture for fast iterative image reconstruction in X-Ray CT, IEEE Trans. Signal Process. 60 (10) (2012) 5508–5518.
[49] K. Kothari, M. de Hoop, I. Dokmanic, Learning the geometry of wave-based imaging, in: 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada, 2020.
[50] V.P. Krishnan, E.T. Quinto, Microlocal analysis in tomography, in: Handbook of Mathematical Methods in Imaging, Springer Verlag, 2015, pp. 847–902.
[51] H. Kudo, T. Saito, Sinogram recovery with the method of convex projections for limited-data reconstruction in computed tomography, J. Opt. Soc. Am. A 8 (1991) 1148–1160.

[52] G. Kutyniok, W.-Q. Lim, R. Reisenhofer, ShearLab 3D: faithful digital shearlet transforms based on compactly supported shearlets, ACM Trans. Math. Softw. 42 (1) (2016) 5:1–5:42.
[53] P.J. LaMontagne, T.L. Benzinger, J.C. Morris, S. Keefe, R. Hornbeck, C. Xiong, E. Grant, J. Hassenstab, K. Moulder, A.G. Vlassenko, M.E. Raichle, C. Cruchaga, D. Marcus, Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease, medRxiv, 2019.
[54] S. Lang, Real and Functional Analysis, vol. 142, Springer Verlag, 2012.
[55] J. Latz, On the well-posedness of Bayesian inverse problems, SIAM/ASA J. Uncertain. Quantificat. 8 (1) (2020) 451–482.
[56] Z. Li, A. Cai, L. Wang, W. Zhang, C. Tang, L. Li, B. Liang, N. Yan, Promising generative adversarial network based sinogram inpainting method for ultra-limited-angle computed tomography imaging, Sensors 19 (18) (2019) 3941.
[57] Y. Long, J.A. Fessler, J.B. Balter, 3D forward and back-projection for X-ray CT using separable footprints, IEEE Trans. Med. Imaging 29 (11) (2010) 1839–1850.
[58] A.K. Louis, Incomplete data problems in X-ray computerized tomography, Numer. Math. 48 (3) (1986) 251–262.
[59] D. Ludwig, The Radon transform on Euclidean space, Commun. Pure Appl. Math. 19 (1966) 49–81.
[60] S. Mall, S. Lewis, P. Brennan, J. Noakes, C. Mello-Thoms, The role of digital breast tomosynthesis in the breast assessment clinic: a review, J. Med. Radiat. Sci. 64 (3) (2017) 203–211.
[61] A. Markoe, Analytic Tomography, Encyclopedia of Mathematics and Its Applications, Cambridge University Press, 2006.
[62] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, M. Ebrahimi, Edgeconnect: generative image inpainting with adversarial edge learning, ArXiv preprint, arXiv:1901.00212.
[63] O. Öktem, Mathematics of electron tomography, in: O. Scherzer (Ed.), Handbook of Mathematical Methods in Imaging, 2nd edition, Springer Verlag, 2015, pp. 937–1031.
[64] A.R. Podgorsak, M.M.S. Bhurwani, C.N. Ionita, CT artifact correction for sparse and truncated projection data using generative adversarial networks, Med. Phys. 48 (2) (2021) 615–626.
[65] E.T. Quinto, Singularities of the X-ray transform and limited data tomography in $\mathbb{R}^2$ and $\mathbb{R}^3$, SIAM J. Math. Anal. 24 (5) (1993) 1215–1225.
[66] E.T. Quinto, Exterior and limited-angle tomography in non-destructive evaluation, Inverse Probl. 14 (2) (1998) 339–353.
[67] E.T. Quinto, Artifacts and visible singularities in limited data X-ray tomography, Sens. Imaging 18 (2017).
[68] E.T. Quinto, O. Öktem, Local tomography in electron microscopy, SIAM J. Appl. Math. 68 (5) (2008) 1282–1303.
[69] E.T. Quinto, U. Skoglund, O. Öktem, Electron lambda-tomography, Proc. Natl. Acad. Sci. USA 106 (51) (2009) 21842–21847.
[70] M. Reed, B. Simon, Methods of Mathematical Physics: Fourier Analysis, Self-Adjointness, Academic Press, 1975.
[71] N.A.B. Riis, J. Frøsig, Y. Dong, P.C. Hansen, Limited-data x-ray CT for underwater pipeline inspection, Inverse Probl. 34 (2018) 034002, 16 pp.
[72] O. Ronneberger, P. Fischer, T. Brox U-Net, Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer Verlag, Cham, 2015, pp. 234–241.
[73] W. Rudin, Real and Complex Analysis, McGraw-Hill, 1987.
[74] L. Ruthotto, E. Haber, Deep neural networks motivated by partial differential equations, J. Math. Imaging Vis. 62 (2020) 352–364.
[75] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, F. Lenzen, Variational Methods in Imaging, Applied Mathematical Sciences, vol. 167, Springer Verlag, 2009.
[76] L. Schwartz, Sur l'impossibilité de la multiplication des distributions, C. R. Acad. Sci. 238 (1954) 847–848.
[77] S. Shalev-Shwartz, S. Ben-David, Understanding machine learning: from theory to algorithms, Cambridge University Press, 2014.
[78] L.A. Shepp, B.F. Logan, The Fourier reconstruction of a head section, IEEE Trans. Nucl. Sci. 21 (3) (1974) 21–43.
[79] A.M. Stuart, Inverse problems: a Bayesian perspective, Acta Numer. (2010) 451–559.
[80] S. Tan, Y. Zhang, G. Wang, X. Mou, G. Cao, Z. Wu, H. Yu, Tensor-based dictionary learning for dynamic tomographic reconstruction, Phys. Med. Biol. 60 (2015) 2803.
[81] J. Teuwen, N. Moriakov, C. Fedon, M. Caballo, I. Reiser, P. Bakic, E. García, O. Diaz, K. Michielsen, I. Sechopoulos, Deep learning reconstruction of digital breast tomosynthesis images for accurate breast density and patient-specific radiation dose estimation, Med. Image Anal. 71 (2021) 102061.
[82] A.N. Tikhonov, A. Goncharsky, V.V. Stepanov, A. Yagola, Numerical Methods for the Solution of Ill-Posed Problems, Springer Verlag, 1995.
[83] R. Tovey, M. Benning, C. Brune, M.J. Lagerwerf, S.M. Collins, R.K. Leary, P.A. Midgley, C.-B. Schönlieb, Directional sinogram inpainting for limited angle tomography, Inverse Probl. 35 (2) (2019) 024004.
[84] M.V. de Hoop, H. Smith, G. Uhlmann, V.R.D. der Hilst, Seismic imaging with the generalized radon transform: a curvelet transform perspective, Inverse Probl. 25 (2) (2009) 25005–25021.
[85] I.A. Vainberg, E.I. Kazak, V.P. Kurozaev, Reconstruction of the internal three-dimensional structure of objects based on real-time integral projections, Sov. J. Nondestruct. Test. 17 (1981) 415–423.
[86] W. van Aarle, W.J. Palenstijn, J. Cant, E. Janssens, F. Bleichrodt, A. Dabravolski, J. De Beenhouwer, K.J. Batenburg, J. Sijbers, Fast and flexible x-ray tomography using the ASTRA toolbox, Opt. Express 24 (22) (2016) 25129–25147.
[87] W. van Aarle, W.J. Palenstijn, J. De Beenhouwer, T. Altantzis, S. Bals, K.J. Batenburg, J. Sijbers, The ASTRA toolbox: a platform for advanced algorithm development in electron tomography, Ultramicroscopy 157 (2015) 35–47.
[88] J. Velikina, S. Leng, G.-H. Chen, Limited view angle tomographic image reconstruction via total variation minimization, Proc. SPIE 6510 (2007).
[89] C. Wang, M. Tao, J.G. Nagy, Y. Lou, Limited-angle CT reconstruction via the l_1/l_2 minimization, SIAM J. Imaging Sci. 14 (2) (2021) 749–777.
[90] C. Wang, L. Zeng, Y. Guo, L. Zhang, Wavelet tight frame and prior image-based image reconstruction from limited-angle projection data, Inverse Probl. Imaging 11 (2017) 917–948.

[91] T. Wang, K. Nakamoto, H. Zhang, H. Liu, Reweighted anisotropic total variation minimization for limited-angle CT reconstruction, IEEE Trans. Nucl. Sci. 64 (2017) 2742–2760.
[92] Y. Wang, W. Zhang, A. Cai, L. Wang, C. Tang, Z. Feng, L. Li, N. Liang, B. Yan, An effective sinogram inpainting for complementary limited-angle dual-energy computed tomography imaging using generative adversarial networks, J. X-Ray Sci. Technol. 29 (1) (2021) 37–61.
[93] J. Xu, Y. Zhao, H. Li, P. Zhang, An image reconstruction model regularized by edge-preserving diffusion and smoothing for limited-angle computed tomography, Inverse Probl. 35 (2019) 085004.
[94] L. Zhang, L. Zeng, Y. Guo, ℓ_0 regularization based on a prior image incorporated non-local means for limited-angle x-ray CT reconstruction, J. X-Ray Sci. Technol. 26 (2018) 481–498.
[95] Z. Zhang, B. Chen, D. Xia, E.Y. Sidky, X. Pan, Directional-TV algorithm for image reconstruction from limited-angular-range data, Med. Image Anal. 70 (2021) 102030.
[96] J. Zhao, Z. Chen, L. Zhang, X. Jin, Unsupervised learnable sinogram inpainting network (SIN) for limited angle CT reconstruction, ArXiv preprint, arXiv:1811.03911.
[97] B. Zhou, K.S. Zhou, J.S. Duncan, C. Liu, Limited view tomographic reconstruction using a cascaded residual dense spatial-channel attention network with projection data fidelity layer, IEEE Trans. Med. Imaging 40 (7) (2021) 1792–1804.