Shearlets as Feature Extractor for Semantic Edge Detection: The Model-Based and Data-Driven Realm

3 authors:

Héctor Andrade Loarca
Technische Universität Berlin
3 PUBLICATIONS **2** CITATIONS

SEE PROFILE

Gitta Kutyniok
Technische Universität Berlin
224 PUBLICATIONS **6,378** CITATIONS

SEE PROFILE

Öktem Ozan
KTH Royal Institute of Technology
65 PUBLICATIONS **505** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project   Convex optimization methods for large-scale non-smooth optimization View project

Project   Scalable Frames View project

Shearlets as Feature Extractor for Semantic Edge Detection: The Model-Based and Data-Driven Realm

Héctor Andrade-Loarca[1], Gitta Kutyniok[1,2,3] and Ozan Öktem[4]

Abstract

Semantic edge detection has recently gained a lot of attention as an image processing task, mainly due to its wide range of real-world applications. This is based on the fact that edges in images contain most of the semantic information. Semantic edge detection involves two tasks, namely pure edge detecion and edge classification. Those are in fact fundamentally distinct in terms of the level of abstraction that each task requires, which is known as the distracted supervision paradox that limits the possible performance of a supervised model in semantic edge detection. In this work, we will present a novel hybrid method to avoid the distracted supervision paradox and achieve high-performance in semantic edge detection. Our approach is based on a combination of the model-based concept of shearlets, which provides probably optimally sparse approximations of a model-class of images, and the data-driven method of a suitably designed convolutional neural netwok. Finally, we present several applications such as tomographic reconstruction and show that our approach signifiantly outperforms former methods, thereby indicating the value of such hybrid methods for the area in biomedical imaging.

Keywords: Multiscale geometric analysis, Harmonic analysis, Deep learning, Feature extraction.
Mathematics Subject Classification: 42Bxx, 35A18, 65T60, 68T10.

1 Introduction

In computer vision, semantic edge detection is the task of detecting edges and object boundaries in natural images and classifying the points in those edges from a finite set of classes, which, for instance, represent the objects the edges belong to [38] or the orientation of the edge at that particular point [3]. The recent interest from the research community in semantic edge detection is mainly driven by its far-reaching applications in imaging related tasks such as object recognition, semantic segmentation, and image reconstruction.

Semantic edge detection combines two different classification tasks. The first is classical *category-agnostic* edge detection, which can be viewed as a pixel-wise binary classification for determining whether a pixel belongs to an edge or not. The second is the recognition of the classes of pixels in an image that belong to edges.

One can perform semantic edge detection using a model-based or a data-driven approach; each comes with their strengths and shortcomings. The main idea presented here is to overcome the shortcomings by combining elements of model-based and data-driven approaches for semantic edge detection. We remark that this objective follows a common thread of current research, namely aiming for an optimal combination of model-based or a data-driven methodologies.

Our conceptually general approach aims to combine advances from several fields. First, it utilizes properties that the representation systems of shearlets has, foremost to optimally represent singularities in signals [13]. Second, it also takes advantage of microlocal analysis for describing how singularities transform when acted

---

[1]Institut für Mathematik, Technische Universität Berlin, 10623 Berlin, Germany, {kutyniok,andrade}@math.tu-berlin.de
[2]Fakultät Elektrotechnik und Informatik, 10587 Berlin, Germany, kutyniok@math.tu-berlin.de
[3]Department of Physics and Technology, University of Tromsø
[4]Department of Mathematics, KTH - Royal Institute of Technology, SE-100 44 Stockholm, Sweden, ozan@kth.se

upon by a wide range of operators [17]. Finally, it in addition leverages on the proven track record that convolutional neural networks has had in image classification [23].

1.1 Model-Based Semantic Edge Detection

Many approaches to identify singularities in images are model-based. These methods usually involve two steps: a filtering step to enhance edge-like features and a classification step to identify pixels belonging to edges.

The aforementioned features are extracted using simple rules and heuristics, e.g., convolution with local difference filters correspond to operating on the image with Roberts [32], Sobel [33], and Prewitt [28] operators. In a similar manner, the well-known Canny edge detector [9] corresponds to convolving the image with a Gaussian kernel to further identify those pixels where the gradient is high.

There have been attempts in the past to also model the semantic information of detected edges as in [15]. This work was in fact among the first publications to propose a principled way to combine generic object detectors with bottom-up contours for semantic edge detection.

Determining the orientation of an edge is particularly important in inverse problems, since this information is essential in relating edges in data to those in the signal [17]. Wavefront set extraction refers to semantic edge detection, where the classification of the edges is based on their orientation. The continuous theory of wavefront set resolution via multiscale directional systems (e.g., shearlets [18]) allows one to design model-based approaches for wavefront set extraction. These are essentially a digital implementation of the continuous theory, which filters the images before performing the corresponding classification. In Section 3, the concept of a wavefront set will be formally defined as well as the ability of shearlets to resolve it.

An example of such an approach is the shearlet-based algorithm in [36], which uses digital shearlets to filter an image in order to highlight the features corresponding to different orientations and scales. One then performs a simple clustering classification algorithm to classify the corresponding directions. A more recent approach is [30, 31], where a general directional system, known as symmetric molecules, is used to filter the directional features of images to then classify them to be edge, ridge, or blob.

These model-based approaches for semantic edge detection rely on 'first principles' from approximation theory and are easy to interpret, hence can also easier to improve upon. On the other hand, the use of rigid heuristics regarding the characterization of singularities makes it difficult to utilize these methods in real world application, where the data represents empirically defined function classes.

1.2 Data-Driven Semantic Edge Detection

More recently, as part of the success stories of machine learning and its successes in addressing various tasks in modern computer vision, a set of deep neural network architectures for semantic edge detection [38, 4, 24, 39] have appeared. One needs to stress that these have set a new state-of-the-art of semantic edge detection.

In broad terms, these methods use similar principles as the model-based ones, i.e., learning filters using convolutional layers and subsequently classifying the corresponding edge pixels by sigmoid or softmax classifiers. Since each convolutional layer represents a level of abstraction of the features in the target images, the initial layers will represent 'simple' edges, whereas deeper layers will represent more 'complex' features. In that sense, the two steps involved in semantic edge detection, which are semantic-agnostic edge detection and classification followed by edge classification, are conceptually far from each other. Therefore, there is no straightforward way of jointly learning how to extract and classify the edges. This limitation in semantic edge detection is known as *the distracted supervision paradox* [24].

1.2.1 The Distracted Supervision Paradox

As mentioned before, semantic-agnostic edge detection involves locating fine detailed edges by capturing discontinuities among image regions. This makes mainly use of low level features, whereas edge classification requires identifying high-level semantics by summarizing different appearance variations of the target classes.

This distracted supervision paradox prevents state-of-the-art end-to-end semantic edge detection method based on deep supervision.

On the other hand, one could directly use high-level convolution features for semantic classification and low-level convolutional ones for non-semantic edge detection by jointly optimizing over the two corresponding losses. However, such a straightforward approach was in [38] shown to actually decrease the performance compared with directly learning semantic edges without deep supervision.

Given the seriousness of this limitation, researchers have tried non-standard ways to avoid the distracted supervision paradox. In [38] the authors propose the CASENet architecture based on a concatenation of convolutional residual networks. They used a new skip-layer architecture, where category-wise edge activations at the top convolutional layers are shared and fused with the same set of bottom layer features. Training such a network with a multi-label loss function on the fused activations leads to state-of-the-art on semantic edge detection. In this case, the low-level features are used to augment top classifications. More recently, a new training approach for CASENet was proposed in [39] that is referred to as the *simultaneous edge alignment and learning* (SEAL). Here, one simultaneously aligns ground truth edges and learns semantic edge detectors. Inspired by these approaches, [24] proposed an end-to-end architecture using convolutional residual layers followed by an information converter layer to transform information coming from low-level features and change it into different representations. This is needed for training category-agnostic edge detection and semantic edge classification, respectively. It allows for a single deep convolutional neural network backbone representation while outperforming CASENet and SEAL. The approach based on using information converter is known as *diverse deep supervision*.

The main drawback of these approaches lies in the complexity of the related deep neural network architectures, which represents an elaborate way to avoid the distracted supervision paradox. Furthermore, the large number of network parameters makes those methods slow and difficult to train.

A new approach, based on a novel architecture DeNSE for wavefront set extraction was recently introduced in [3]. In this work the digital shearlet transform is used to map the input image to a representation that is well adapted for edge detection and further classification. The power of multiscale directional system for representing singularities allows one to use a significantly smaller architecture to obtain state-of-the-art results in semantic edge detection. Furthermore, if the specific semantic edge detection task has N target classes, the full problem can be seen as a multi-label classification class with $N + 1$ target classes with the extra class corresponding to the class of edge pixels. In addition to the heavy lifting performed by the shearlet transform on the classification task by offering such a convenient representation of edges, the DeNSE architecture also separates the multi-label classification task into $N + 1$ individual binary classifiers.

This strategy allows the network to achieve high accuracy, since binary classification is significantly easier than the multi-label classification. In addition, the separation of the multi-label classification problem into small binary classifiers avoids the model to encounter the distracted supervision paradox. The edge detection and the edge classification, which originally needed different levels of abstraction, are now learned independently. DeNSE was for these reasons able to outperform all other available semantic edge detection methods where the classes were the orientations of the edges. In Section 3.3 we present a way to avoid the distracted supervision paradox in the case of wavefront set extraction. Moreover, this concept is extended to general semantic edge detection in Section 4.

We will however first introduce the important ideas that motivate using multiscale directional representations to sense a signal on different scales and directions. This will also show how these representations optimally represent multidimensional data such as images by extracting information on their singularities. This will provide the necessary background in understanding how multiscale directional representations, like shearlets, help in reducing the complexity of edge detection and classification.

1.3 Multiscale Directional Systems as Feature Extractors

Multiscale systems play an important role in applied and computational harmonic analysis for the analysis of multiscale features of a signal. In the case of 1D signals, wavelets and similar systems have been used to detect the singular support [25]. This is achieved by studying the asymptotic behavior of the corresponding

localized coefficients as the scale converges to zero. Rapid decay then translates to regularity properties of the function itself.

One can in this sense regard the wavelet transform as a feature extractor that is well suited for detecting singularities. The feature vector would then be the wavelet coefficients. However, singularities of 2D signals are not necessarily point-like. In fact, they may very well be curve-like structures (curvilinear singularities). Wavelet type expansions are, due to their isotropic nature, therefore less suitable for representing such singularities.

Geometric multiscale analysis.   The area of geometric multiscale analysis can be regarded as a sub-field of applied and computational harmonic analysis for the study of directional multiscale systems, which address the aforementioned limitation of wavelets. Among the first examples are curvelets [8], which use parabolic (non-isotropic) scaling and rotation to efficiently represent curvilinear singularities.

Curvelets represented a breakthrough in the optimal approximation of curvilinear singularities — formalized as the model of cartoon-like functions — and can be seen as the starting point of geometric multiscale analysis. However, the system comes with a major practical drawback. The rotation operator used for varying the orientation of the coefficients in the system does not have a digitization that results in an implementation, which is consistent with the continuous theory [7].

One main reason for the development of the *shearlet* system [13] was in fact to overcome the aforementioned limitation associated with curvelets. shearlets use the shearing operator — instead of the rotation operator — to vary the orientation of the coefficients. This allows for a consistent and faithful digitization. In this sense shearlets provide a uniform concept for the continuum and digital realm.

The implementation of shearlets indeed inherits most of the theoretical properties of the continuous transform, including the optimal representation of the model class of cartoon-like functions [19]. Furthermore, there even exists a set of shearlet generators with compact support that also provide optimally sparse approximations of cartoon-like functions, thereby allowing high spatial localization [20]. Similar to regarding the wavelet transform as a feature extractor for point singularities, the shearlet transform can be used to extract information relevant for curvilinear singularities detection, like edges in images [22]. At the same time, shearlets do allow the computation of the orientation of curvilinear singularities as well. This establishes a connection to the concept of microlocal wavefront sets. One can in fact prove that the microlocal wavefront set can be characterised through the asymptotic decay of the shearlet coefficients [18]. Fixing the shearing and location parameters amounts to characterizing the wavefront set at a specific scale (resolution).The formal definition of continuous shearlets and their connection to the wavefront set is given in Section 3.1.

Digital signals.   Once an analog signal represented by a real valued function — one can think of an image — is digitized, it is not meaningful to regard its singularities in the sense of lack of regularity in analysis. For such a signal, the notion of singularity refers to an abrupt change in intensity values, which in imaging applications typically corresponds to the presence of an edge. It is indeed challenging to define a precise notion of a digital edge extractor that is practically usable and consistent with a corresponding continuous version [3]. Despite this problem, it is still possible to approximate a singularity detector in the digital case, by using diverse heuristics. An approach based on the shearlet transform of an image is presented in detail in Section 3.2, see also [3].

In general, deep learning models employed for imaging tasks — including those for semantic edge detection — which involve convolutional layers can be divided in different separate stages, where feature extraction step is done at first using convolutional filters playing the role of dictionary elements, to transform the data in a particular feature vector which is convenient for the specific task. In most of the cases, the harder part of the whole process is precisely the feature extraction containing most of model parameters.

It is conceivable that applying a specific, analytically defined, feature extraction such as the shearlet transform as a pre-processing step can benefit the learning process by reducing the complexity of the overall task. Intuitively, such as pre-processing should perform most of the heavy lifting and reducing the number of parameters to be learned. This idea was applied to semantic edge extraction for the case of wavefront set

4

extraction (see [3]). But it seems natural to extend this concept to other tasks of semantic edge extraction, and we will cover this idea carefully in the next sections.

1.4 Applications of Semantic Edge Detection

The concept of edge detection is one of the principal problems in the field of image processing and computer vision. This is due to the fact that edges in images represent boundaries of objects and carry most of the information of the associated physical scene [34, 27, 5, 6]. It has also been shown in [27] that the human visual cortex performs multiple operations of image processing, the first of which is rough sketching involving edge detection in order to reduce the amount of information that needs to be fully acquired having that the visual cortex is sparsely connected.

Based on its important role in information processing for imaging, solely edge detection is utilized in a wide range of applications in computer vision and other fields. In particular, detecting edges allows to track objects within different frames of a video by tracking the points in their boundaries. This can, for instance, be used for object depth estimation [2] an 3D image reconstruction.

Moreover, semantic edge detection has a deeper impact in computer vision than solely edge extraction. If the edge classes are defined by the object the corresponding edge belongs to, one can, in particular, perform object proposal generation [4], occlusion reasoning [1], object detection [11, 12], and image-based localization [29].

In the special case when the performed semantic edge detection includes the extraction of the wavefront set as well, the acquired semantic edges can be used to solve a wide range of inverse problems. One particular class of examples are inverse problems whose forward operator is a Fourier integral operator, which are operators arising from biomedical imaging problems such as computed tomography (CT) and magnetic resonance imaging (MRI). Section 7.3 is dedicated to an application in inverse problems of special interest in biomedical imaging, namely CT image reconstruction. To solve this, we exploit the fact that the X-ray transform is a Fourier integral operator, so the wavefront set of the image is prescribed by the wavefront set of the projected data. We refer to Section 2.4 for more details.

1.5 Contributions

The main contributions of our work are three-fold:

- **Hybrid semantic edge detection** by using the carefully designed model-based shearlet transform to perform semantic edge detection on a convenient image representation for directional edge extraction. This concept is first introduced for the particular case of wavefront set extraction (Section 3) and then extended to general semantic edge detection (Section 4).

- **Distracted supervision paradox avoidance** by splitting the N-classes semantic edge detection into $N + 1$ individual classifiers (Section 33.3).

- **Applications of semantic edge detection to inverse problems** by introducing the micro-canonical relation that prescribes the wavefront set of an image from the wavefront set of its transformation under a forward model. We, in particular, apply this to the inverse problem of Computed Tomography (CT) (Section 7).

In order to evaluate our novel method, we present benchmarks on the human-annotated Semantic Boundaries Dataset (SBD) [14] and a custom dataset made of random ellipses and analytically defined semantic edges.

2   Microlocal Analysis

2.1   Basic Definitions and Properties of Distributions

Distributions appear in physics in various forms, e.g., the Dirac δ-function is widely used for describing the density of a point mass. The δ-function is described as a positive function with support at a single point that integrates to 1. It is however easy to see that there are no functions with these properties. Still, one can compute with such function and get reasonable results. Next, distributions also arise in mathematics and in particular so in the theory of partial differential equations (PDEs). For example, fundamental solutions to PDEs are usually singular distributions and the behaviour of the singularities of these distributions encodes the behaviour of the solutions. The aim of distribution theory is to formalize the definition and mathematical calculus of distributions. In the following, we outline the basic parts of distribution theory that are necessary for microlocal analysis.

Let $\Omega \subset \mathbb{R}^n$ be an open set and $\mathcal{E}(\Omega) := C^\infty(\Omega)$ is the vector space of real (or complex) valued smooth functions on Ω. Next, $\mathcal{D}(\Omega) := C_0^\infty(\Omega)$ is the corresponding set of smooth functions that are compactly supported in Ω. Distributions are now continuous linear functionals (with the weak-* topology) on these function spaces.

A more precise definition requires describing the topologies for $\mathcal{E}(\Omega)$ and $\mathcal{D}(\Omega)$. The space $\mathcal{E}(\Omega)$ is a Fréchet space with the family of semi-norms

$$d_{\alpha,K}(f) := \sup_{x \in K} \left| \partial^\alpha f(x) \right| \quad \text{where } K \text{ runs over all compact subsets } K \subset \Omega. \tag{2.1}$$

Hence, a sequence of functions f_n converges to f in $\mathcal{E}(\Omega)$ if and only if all its derivatives converge uniformly on compact subsets. In particular, $\mathcal{E}(\Omega)$ is Fréchet space. The topology on $\mathcal{D}(\Omega)$ is more complicated to define. A sequence f_n converges in $\mathcal{D}(\Omega)$ to f if there exists a compact set K such that supp $f_n \subset K$, supp $f \subset K$, and all derivatives of f_n converge uniformly in K.

Definition 2.1. *The set $\mathcal{E}'(\Omega)$ is the topological dual of $\mathcal{E}(\Omega)$, i.e., the space of continuous linear functionals on $\mathcal{E}(\Omega)$. Elements in $\mathcal{E}'(\Omega)$ are called* compactly supported distributions *on Ω. Likewise, the set $\mathcal{D}'(\Omega)$ is the topological dual of $\mathcal{D}'(\Omega)$, i.e., the space of continuous linear functionals on $\mathcal{D}(\Omega)$. Elements in $\mathcal{D}'(\Omega)$ are called* distributions *on Ω.*

Next, to define the support of a distribution we first observe that if $\Omega_0 \subset \Omega$ is an open subset, then $\mathcal{D}(\Omega_0)$ is a closed subspace of $\mathcal{D}(\Omega)$. Furthermore, there is a natural restriction map $\mathcal{D}'(\Omega) \to \mathcal{D}'(\Omega_0)$ for any open subset. This ensures that the following definition is well-defined.

Definition 2.2. *Let $f \in \mathcal{D}'(\Omega)$. The* support *supp f of f is the smallest closed set K such that the restriction of f to $\Omega \setminus K$ is 0.*

Distributions are frequently called 'generalised functions' and the next example motivates this alternative terminology. If $f \in L^1_{\text{loc}}(\Omega)$ (a locally integrable function), then f is a distribution with the standard definition

$$f(u) := \int_\Omega u(x) f(x) \, dx \quad \text{for } u \in \mathcal{D}(\Omega).$$

The map $L^1_{\text{loc}}(\Omega) \to \mathcal{D}'(\Omega)$ is injective, which means that f is almost everywhere determined by the distribution. In particular, every smooth function defines a distribution and the support of a function as a distribution coincides with its support as a function. Thus, we obtain the following inclusions:

$$\mathcal{D}(\Omega) \subset \mathcal{E}'(\Omega) \subset \mathcal{D}'(\Omega) \quad \text{and} \quad \mathcal{E}(\Omega) \subset \mathcal{D}'(\Omega).$$

The most common example of a distribution that is not a function is the Dirac δ-distribution δ_{x_0} at a point $x_0 \in \Omega$. It is defined by $\delta_{x_0} \colon \mathcal{D}(\Omega) \to \mathbb{R}$ as

$$\delta_{x_0}(f) = \langle \delta_{x_0}, f \rangle = f(x_0).$$

This is a distribution with support and singular support equal to $\{x_0\}$.

Next, we define derivatives of distributions. From the definition, it is clear that any distribution can be arbitrarily often differentiated and the result will again be a distribution. For example, any function in $L^1_{loc}(\Omega)$ has distributional derivatives of any order. Thus, distribution theory can be thought of as the completion of differential calculus, similar to how Lebesgue integration theory is a completion of integral calculus.

Definition 2.3. *The partial derivative $\partial^\alpha u$ of $u \in \mathcal{D}'(\Omega)$ is defined by*

$$(\partial^\alpha u)(f) := (-1)^{|\alpha|} u(\partial^\alpha f) \quad \text{for } f \in \mathcal{D}(\Omega).$$

Finally, to extend the notion of a Fourier transform to distributions, it is convenient to introduce a third space of distributions. This is the space of Schwartz distributions $\mathcal{S}'(\mathbb{R}^n)$ (or tempered distributions) that is defined as the topological dual of the space $\mathcal{S}(\mathbb{R}^n)$, which are smooth functions in $\mathbb{R}^n$ where the following semi-norms are finite:

$$d_{\alpha,\beta}(f) := \sup_x |x^\alpha \partial^\beta f(x)| \tag{2.2}$$

This space is a Fréchet space and tempered distributions is the dual of $\mathcal{S}(\mathbb{R}^n)$. One can now define the Fourier transform of a tempered distribution $u \in \mathcal{S}'(\mathbb{R}^n)$ as

$$\widehat{u}(f) := u(\widehat{f}) \quad \text{for } f \in \mathcal{S}(\mathbb{R}^n).$$

By duality (using the Plancherel formula) one can show that the Fourier transform with the above definition extends to a weak-* continuous linear map from $\mathcal{S}'(\mathbb{R}^n)$ to $\mathcal{S}'(\mathbb{R}^n)$.

As a final note, the above constructions extend to functions on manifolds. Microlocal analysis deals with the detailed analysis of how singularities of distributions can be localized in phase space. This leads to the notion of wavefront sets, which refines the notion of singular support.

2.2 The Wavefront Set

The wavefront set of a distribution describes simultaneously the location and 'direction' of its singularities. The starting point is the well-known fact that a compactly supported function or distribution is infinitely differentiable if and only if its Fourier transform decays as $O(|\xi|^{-m})$ as $|\xi| \to \infty$ for every $m = 1, 2, \ldots$. This characterizes the singular support $\operatorname{sing\,supp}(f) \subset \mathbb{R}^n$, which is defined as the complement of the largest open set where f is C^∞.

The singular support is however not invariant under smooth change of coordinates, so it is difficult to use it for understanding how an operator transforms the singularities of a distribution. To address this drawback, we first localize the above characterization to a point $x_0 \in \mathbb{R}^n$ by multiplying with a smooth cut-off that is non-zero at x_0. A further localization (micro-localization) is obtained by identifying those 'directions' in the frequency space where the already localized Fourier transform does not decay sufficiently fast. This singles out those directions (if there are any) that causes f to be singular at x_0. In particular, $f \in \mathcal{D}'(\Omega)$ has a singularity at $x_0 \in \Omega$ in direction $\xi_0 \in \mathbb{R}^n \setminus \{0\}$ if for any smooth cutoff function ψ at x_0, the Fourier transform $\widehat{\psi f}$ does not decay rapidly in any open conic neighbourhood of the ray $\{s\xi_0 : s > 0\}$. The wavefront set of $f \in \mathcal{D}'(\Omega)$ can now be defined as the set of all tuples (x_0, ξ_0) such that f is singular at x_0 in direction ξ_0. With this definition, the wavefront set is a closed subset of $\mathbb{R}^n \times (\mathbb{R}^n \setminus \{0\})$ that is conic in the second variable, i.e., it can be seen as a subset of $\mathbb{R}^n \times S^{n-1}$.

The wavefront set can also be defined in an equivalent manner as a subset of the cotangent bundle of $\mathbb{R}^n$ with the zero section removed, i.e., as a subset of $T^*(\mathbb{R}^n) \setminus \{0\}$. Such a definition has the advantage that it extends readily to distributions on smooth manifolds, which is required in many applications such as tomographic imaging. We will therefore adopt the viewpoint where the wavefront set is a subset of the cotangent bundle. Hence, before stating the formal definition, we make a small digression into differential geometry.

Let M be a smooth manifold and let O_x denote the algebra of smooth functions defined in a neighbourhood of $x \in M$. Any functional $t \colon O_x \to \mathbb{R}$ such that $t(fh) = t(f)f(x) + f(x)t(h)$ for $f, h \in O_x$ is called a tangent vector in M at x. The vector space of all tangent vectors at x, which is denoted by $T_x(M)$, is called the tangent space of M at x. Its dual vector space, denoted by $T_x^*(M)$, is called the cotangent space of M at x. An element τ of the cotangent space is called a covector (or differential form of degree 1) at x. Next, if $M_0 \subset M$ is sub-manifold of M, then t is tangent to M_0 if $t(f) = 0$ for any function $f \in O_x$ that vanishes in M_0. Any functional $\tau \in T_x^*(M)$ such that $\tau(t) = 0$ for any tangent vector t to M_0 is called a conormal covector to M_0 at x. Finally, the tangent bundle $T(M)$ and the cotangent bundle $T^*(M)$ are defined as the union of the tangent and cotangent spaces as $x \in M$ varies, i.e.,

$$T(M) := \bigcup_{x \in M} T_x(M) \quad \text{and} \quad T^*(M) := \bigcup_{x \in M} T_x^*(M) = \big\{(x, \eta) : x \in M, \eta \in T_x^*(M)\big\}.$$

Definition 2.4. *Let M be a smooth manifold and $f \in \mathcal{D}'(M)$. We say that f is* microlocally smooth *at $(x_0, \xi_0) \in T^*(M) \setminus \{0\}$ if there exists a neighbourhood $U \subset V$ of x_0 and $\psi \in C_0^\infty(U)$ with $\psi(x_0) \neq 0$ and a conic neighbourhood Γ of ξ_0 such that for constants C_m we have*

$$\big|(\widehat{\psi f})(\xi)\big| \leq C_m \big(1 + |\xi|\big)^{-m} \quad \text{for } \xi \in \Gamma \text{ with } m = 1, 2, \ldots. \tag{2.3}$$

The C^∞-wavefront set $\mathrm{WF}(f) \subset T^*(M) \setminus \{0\}$ is the set of (x_0, ξ_0) where f is not microlocally smooth.

The wavefront set is a closed set that is conic in the ξ-variable, i.e., $(x, \xi) \in \mathrm{WF}(f)$ if and only if $(x, \lambda\xi) \in \mathrm{WF}(f)$ for any $\lambda > 0$. Next, it is clear that f is equal to a C^∞-function near x whenever $(x, \xi) \notin \mathrm{WF}(f)$ for all $\xi \neq 0$. In fact, the x-projection of $\mathrm{WF}(f)$ equals the singular support of f [16, Proposition 8.1.3], so the wavefront set contains the information about the location of singularities. Furthermore, $\mathrm{WF}(f)$ is also invariant under a diffeomorphic change of variables, which is not true for the singular support.

When $M \subset \mathbb{R}^n$ is an open set, then the differentials $\mathrm{d}x_1, \mathrm{d}x_2, \ldots, \mathrm{d}x_n$ are a basis of $T_x^*(\mathbb{R}^n)$ for any $x \in \mathbb{R}^n$. Hence, any element in $T_x^*(\mathbb{R}^n)$ can be written as

$$\xi \mathrm{d}x = \xi_1 \mathrm{d}x_1 + \ldots + \xi_n \mathrm{d}x_n \in T_x^*(\mathbb{R}^n) \quad \text{with } \xi \in \mathbb{R}^n.$$

The tangent bundle of M is therefore isomorphic to $M \times \mathbb{R}^n$, so the cotangent bundle has the structure $T^*(M) = M \times (\mathbb{R}^n)^*$ where $(\mathbb{R}^n)^*$ is the space dual to $\mathbb{R}^n$.

To give some examples, if $f \in \mathcal{D}'(\mathbb{R}^n)$ is a smooth density on the hypersurface $x_n = 0$, i.e., $f(x) = f(x', x_n) = g(x')\delta_0(x_n)$ for some $g \in C^\infty(\mathbb{R}^{n-1})$, then

$$\mathrm{WF}(f) = \{(x', 0; 0, \xi_n) : x' \in \mathrm{supp}(g) \text{ and } \xi_n \neq 0\}.$$

Another example is when f is the characteristic function of a domain $\Omega \subset \mathbb{R}^n$ with smooth boundary. Then $\mathrm{WF}(f)$ is

$$\mathrm{WF}(f) = \big\{(x, \xi \mathrm{d}x) : x \in \partial\Omega \text{ and } \xi \text{ is normal to } \partial\Omega \text{ at } x\big\}.$$

Furthermore, if f is a linear combination of characteristic functions on sets with smooth boundaries, then $\mathrm{WF}(f)$ is the union of the conormal bundles to the individual sets unless cancellation occurs along shared boundaries. Further examples are given in [17, Section 4].

2.3 Characterization of Visible Singularities

Having defined the notion of a wavefront set, we now turn our attention to characterizing how an operator transforms the wavefront set. This is one of the central questions in microlocal analysis and it turns out that useful results can be obtained for certain classes of operators.

Let M and N be smooth manifolds and consider an operator $P \colon \mathcal{D}'(M) \to \mathcal{D}'(N)$. If P is a differential operator, then it one can show that it does not increase the wavefront set. This is a special case of a more general result that shows the same result when P is a pseudodifferential operator [17, Theorem 14]:

$$\mathrm{sing\,supp}(P(f)) \subset \mathrm{sing\,supp}(f) \quad \text{and} \quad \mathrm{WF}\big(P(f)\big) \subset \mathrm{WF}(f)$$

with equality if P is in addition elliptic, i.e.,

$$\text{sing supp}(P(f)) = \text{sing supp}(f) \quad \text{and} \quad \text{WF}(P(f)) = \text{WF}(f).$$

Hence, a pseudodifferential operator may spread the support of a function f, but it does *not spread* its singular support or *wavefront set*. This is referred to as the *pseudolocal property* and it has important consequences for inverse problems. In an inverse problem one often seeks to (approximately) invert a pseudodifferential operator P by another operator Q with an integral representation that is similar to P. The pseudolocal property would then imply that the singularities of QPf are identical to those of f, so this approximate inversion can be designed to recover the singularities of f.

Fourier integral operators (FIOs) is an even wider class of operators that contain the pseudodifferential operators as a special case. FIOs are linear and defined in terms of amplitudes and phase functions, see [17, Definition 7] for the precise definition. They arise naturally when one seeks to represent solutions of PDEs and in inverse problems, e.g., FIOs model the solution operators of hyperbolic PDEs as well as a variety of integral transforms, like the Radon transform and its generalizations. One of the main results of microlocal analysis is the Hörmander-Sato Lemma [16, Theorem 5.4], which states that if $P \colon \mathcal{D}'(M) \to \mathcal{D}'(N)$ is a FIO, then

$$\text{WF}(P(f)) \subset C \circ \text{WF}(f) \quad \text{for } f \in \mathcal{E}'(M). \tag{2.4}$$

Here, $C \subset T^*(N) \times T^*(M)$ is the (microlocal) canonical relation associated to P, see [17, eq. (52)] for its formal definition and [17, Definition 8] for how to interpret $C \circ D$ for $D \subset T^*(M)$. In (2.6) we write out the canonical relation for the Radon transform, which is an example of a FIO that is not a pseudodifferential operator.

2.4 Applications to Tomographic Imaging

Data in planar X-ray transmission tomography can in a simplified setting be modelled as samples of the (classical) 2D Radon transform. This is an integral transform that maps a function/distribution $f \in \mathcal{D}'(\mathbb{R}^2)$ to a function/distribution g on lines in $\mathbb{R}^2$. A more explicit definition requires introducing coordinates on the manifold of lines in $\mathbb{R}^2$. Any line $\ell \subset \mathbb{R}^2$ can be described uniquely by $(\theta, p) \in [0, \pi] \times \mathbb{R}$ through

$$\ell = \left\{ x \in \mathbb{R}^2 : x = s \mapsto p\omega(\theta) + s\omega(\theta)^\perp \right\} \quad \text{where} \quad \begin{cases} \omega(\theta) := (\cos\theta, \sin\theta) \in S^1 \\ \omega(\theta)^\perp := (-\sin\theta, \cos\theta). \end{cases}$$

Hence, $(\theta, p) \in [0, \pi] \times \mathbb{R}$ serves as coordinates for a sub-manifold $\mathbb{M}$ of lines in $\mathbb{R}^2$, i.e., $\mathbb{M} \subset [0, \pi] \times \mathbb{R}$. Using these coordinates, we can express the Radon transform of $f \in \mathcal{D}'(\mathbb{R}^2)$ as

$$g = R(f)(\theta, p) := \int_{-\infty}^{\infty} f\left(p\omega(\theta) + s\omega(\theta)^\perp\right) ds \quad \text{for } (\theta, p) \in \mathbb{M}.$$

One can show that R is a continuous map from $L^1(\mathbb{R}^2)$ to $L^1(\mathbb{M})$. There is a rich mathematical theory that investigates properties of such transforms. Particular focus is on injectivity, deriving explicit inversion formulas, and characterizing the range, see [26] for a nice survey.

A natural task in tomographic imaging is to reconstruct f from $g = R(f)$, preferably by using a stable recovery scheme. A closely related task is to recover the wavefront set of f from the wavefront set of data $g \in \mathcal{D}'(\mathbb{M})$, which is a subset of $T^*(\mathbb{M})$. This involves taking the Fourier transform of a localized version of g. To do that, we make use of the aforementioned coordinates on $\mathbb{M}$. One can extend g periodically in θ and take localizing functions ψ with support in θ less than a period, which allows us to view ψg as a function on $\mathbb{R}^2$. Hence, its two-dimensional Fourier transform can be calculated using these coordinates. Next, let $d\theta$ and dp be the standard basis of $T^*_{(\theta, p)}(\mathbb{M})$ where the basis covector $d\theta$ is the dual covector to $\partial/\partial\theta$ and dp is the dual covector to $\partial/\partial p$. One can also view $(x; rdx)$ as the vector $(x; r)$ where $x \in \Omega$ and r is a tangent vector at x. In a similar way, $(\theta, s; ad\theta + bdp)$ can be viewed as the vector $(\theta, p; a, b)$.

The next theorem characterises those singularities of $f \in \mathcal{D}'(\mathbb{R}^2)$ that can be recovered from Randon transform data $g = R(f)$, see [17, Corollary 1] for its proof.

Theorem 2.5. *Let $f \in \mathcal{D}'(\mathbb{R}^2)$ and $g = R(f)$. Then*

$$(x_0, \xi_0 dx) \in \mathrm{WF}(f) \quad \Longrightarrow \quad \Big(\theta_0, x_0 \cdot \omega(\theta_0); s\big(-x_0 \cdot \omega(\theta_0)^\perp d\theta + dp\big)\Big) \in \mathrm{WF}(g)$$

whenever $\theta_0 \in [0, \pi]$ and $s \neq 0$ are such that $\xi_0 = s\omega(\theta_0)$. Likewise, if $(\theta_0, p_0) \in [0, 2\pi] \times \mathbb{R}$ and $q \in \mathbb{R}$, then

$$\Big(\theta_0, p_0; s(-q d\theta + dp)\Big) \in \mathrm{WF}(g) \quad \Longrightarrow \quad (x_0, \xi_0 dx) \in \mathrm{WF}(f)$$

whenever $x_0 := p_0 \omega(\theta_0) + q\omega(\theta_0)^\perp$ and $\xi_0 := s\omega(\theta_0)$.

A key step in the proof is to show that the Radon transform is an elliptic Fourier integral operator. One can then refer to a stronger version of the Hörmander-Sato Lemma in (2.4) for elliptic FIOs [35] and conclude that

$$\mathrm{WF}\big(R(f)\big) = C \circ \mathrm{WF}(f) \quad \text{whenever } f \in \mathcal{D}'(\mathbb{R}^2). \tag{2.5}$$

Here, C is the associated canonical relation and the next step is to explicitly calculate it for the Radon transform [17, Theorem 17]:

$$C = \Big\{ \big(\theta, p, s(-x \cdot \omega(\theta)^\perp d\theta + dp); x, s\omega(\theta) dx\big) \in T^*(\mathbb{M}) :$$

$$(\theta, p) \in \mathbb{M}, x \in \mathbb{R}^2, s \neq 0, x \cdot \omega(\theta) = p \Big\}. \tag{2.6}$$

The claims in Theorem 2.5 now follows from combining (2.5) and (2.6).

Theorem 2.5 implies in particular that the Radon transform R detects singularities of f perpendicular to a line included in the manifold $\mathbb{M}$. These singularities a referred to as visible, whereas singularities of f in other directions do not show up in data $g = R(f)$ (invisible singularities). In particular, if f is a sum of characteristic functions of sets with smooth boundaries, then the tangent line to any point on the boundary of a region is normal to the wavefront direction of f at that point. Hence, a singularity of f at a boundary point x is detectable if there is a line in $\mathbb{M}$ that is tangent to x.

The above analysis for the Radon transform can also be applied to other operators relevant for tomographic reconstruction. One is the backprojection R^* which maps data to image space by integrating $g \in L^1(\mathbb{M})$ over all lines in $\mathbb{M}$ that goes through x:

$$R^*(g)(x) := \int_0^{2\pi} g\big(\theta, x \cdot \omega(\theta)\big) \, d\theta \quad \text{for } g \in L^1(\mathbb{M}).$$

Clearly, R^* maps data to an image but it is not an inverse or R. However, applying R^* to data $g = R(f)$ is the same as $R^*R(f)$, and it turns out that R^*R is an elliptic pseudodifferential operator of order -1 [17, Theorem 13 and eq. (42)]. Hence, by the general theory of microlocal analysis, the backprojection may spread out the support of the function f, but it does not spread out its singular support or wavefront set.

Another related operator that maps the data to an image is the filtered backprojection (FBP) operator:

$$\mathrm{FBP}(g) := R^*(k \circledast g) \quad \text{where } k = R^*(K).$$

The '$\circledast$' above is a 1-dimensional convolution along p-variable and if $g = Rf$, then $\mathrm{FBP}(g) = f * K$ where '$*$' is the 2-dimensional convolution. The idea in FBP is now to choose the reconstruction kernel k such that $K \approx \delta$. The FBP operator is, just like the backprojection, a pseudodifferential operator [17, Example 8], so it could spread the support of the function f, but it does not spread its singular support or wavefront set.

We conclude with mentioning the lambda-reconstruction operator [17, eq. (24)]. This operator is similar to FBP, but instead of recovering f it aims to to recovers its singular support and wavefront set. The main advantage of the lambda-reconstruction operator is that it is local, so one only needs to include lines through a neighbourhood of a point in order to recover the behaviour of f at that point. The lambda-reconstruction operator is an elliptic pseudodifferential operator of order one, so it recovers the singular support and wavefront set f.

3 Shearlets and Wavefront Sets

Due to a lack of a directional component, a (continuous) wavelet system is isotropic and hence not able to resolve the wavefront set of a distribution, in the sense of detecting it by decay properties of the wavelet coefficients of the distribution. Shearlet systems were introduced in as an anisotropic representation system with the structure of an affine system, consisting of a scaling operator to change the resolution, a translation operator to change the position, and a shearing operator to change the orientation, applied to a "mother shearlet". This construction indeed allows precise resolution of wavefront sets as we will discuss (see also [18]).

3.1 Continuous Case

For the definition of a continuous shearlet system, let A_a, $a \in \mathbb{R}^* := \mathbb{R} \setminus \{0\}$ be a *parabolic scaling matrix* A_a, $a \in \mathbb{R}^* := \mathbb{R} \setminus \{0\}$ and S_s, $s \in \mathbb{R}$ be a *shearing matrix* given by

$$A_a = \begin{pmatrix} a & 0 \\ 0 & |a|^{1/2} \end{pmatrix} \quad \text{and} \quad S_s = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}.$$

Moreover, for $M \in \mathbb{R}^{2 \times 2}$, let the *dilation operator* be defined by

$$D_M : L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2), \quad (D_M f)(x) \mapsto |\det(M)|^{-1/2} f(M^{-1} x).$$

Choosing M in the dilation operator as A_a and S_s yields the set of scaling and shearing operators, respectively. Let us also emphasize that the choice of shearing instead of rotation is key to allowing a faithful digitalization due to the fact that the discrete versions S_k, $k \in \mathbb{Z}$, leave the digital grid $\mathbb{Z}^2$ invariant. Finally, let T_t, $t \in \mathbb{R}^2$, denote the *translation operator* as defined by

$$T_t : L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2), \quad (T_t f)(x) \mapsto f(x - t).$$

This now leads to the following definition of continuous shearlet systems.

Definition 3.1. *For $\psi \in L^2(\mathbb{R}^2)$, the* continuous shearlet system $\mathcal{SH}(\psi)$ *is defined by*

$$\mathcal{SH}(\psi) = \{\psi_{a,s,t} := T_t D_{S_s} D_{A_a} \psi = a^{-\frac{3}{4}} \psi(A_a^{-1} S_s^{-1}(\cdot - t)) : a \in \mathbb{R}^*, s \in \mathbb{R}, t \in \mathbb{R}^2\}.$$

Group representation theory leads to conditions under which the associated shearlet transform is even an isometry. For this, let $\mathbb{S} := \mathbb{R}^* \times \mathbb{R} \times \mathbb{R}^2$ be endowed with the group operation

$$(a, s, t) \circ (a', s', t') = (aa', s + \sqrt{|a|} s', t + S_s A_a t').$$

This is a locally compact group with left Haar measure $d_\mu(a, s, t) = da/|a|^3 ds dt$.

Theorem 3.2 ([10]). *Let $\psi \in L^2(\mathbb{R}^2)$ be admissible, i.e., it satisfies*

$$\int_{\mathbb{R}^2} \frac{|\hat{\psi}(\xi)|^2}{|\xi_1|^2} d\xi < \infty.$$

Then the continuous shearlet transform $SH_\psi : L^2(\mathbb{R}^2) \to L^2(\mathbb{S})$ *given by*

$$SH_\psi f(a, s, t) = \langle f, \psi_{a,s,t} \psi \rangle.$$

is an isometry.

One example for a suitable function $\psi \in L^2(\mathbb{R}^2)$ are *classical shearlets*, which are defined by

$$\hat{\psi}(\xi) = \hat{\psi}(\xi_1, \xi_2) = \hat{\psi}_1(\xi_1)\, \hat{\psi}_2(\tfrac{\xi_2}{\xi_1}).$$

Here, $\psi_1 \in L^2(\mathbb{R})$ is a wavelet, i.e., it satisfies the discrete Calderón condition given by

$$\sum_{j \in \mathbb{Z}} |\hat{\psi}_1(2^{-j}\xi)|^2 = 1 \quad \text{for a.e. } \xi \in \mathbb{R},$$

with $\hat{\psi}_1 \in C^\infty(\mathbb{R})$ and supp $\hat{\psi}_1 \subseteq [-\frac{5}{4}, -\frac{1}{4}] \cup [\frac{1}{4}, \frac{5}{4}]$, and $\psi_2 \in L^2(\mathbb{R})$ is a 'bump function', namely

$$\sum_{k=-1}^{1} |\hat{\psi}_2(\xi + k)|^2 = 1 \quad \text{for a.e. } \xi \in [-1, 1],$$

satisfying $\hat{\psi}_2 \in C^\infty(\mathbb{R})$ and supp $\hat{\psi}_2 \subseteq [-1, 1]$.

Since the just defined continuous shearlet system exhibits a directional bias and thus is in this pure form not able to resolve any wavefront set, we require s slightly adapted version. This is based on a suitable splitting of the frequency domain into four conic regions and a low frequency part as illustrated in Figure 1. This leads to the so-called cone-adapted continuous shearlet system.
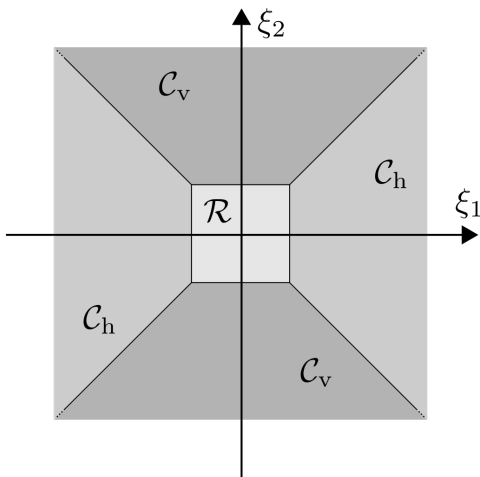

Figure 1: Frequency cones for the cone-adapted shearlet system.

Definition 3.3. *For $\phi, \psi, \tilde{\psi} \in L^2(\mathbb{R}^2)$, the* cone-adapted continuous shearlet system *is defined by $\mathcal{SH}(\phi, \psi, \tilde{\psi}) = \Phi(\phi) \cup \Psi(\psi) \cup \tilde{\Psi}(\tilde{\psi})$, where*

$$
\begin{aligned}
\Phi(\phi) &= \{\phi_t := \phi(\cdot - t) : t \in \mathbb{R}^2\}, \\
\Psi(\psi) &= \{\psi_{a,s,t} = a^{-\frac{3}{4}} \psi(A_a^{-1} S_s^{-1}(\cdot - t)) : a \in (0, 1], |s| \leq 1 + a^{1/2}, t \in \mathbb{R}^2\}, \\
\tilde{\Psi}(\tilde{\psi}) &= \{\tilde{\psi}_{a,s,t} := a^{-\frac{3}{4}} \tilde{\psi}(\tilde{A}_a^{-1} S_s^{-T}(\cdot - t)) : a \in (0, 1], |s| \leq 1 + a^{1/2}, t \in \mathbb{R}^2\},
\end{aligned}
$$

and $\tilde{A}_a = diag(a^{1/2}, a)$.

Similar to continuous shearlet systems as defined in Definition 3.1, also for cone-adapted continuous shearlet systems an associated transform can be defined, namely

$$SH_\psi f(a, s, t) := \langle f, \psi_{a,s,t}\psi \rangle \quad \text{and} \quad SH_{\tilde{\psi}} f(a, s, t) := \langle f, \tilde{\psi}_{a,s,t}\psi \rangle,$$

and isometry conditions can be proven [18]. With this system, we are also now able to precisely resolve a wavefront set of a distribution in the following way.

Theorem 3.4 (Theorem 5.1. of [18]). *Let $\psi \in L^2(\mathbb{R}^2)$ be admissible, and $f \in L^2(\mathbb{R}^2)$. Let $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$, where $\mathcal{D}_1 = \{(t_0, s_0) \in \mathbb{R}^2 \times [-1, 1] :$ for (s, t) in a neighborhood U of (s_0, t_0), $|SH_\psi f(a, s, t)| = O(a^k)$ as $a \to 0$, for all $k \in \mathbb{N}$, with the $O(\cdot)$–term uniform over $(s, t) \in U\}$ and $\mathcal{D}_2 = \{(t_0, s_0) \in \mathbb{R}^2 \times [1, \infty) :$ for $(\frac{1}{s}, t)$ in a neighborhood U of (s_0, t_0), $|SH_{\tilde\psi} f(a, s, t)| = O(a^k)$ as $a \to 0$, for all $k \in \mathbb{N}$, with the $O(\cdot)$–term uniform over $(\frac{1}{s}, t) \in U\}$. Then*

$$WF(f)^c = \mathcal{D}.$$

3.2 Digital Case as a Semantic Edge Detection Problem

Since we are now familiar with the approach to extract the wavefront set of a continuous distribution by analyzing the asymptotic behavior of the shearlet coefficients at a fixed position-orientation pair, we now aim to extend it to real world data. To be precise, we will now consider data coming from a finitely sampled function such as images formed by pixels representing point samples of a real-valued function.

In the most general case, we showed in [3] that the previously discussed approach is not directly transferable to the situation of functions being defined on a grid on a bounded domain. Coarsely speaking, the reason for this is the fact that we just have access to a finite number of Fourier samples as well as finite number of shearlet coefficients.

In order to overcome this limitation, we are now assuming that a digital image arises from the finite sampling of a continuous model in that sense that the image has itself a wavefront set in the sampling limit. As in our paper [3], we aim to approximate the wavefront set by a sequence of what we call *the digital wavefront sets*.

Let us now define what we mean by a digital wavefront set (see also [3]). In the sense of Shannon's sampling theorem one can make use of Paley-Wiener spaces to define a sampling space of $L^2(\mathbb{R}^2)$ for the coarsest scale $\Lambda > 0$, namely $\mathcal{PW}_\Lambda \subset L^2(\mathbb{R}^2)$ defined by

$$\mathcal{PW}_\Lambda := \left\{ f \in L^2\left(\mathbb{R}^2\right) : \operatorname{supp}\left(\widehat{f}\right) \subset [-\Lambda, \Lambda]^d \right\}.$$

Using this definition of the sampled space, we now define the notion of a *digital wavefront set extractor* for the given coarsest scale $\Lambda > 0$ to be the map

$$\mathrm{DWF}_\Lambda : \mathcal{PW}_\Lambda \to P\left(\mathbb{R}^2 \times \mathbb{S}^1\right) \quad \text{such that} \quad \mathrm{DWF}(P_\Lambda f) = \mathrm{WF}(f) \text{ for all } f \in L^2(\mathbb{R}^2). \tag{3.1}$$

The existence of a so-called *faithful sequence of digital wavefront set extractors* assumes that the sequence of maps $\{\mathrm{DWF}_j\}_{j \in \mathbb{N}}$ converges to the continuous wavefront set extractor WF in the Hausdorff sense, i.e., we have

$$d_H\left(\overline{\mathrm{DWF}_j(P_j(f))_x}, \overline{\mathrm{WF}(f)_x}\right) \to 0. \tag{3.2}$$

In [3], we make use of Shannon's sampling theorem to show that it is not possible to analytically define a sequence of digital wavefront set extractors for general function classes, which are dense on $L^2(\mathbb{R}^2)$ functions (see [3, Thm. 3.3]). In this work, we also indicated that, although for most classical function classes it is impossible to construct an analytical digital wavefront set extractor on their Paley-Wiener space projections, for function classes in applications such as natural images, which are typically empirically defined, such wavefront set extractor could exist. However, those will certainly be highly sensitive to the choice of the class.

Inspired by these results, it is natural to have as *guiding principle* of the construction of such digital wavefront set extractors, in our case the wavefront set extractor should be closely adapted to the underlying function class.

As mentioned, typical function classes arising from real-world applications are empirically defined. Thus most of model-based heuristics that could potentially be used to construct analytically a digital wavefront set extractor for each class, will fail with high probability.

Using solely, model-based heuristics will limit the model with rigid assumptions and the guiding principle wont be followed. Although the best choice for our guiding principle, will be to learn the data representation

from scratch this approach will typically be computationally intractable, requiring a lot of data and complicated architecture, having as example, CASENet [38] and SEAL [39].

Summarizing, our approach presented in [3] follows the main thrust of current deep learning-based methodologies in imaging science. It leads to a state-of-the-art wavefront set extraction by combining the model-based digital shearlet transform as a pre-processor for edge detection and a convolutional neural network that locally learns the wavefront set by classifying patches of the digital shearlet coefficients with the potential singular point at the center.

3.2.1   Digital Shearlets

Let us now delve more deeply into the digital shearlet transform [21] for a digital domain of pixel images, which — as explained before — is used in the classifier proposed in [3]. The construction of the digital shearlets is inspired by the fast wavelet transform. In fact, the digital function system can be seen as a filter bank, for which the novelty in the shearlet case resides in the definition of a faithful digital shearing operator.

To explain the digital shearlet transform in detail, let $M \in \mathbb{N}$, $J \subset \mathbb{N}$ be finite, $k_j \subset \mathbb{N}$ for all $j \in J$ and $K_j := [-k_j, \ldots, 0, \ldots, k_j]$. We then pick $2 \sum_{j \in J} K_J + 1$ matrices in $\mathbb{R}^{M \times M}$, and denote these matrices by ϕ^{dig} and $\psi^{dig}_{j,k,\iota}$ for $j \in J, k \in K_j, \iota \in \{-1, 1\}$. To make the connection to the continuous shearlet transform, we can think of $\psi^{dig}_{j,k,\iota}$ as a digitized version of $\psi_{2^{-j}, 2^{-j/2}k, 0, \iota}$ and of ϕ^{dig} as a digitized version of a low frequency filter. An explicit — and highly technical — construction of the matrices ϕ^{dig} and $\psi^{dig}_{j,k,\iota}$ can be found in [21]. Those are then exploited to define the *digital shearlet transform* of an image $I \in \mathbb{R}^{M \times M}$ by

$$
\mathrm{DSH}(I)(j, k, m, \iota) := \begin{cases} \langle I, T_m \psi^{dig}_{j,k,\iota} \rangle, & \text{if } \iota \in \{-1, 1\}, \\ \langle I, T_m \phi^{dig} \rangle, & \text{if } \iota = 0, \end{cases}
$$

where $j \in J, k \in K_j$, $m \in \{1, \ldots, M\}^2$, and $T_m : \mathbb{R}^{M \times M} \to \mathbb{R}^{M \times M}$ circularly shifts the entries of the elements of a matrix by m. Thus, from a structural viewpoint, the digital shearlet transform of an image $I \in \mathbb{R}^{M \times M}$ is a stack of $2 \sum_{j \in J} (K_j - 1) + 1$ matrices of dimension $M \times M$. We will refer to this stack of images, as the shearlet volume.

3.2.2   DeNSE Algorithm

Using the digital shearlet transform as the input of a convolutional neural network classifier, the Deep Network shearlet Edge Extractor (DeNSE) is able to perform wavefront set extraction with high accuracy and closely adapted to the class of the training data. In addition to follow the guiding principle, the main purpose of this approach was to achieve high accuracy concerning the task of wavefront set extraction for inverse problem regularization. With this in mind, the problem of classifying pixels of a digital image into $N + 1$-classes, where N of these classes correspond to orientations of singular points and the additional ($N + 1$th) class is the binary class of a pixel being an edge, is splitted into $N + 1$ separate binary classifiers. In addition, each pixel is classified independently as well by using a patch-based approach in the sense that a pixel will be classified to be an edge of a certain orientation by classifying a patch of the shearlet coefficients where the pixel is centered at.

The architecture used for each classification was composed of four convolutional layers, with 2×2 max pooling, ReLU activation, and batch normalization, followed by a fully connected layer with 1024 neurons, softmax activation function, and a one dimensional output. We chose this architecture since it performed well in a series of tests while being of moderate size. The network architecture is illustrated in Figure 2.

As in [3], if we pick $J = 4$ as the coarsest scale, the digital shearlet coefficients $DSH(I)$ will form a three dimensional array composed by 49 stacked digital images of the same size as the original image. Having these coefficients, if the patches to classify are as in the original work, of size 21×21, the input of DeNSE will be a stacked set of patches of size $21 \times 21 \times 49$.
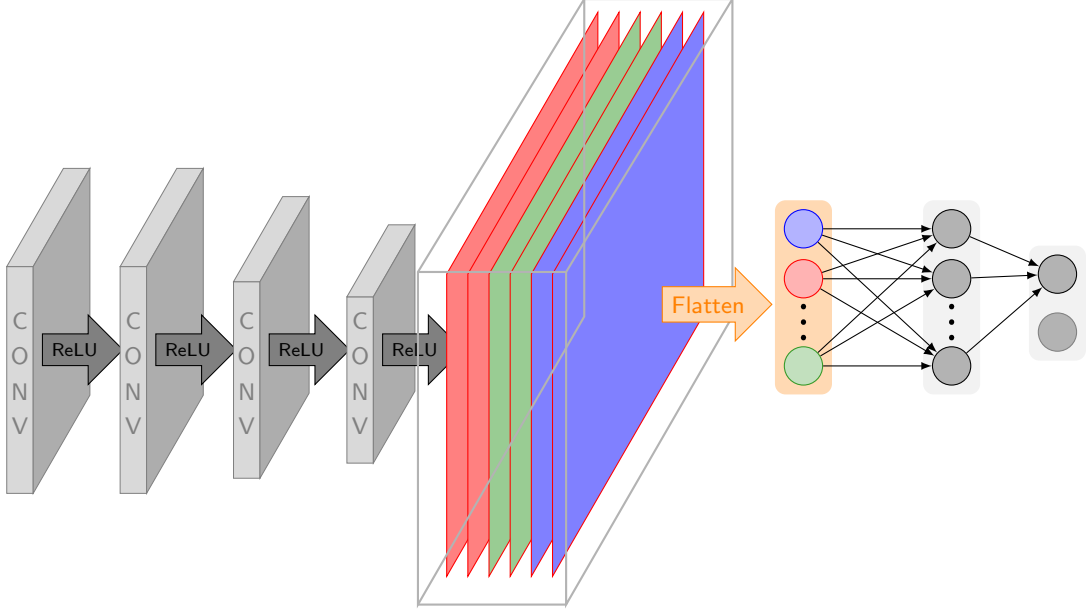
14

Figure 2: Illustration of the network architecture forming the foundation of the classifier. This network consists of four convolutional layers and one fully-connected layer. The colored block in the middle represents a stack of the output of the last convolutional layer. The colors correspond to the different channels.

If for example, we have 180 orientations $\{\theta_i\}_{i=1}^{180}$, for each θ_i, the network is trained on patches of shearlet coefficients of images $I \in \mathbb{R}^{M \times M}$ of the form

$$(\mathrm{DSH}(I)(j,k,m,\iota))_{j \in J, k \in K_j, \iota \in \{-1,0,1\}, m \in [m_1^*-10, m_1^*+10] \times [m_2^*-10, m_2^*+10]}, \tag{3.3}$$

with $m^* \in \{11, \ldots, M-10\}^2$ being the center point of the patch to which the classified orientation is assigned. By using a softmax binary classifier, the label associated to a batch of (3.3) is 1 if I has a singularity with orientation θ_i at m^*. In addition, there exists a separate classifier, which uses the same data and assigns 1 to a patch whose central pixel corresponds to a singular point (edge point).

The DeNSE algorithm for wavefront set extraction [3] uses as coarsest scale $J = 4$ with the corresponding shearing levels for each scale given by $K_j = 2^{\lceil j/2+1 \rceil} + 1$, obtaining digital shearlet coefficients arrange by L stacked images, where $L = 2 \sum_{j \in J}(K_j - 1) + 1 = 49$. The implementation of choice was the julia API of the software ShearLab [20] (`www.shearlab.org/software`).

The Deep Network shearlet Edge Extractor (DeNSE) was in fact shown to outperform other methods, for example, CASENet and SEAL, on the standard datasets such as BSDS500 (Berkeley segmentation dataset) with 503 natural images, the semantic boundaries dataset (SBD) with 11355 natural images, and a set of phantoms formed by ellipses and paralellograms with analytically defined wavefront set. We present some of those results in Section 7.

3.3 Avoiding the Distracted Supervision Paradox

The distracted supervision paradox as introduced in Section 1.2 refers to the fact that semantic edge detection such as wavefront set extraction requires the supervision of two fundamentally different tasks:

- **Category-agnostic edge detection** requires the detection of pixels corresponding to edges and other singularities, thus the use of low-level features.

- **Semantic edge classification** requires the classification of edges with abstracted high-level semantics, therefore relies on high-level features.

15

Intuitively accomplishing both tasks jointly seems infeasible as both problems require very different features. Yu et al. confirmed in [38] with the CASENet architecture that a naive joint supervision of both tasks, performs less well than directly learning the semantic edges with no deep supervision that combines both features.

This paradox imposes an upper bound on the performance for methods that are aiming to directly learn (in an end-to-end fashion) semantic edges. Since the initial work [38], there have been several approaches to avoid this paradox. One particularly remarkable approach is the one presented by Liu et al. in [24]. In their paper, they propose a network architecture containing a backbone based on residual convolutional neural networks — similar to the CASENet, but with the introduction of a novel information converter layers —, which allows to combine information coming from lower levels used for edge supervision with information from higher levels used for semantic supervision, which is ultimately guided by the detected edges. This approach successfully established a new state-of-the-art in semantic edge detection, with a significant performance improvement.

The key to the Deep Network Shearlet Edge Extractor (DeNSE) resides on the splitting of the multi-label classification task into individual binary classifiers inspired on the performance increment. In addition DeNSE separates the category-agnostic edge detection and the semantic edge classification, which already avoids the distracted supervision paradox.

4 General Semantic Edge Detection Using Shearlets and Deep Supervision

The use of shearlets for achieving high precision in digital wavefront set extraction motivates the introduction of the shearlet transform in general semantic edge detection, where classes of edges are coming typically from the particular object the edge belongs to. In order to illustrate the power of the shearlet transform as a preprocessing step, we are going to make use of the backbone of the current state-of-the-art approach on general semantic edge detection, based on the CASENet architecture [38]. As already mentioned, after the introduction of CASENet and the distracted supervision paradox, most of the approaches to perform semantic edge detection, made use of the same architecture backbone. Each model, introduced an alternative way to train it, mostly to avoid the distracted supervision paradox.

At the following we explore two alternatives architectures, namely the original CASENet and the deep diverse supervision approach presented by Liu et al in [24], where information converters are used for the joint training of low-level edge supervision and high-level edge classification. In order to show the heavy lifting that the shearlet transform is able to do, we use as the input for these networks the shearlet coefficients of the images, and extend the channels of the first convolutional layer by the number of slices of the particular shearlet coefficients. We also decrease the depth of the resulting network by removing the last category-agnostic edge feature map resnet subnetworks, resulting in an overall smaller network, which will be addressed in detail in the next sections.

5 Shear-CASENet: Deep Shearlet Category-Aware Semantic Edge Detection

The CASENet architecture [38] is based on the already known Residual Neural Network architecture, known as ResNet (see Figure 3). This architecture has shown tremendous success in different image processing task, including image classification. It in fact did even won the ImageNet challenge in 2015. CASENet receives as input the image and it produces as output a two-dimensional array of the same size with the classified edges represented as pixels with value given by the corresponding category.

We next explain the CASENet architecture, which is displayed in Figure 4, in more detail. The input image is connected to a $1 - channel$ convolutional layer (conv1), which is followed by four stacked ResNet subnetworks; res2c, res3b, res4b22 and res5c correspondingly. Each of those sub-networks is a block of the
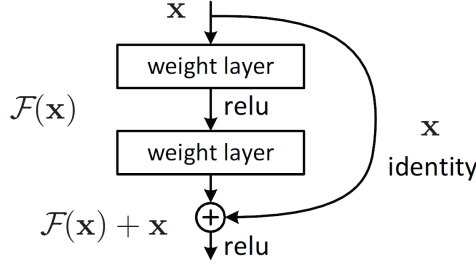
Figure 3: Illustration of the principal block in ResNet, namely the skip connection from the input to the output is the main characteristic of this architecture (taken from `https://neurohive.io/en/`).

network ResNet-101, where res(N)(M) represents the M-th layer (represented by the letter "a", "b" and "c") of the N-th stage of ResNet-101.

The first three stages of CASENet (i.e. conv1, res2c, res3b) produce a single channel feature map $F^{(m)}$, which is used to perform the edge detection part. The last stage, res5c, is connected to a 1×1 convolutional layer to produce a K-channel class activation map $A^{(5)} = \{A_1^{(5)}, A_2^{(5)}, \ldots, A_K^{(5)}\}$, where K is the total number of categories. In order to combine the edge information coming from the first stages, at the end of the network, one replicates the bottom features $F^{(m)}$, by concatenating them in each channel of the class activation map at the last stage, namely:

$$A^f = \{F^{(1)}, F^{(2)}, F^{(3)}, A_1^{(5)}, \ldots, F^{(1)}, F^{(2)}, F^{(3)}, A_K^{(5)}\}.$$

At the end, a K-grouped 1×1 convolutional layer is applied to A^f, generating a semantic edge map with K channels, whear the k-th channel represents the edge map for the k-th category. Summarizing, the first four stages of CASENet produce category-agnostic edge feature maps with different levels of refinement. This depth becomes necessary to produce edges fine enough to be classified by the last stage.
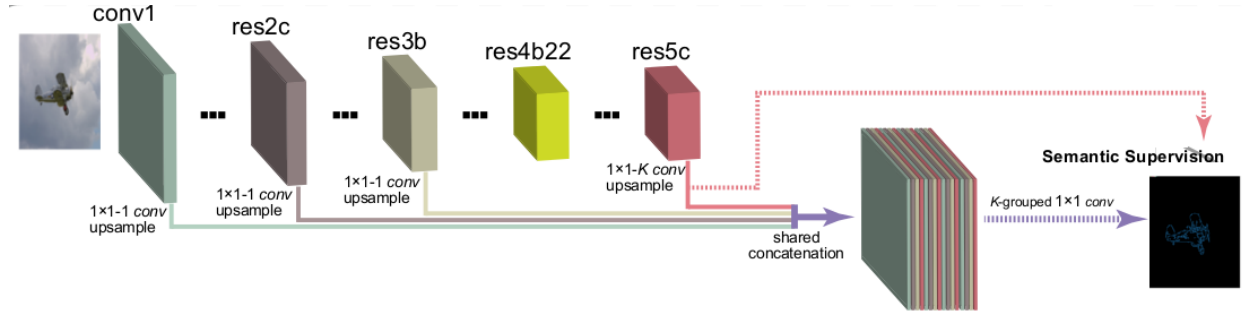

Figure 4: Illustration of the classical CASENet architecture (inspired by the figures of [24]).

Recall that he shearlet transform has shown good performance as a feature extractor for edge detection, due to its properties on wavefront set resolution. With this in mind, it seems conceivable that applying CASENet to the shearlet coefficients instead of the image itself, the produced edge in the first feature maps will be fine enough already in first stages.

By now combining shearlets and the CASENet architecture, we introduce the Shear-CASENet architecture, which takes as input the shearlet coefficients of an image and produces as output an array with the same size as the image with the classified edges, just like CASENet. Shear-CASENet, uses the same backbone as CASENet, extending the first convolutional layer with the same number of channels as slices of the shearlet volume and removing the fourth stage. Figure 5 depicts the Shear-CASENet architecture.

Shear-CASENet has significantly less parameters to train than the classical CASENet. Indeed, if the input is an $N \times N$-image, the fourth stage of CASENet will have $1024 \times N^2/8$ parameters. By having that
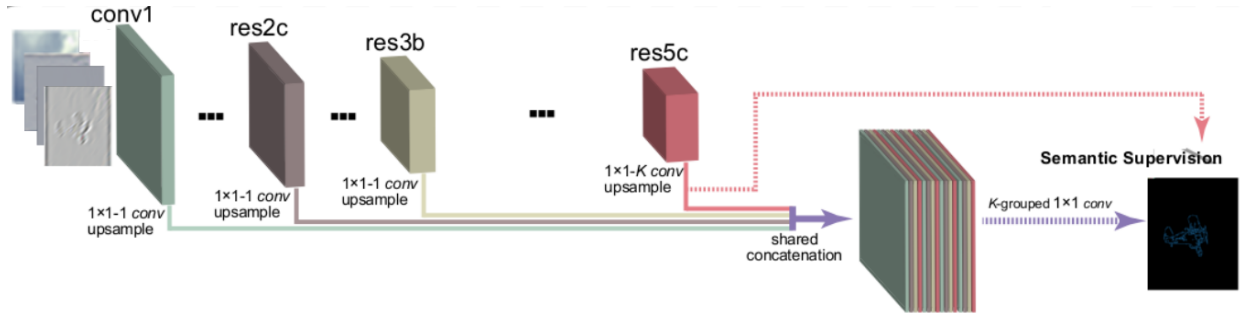
17

Figure 5: Illustration of the Shear-CASENet architecture (inspired in the figures of [24]).

stage removed and extending the first convolutional layer by J shearlet channels, one obtains a reduction of $1024 \times N^2/8 - J$ trainable parameters. Typically four scales in the shearlet transform are used, i.e., $J = 49$, leading to an image of size 256×256. Shear-CASENet has 37259387 parameters while the classical CASENet has 42436731, with around 13% less parameters to train.

In addition, Shear-CASENet presented also better performance in semantic edge detection on the Semantic Boundaries Dataset (`http://home.bharathh.info/pubs/codes/SBD/download.html`). We will present the corresponding numerical experiments in detail in Section 7.2.

6   Shear-DDS: Shearlet Diverse Deep Supervision

In 2017 the CASENet architecture imposed by its novel approach a new state-of-the-art in semantic edge detection. The authors also introduced the concept of distracted supervision paradox, by noticing the fundamental limitation of training jointly, with deep supervision, the category-agnostic edges and the edge classification. In CASENet, even the fourth stage is not used in the supervision. It was in fact introduced as a way to alleviate the information conflicts coming from the first three stages and the fifth stage, also known as a buffer unit.

After the introduction of CASENet, different alternatives to train the network have been introduced. Yu et al. [39] introduced the Simulatenous Edge Alignment and Learning (SEAL) architecture, which is a new training approach for the CASENet architecture. It simultaneously aligns the ground truth edges and learns the corresponding classifier, with the downside of being time consuming due to the necessary CPU usage by the alignment step.

Recently Liu et al. introduced a novel way to train CASENet [24], also known as the deep diverse supervision. This approach makes use of an information converter based on a convolutional residual block (see figure 3), where the output of each stage of CASENet is fused in a final shared concatenation. Figure 6 depicts this architecture, it is worth to notice that in this case, stage four is not anymore a buffer, but it is already used in the supervision.

The information converters help to assist low-level feature learning (stage one-four) in order to generate consistent gradient signals from the higher levels (stage five), producing a highly discriminative feature map for high performance semantic edge detection.

Having the category-agnostic edge maps obtained from the information converter applied to each of the first four stages, namely $E = \{E^{(1)}, E^{(2)}, E^{(3)}, E^{(4)}\}$, the final map will be given by the information conversion of the fifth stage and the shared concatenation, i.e.,

$$E^f = \{E, A_1^{(5)}, E, A_2^{(5)}, \dots E, A_K^{(5)}\}$$

This network is trained with a multi-task loss, meaning, two different losses, corresponding to category-agnostic and category-aware edge detection, are optimized jointly. Both losses are based on reweighted sigmoid cross-entropy loss, which is typically used for multi-label classification. For further details, we refer to [24].
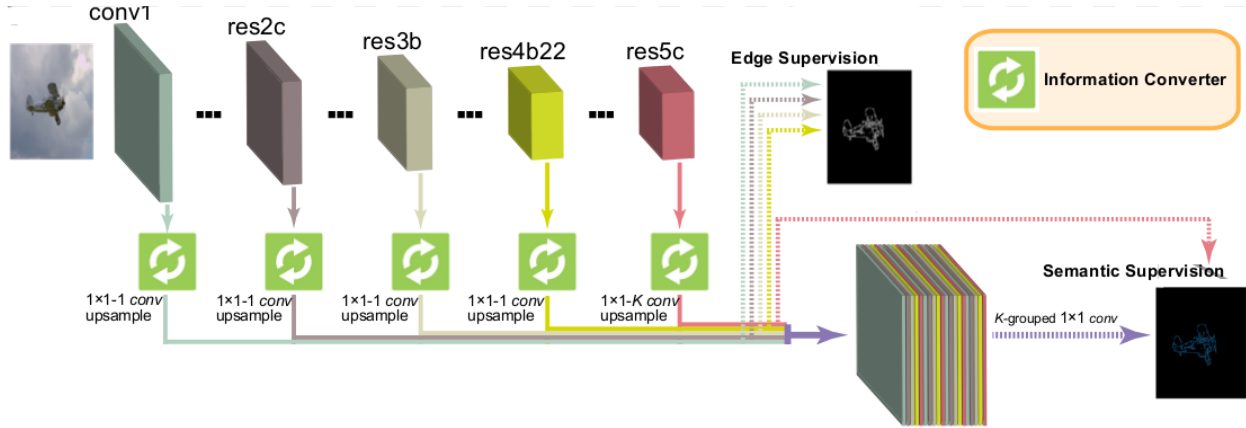
18

Figure 6: Illustration of the classical Diverse Deep Supervision architecture (inspired by the figures of [24]).

Using a similar approach as with the classical CASENet, we introduce as a new architecture the shearlet Diverse Deep Supervision (Shear-DDS). This architecture will accept as input the shearlet volume of an image. The output consists of the same activation map characterizing the classified edges as the original DDS.

As we did with the Shear-CASENet, we reduced the overall number of parameters by removing stage four of the DDS architecture, with a similar 13% reduction of learnable parameters. Figure 7 shows the new Shear-DDS architecture.
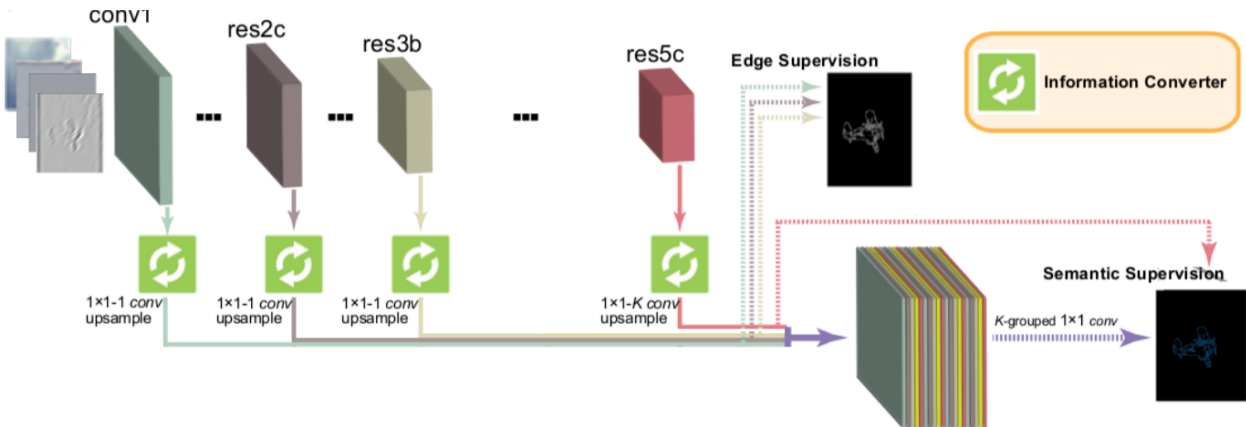
Figure 7: Illustration of the shearlet Diverse Deep Supervision architecture (inspired by the figures of [24]).

Shear-DDS was trained with the same multi-task loss function as the original architectrue, obtaining an improvement in performance on the SBD dataset, the corresponding numerical results are presented in Section 7.2.

7   Numerical Results and Applications

To show the true impact of the approach presented in this work, we performed several numerical experiments, targeted to three specific applications, namely, wavefront set extraction, general semantic edge detection, and Computed Tomography reconstruction. In each of the experiments the networks were trained and evaluated on datasets for the specific purpose. On one hand, the general semantic edge detection application was

19

trained and evaluated on the semantic boundaries dataset (SBD), which is already and standard benchmark for this application. On the other hand, since wavefront set extraction is well suited to be used in Computed Tomography reconstruction (see Section 2.4), the DeNSE sub-networks were trained on images formed by random ellipses which resemble human-head phantoms such as the well-known Shepp-Logan phantom.

All experiments show that our hybrid approach, namely combining shearlets with carefully designed network architectures, provide a significant improvement in performance. This indicates that this conceptual approach should be also beneficial for other image processing tasks.

7.1 Wavefront Set Extraction

For wavefront set extraction, we trained the DeNSE architecture with the procedure presented in [3]. We trained the networks on head-like phantom images, inspired by the Shepp-Logan phantom. Our training data consists of a specific selection of random ellipses, two big ellipses, representing the inner and outer skull and small ellipses inside the skull, with different sizes, pixel intensity value, and orientations. We also vary the gradient of the intensity of the ellipses in order to obtain curves with different levels of regularity.

The advantage of using this type of phantom is two-folded: It allows to have access to the analytical wavefront set and the resulting network can be used for Computed Tomography applications in head-like phantoms.

We used the tensorflow implementation of the DeNSE architecture, publicly available in `http://www. shearlab.org/applications`. For the shearlet transform, we used the python interface of the julia API of shearlab, available in `http://www.shearlab.org/software`.

The DeNSE architecture classifies patches of an image for an specific given orientation. We used a resolution of 180 distinct orientations and trained the network for each orieantion separately. For each training we used 10,000 images. The evaluation and test was then done over 2,000 images each. We made use of a total of four scales on the shearlet transform, producing a shearlet coefficients volume with 49 slices. From the shearlet coefficients of each image, we extracted 10 distinct patches randomly. This was done in such a way that the classes are balanced, meaning, in each binary classifier, the number of positive cases is roughly the same as the number of negative cases.

We then used the standard MF-score to measure the performance of the classifiers, which is the mean of the F-score over all the orientations. We compared the performance with other semantic edge detection models, and for this made use of the publicly available python code given for the CoShREM [30], CASENet [38] and SEAL [39]. For the Yi-Labate-Easley-Krim [37] and the DDS [24] models, we used our own implementation. The performance benchmarks with these models are presented in Table 1.

Figure 8 then shows the results of wavefront set extraction on an example of the head-phantom dataset using three different models, the CoShREM model, and the CASENet architecture. Judging from the images, it is clear that DeNSE shows significantly better performance than CoSHREM, where the latter is not able to find the ellipses with smooth boundaries.

Method	MF-score
Yi-Labate-Easley-Krim[37]	75.7
CoShREM[30]	70.4
CASENet[38]	78.6
SEAL[39]	83.4
DDS[24]	85.6
DeNSE[3]	*95.7*

Table 1: Performance of wavefront set extraction on the head-phantom data set. All values are in percentage.
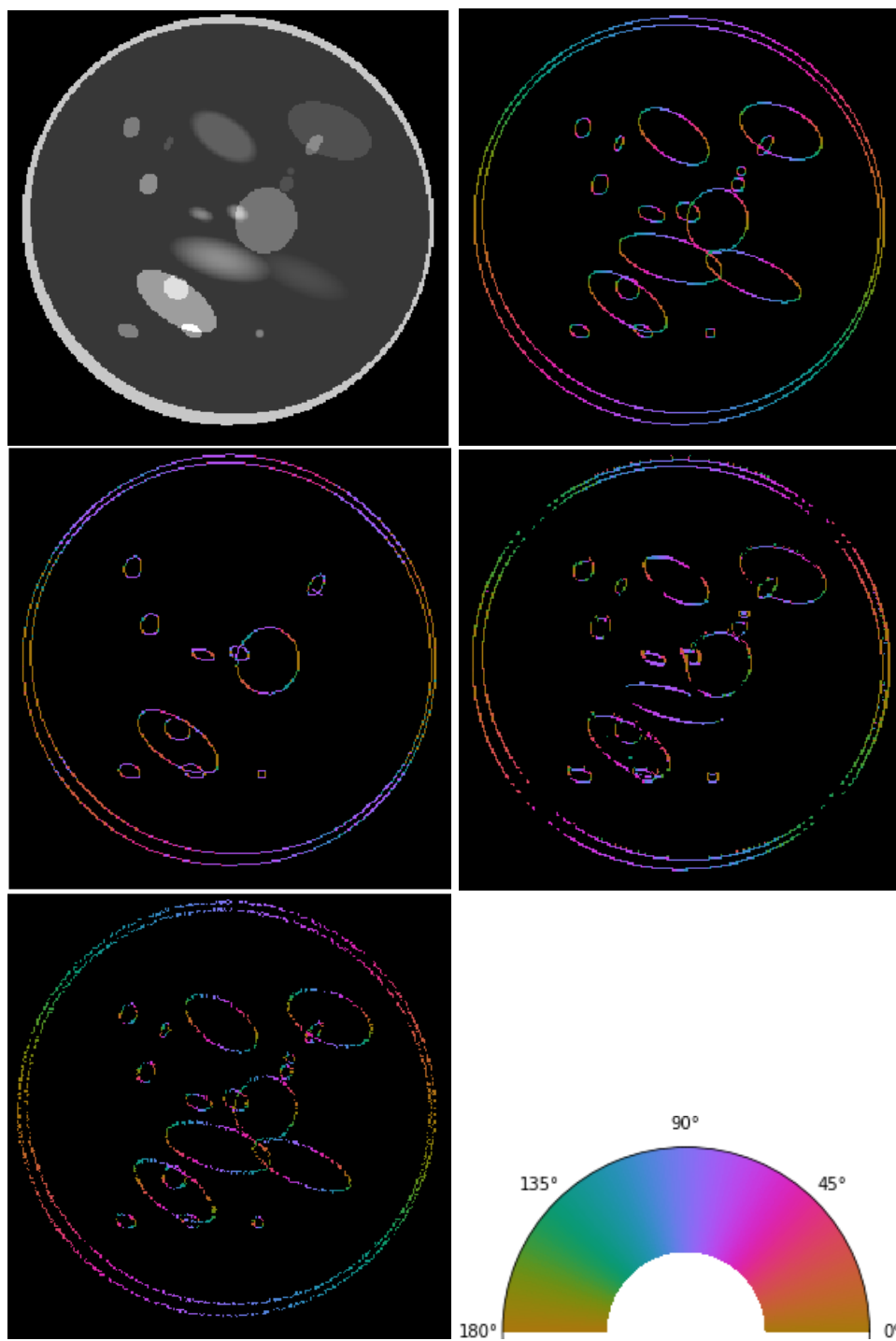
Figure 8: Computed edges and orientations of an example of the head-phantom data set. Top-left: Input image. Top-right: Orientations, analytical ground truth. Middle-left: Orientations predicted by CoShREM algorithm. Middle-right: Orientations predicted by CaSENet. Bottom-left: Orientations predicted by DeNSE algorithm. Bottom-right: Color code for normal-directions.

7.2 General Semantic Edge Detection

For the numerical experiments of the general case of semantic edge detection, we trained the Shear-CASENet and Shear-DDS architectures on the Semantic Boundaries Dataset (SBD). This database consists of 11,355 images, from which we used 9,035 images for training, 1,050 for evaluation, and 1,050 for testing. Each image has a human-annotated array of edge-pixels with the intensity value as the category number of the object, where this edge belongs to. The SBD dataset consists of a total of 20 categories, including vehicles, animals, and plants.

Both Shear-CASENet and Shear-DDS were trained on the full shearlet coefficients. Similar to the case of the wavefront set extraction, we use the digital shearlet transform [21] implemented on julia, with a total of four scales. This produces a shearlet coefficients volume of 49 slices, which was then fed to the proposed architectures.

We use the publicly available implementation of the CASENet architecture (`https://github.com/lijiaman/CASENet`). This implementation makes use of the deep learning framework pytorch, making it compatible with our shearlet implementation. Based on this code, we implemented the deep diverse supervision architecture by introducing the information converters and the proposed multi-task loss. Also based on this code, we implemented the Shear-CASENet and Shear-DDS architectures by extending the first convolutional layer with the corresponding shearlet channels (see Figures 7 and 5) and removing the fourth stage of the original architectures.

In addition to CASENet and DDS, we also compared our methods to deeply supervise version of CASENet [38] and SEAL [39]. The performance benchmarks presented in Table 2 are done in terms of the mean F-score, in a similar fashion as with the wavefront set extraction benchmarks, by computing the mean of the F-score over all the categories. It is visible that the mean-F value is slightly better on the the Shear-CASENet and Shear-DDS than on the other architectures. The improvement is not as significant as in the case of the wavefront set extraction, since DeNSE was specifically designed for this task and the existing models have general semantic edge detection applications. It is though worth to stress that Shear-CASENet and Shear-DDS have significantly less parameters than their non-shearlet counterparts.

In Figure 9, we depict the results obtained using an example of the SBD dataset. It shows the semantic edges obtained by both CASENet and DDS and their respective shearlet extension. In all the cases the airplane in the picture was correctly classified, but the refinement of the obtained edges is improved in the shearlet version. This strongly suggests that the use of shearlets is well-suited for high performance in semantic edge detection.

Method	MF-score
DSN[38]	65.2
SEAL[39]	75.3
Classical CASENet[38]	71.4
Classical DDS[24]	78.6
Shear-CASENet	75.7
Shear-DDS	80.1

Table 2: Semantic edge detection performance on the SBD dataset. All values are in percentage.

7.3 Tomographic Reconstruction

Many inverse problems that arise in tomographic imaging involve recovering object boundaries. It is obvious that determining those boundaries that are visible or not is important. In fact, a close look at the reconstructions reveals that in each case the only feature boundaries that are clearly defined are those tangent to lines in the data set for the problem. Example 1 illustrates this in a simple way: one detects singularities in the Radon data exactly when the lines of integration are tangent to the boundary of the object. The goal of this chapter is to make the idea mathematically rigorous.

As shown in Section 2.4, microlocal analysis can be used to describe how the Radon transform transforms wavefront sets (singularities). To show this in the context of digitized images and data we make use of the digital wavefront set extraction applied to the head-phantom dataset introduced on Section 7.1 and the microlocal canonical relation of the radon transform (Section 2.3), in order to obtain the digital wavefront set an image from the digital wavefront set of its sinogram without a previous inversion. This approach is based on the work presented by the authors in [3]. We simulate tomographic data (sinograms) using the python implementation of the digital Radon transform in the operator discretization library (ODL, http://github.com/odlgroup/odl.

Let us now explain the training procedure. To label each sinogram with the correct wavefront set, we used a digitized version of the canonical relation for the Radon transform, presented in [3, Def. 6.1]. This definition was then taken as an Ansatz for the definition of the digital wavefront set of the sinograms, which came from a phantom whose the wavefront set we know explicitely. We then trained the DeNSE model on the sinogram wavefront sets, with the same training, test, and evaluation set as in Section 7.1. Using the results of the canonical relation as the ground truth for the sinogram wavefront sets, we obtained a MF-score over the test datset of 95.7%, which is comparable to the performance of DeNSE on the head-phantom image class. We will not present any performance comparison, since we are not aware of any competing algorithm for the detection of wavefront set on sinograms.

Since applying the inverse canonical relation to the wavefront set of the sinogram will give us access to part of the wavefront set of the original image, without performing yet any inversion, having a method to detect the wavefront set of a sinogram becomes useful for inverse problem regularization.

To show this potential, we present in Figure 11, the phantom wavefront set obtained when applying the inverse canonical relation to the low-dose sinogram, when measuring every six degrees. This low-dose problem is highly ill-posed, and will require significant effort to be inverted. Nut having at hand the digital canonical relation will give us almost for free the part of the phantom's wavefront set associated with the measured angles.

To further show the use of this method, we performed a the low-dose sinogram example, three standard inversion schemes, filtered backprojection, Tikhonov regularization and total variation (TV) regularization. We first compute the associated image reconstruction from the low-dose sinogram, measured each six angles, and then compute the associated wavefront set of the reconstruction with DeNSE. We next compute the average mean square error to the true wavefront set of the data point. Via the inverse canonical relation, we compare this with the error resulting from computing the wavefront set of the sinogram and mapping it back to the image.

Figure 12 shows the wavefront sets corresponding to the different reconstruction schemes. In Table 3, the obtained errors are presented, clearly showing the advantage of first extracting the wavefront set of the sinogram and then applying the canonical relations, over any first-invert-then-extract strategy. In additio,n this a-priori information can be used as a regularizer on any variational regularization scheme.

Inversion technique	Mean square error
Tikhonov	443.0
Total variation	380.9
Filtered backprojection	504.3
Canonical relations	168.1

Table 3: Error of wavefront set estimation by different inversion techniques.

7.4 Discussion of Numerical Experiments

We observed in the numerical experiments an improvement on edge extraction and classification performance when using the shearlet transform for model-based feature extraction previous to feeding the network. In particular, the DeNSE model for wavefront set extraction has shown a significant improvement with respect of traditional methods designed for wavefront set extraction ([30], [37]), due to the combination of both the

23

model-based shearlet representation and the data-driven high performance classification. On the other hand, it also presents better performance with respect of deep-learning based models ([38], [39], [24]), although is worth to mention that this methods were originally designed for semantic edge detection with a lower variation in the classes over the contiguous pixels. In addition DeNSE, by definition, is able to avoid successfully the distracted supervision paradox, without much effort.

For general semantic edge detection, introducing a change of representation system, to the shearlet system, improved the performance of the standard semantic edge detection models (CASENet and DDS), but not at the same level. This suggests, that these models are not well suited to the problem of wavefront set extraction, and a patch-based strategy will reach to a similar performance leap as in the wavefront set extraction case.

The results also suggest that the use of shearlets in image processing tasks involving edge detection, helps to do heavy-lifting in a model-based fashion, reducing the amount needed parameters, resulting in a reduction of complexity. This is clearly shown by the shear-CASENet and shear-DDS architecture which perform better than their non-shearlet version, but have less learnable parameters. This combination of model-based and data-driven approaches is a strategy that more researchers have adopted in the last few years.

In addition, the applications of semantic edge detection in inverse problems, are not yet explored in depth, but the results presented on Section 7.3, are a clear indicative that this area will be fruitful in the next few years. As part of the future developments, the authors will work on the use for the wavefront set extractor to improve the existing hybrid methods on tomographic reconstruction.

Acknowledgements

References

[1] M. R. Amer, S. Yousefi, R. Raich, and S. Todorovic. Monocular extraction of 2.1 d sketch using constrained convex optimization. *Int. J. Comput. Vis.*, 112(1):23–42, 2015.

[2] H. Andrade-Loarca. Sparse light field reconstruction via shearlet-based inpainting. *Master Thesis TU Berlin*, 2017.

[3] H. Andrade-Loarca, G. Kutyniok, O. Öktem, and P. Petersen. Extraction of digital wavefront sets using applied harmonic analysis and deep neural networks. *SIAM J. Imaging Sci.*, 12(4):1936–1966, 2019.

[4] G. Bertasius, J. Shi, and L. Torresani. Deepedge: A multi-scale bifurcated deep network for top-down contour detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4380–4389, 2015.

[5] T. O. Binford. Inferring surfaces from images. *Artif. Intell.*, 17(1):205–244, 1981.

[6] M. Brady. Computational approaches to image understanding. *ACM Comput. Surv.*, 14(1):3–71, 1982.

[7] E. Candes, L. Demanet, D. L. Donoho, and L. Ying. Fast discrete curvelet transform. *SIAM Multiscale Model. Simul.*, 5:861–899, 2006.

[8] E. Candes and D. L. Donoho. New tight frames of curvelets and optimal representation with C^2 singularities. *Comm. Pure. Appl. Math.*, 57:219–266, 2004.

[9] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, 1986.

[10] S. Dahlke, G. Kutyniok, G. Steidl, and G. Teschke. Shearlet coorbit spaces and associated banach frames. *Applied and Computational Harmonic Analysis*, 27(2):195 – 214, 2009.

[11] V. Ferrari, F. L., F. Jurie, and C. Schmid. Groups of adjacent contour segments for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(1):35–51, 2008.

[12] V. Ferrari, F. L., and C. Schmid. From images to shape models for object detection. *Int. J. Comput. Vis.*, 87(3):284–303, 2010.

[13] K. Guo, G. Kutyniok, and D. Labate. Sparse multidimensional representations using anisotropic dilation and shear operators. In *Wavelets and Splines*, pages 189–201, Nashville, TN, 2005. Nashboro Press,.

[14] B. Haniharan, L. D. Arbelaez, Bourdev, S. Maji, and J. Malik. Semantic contour from inverse detectors. In *IEEE International Conference on Computer Vision, ICCV*, 2011.

[15] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 991–998, Washington, DC, USA, 2011. IEEE Computer Society.

[16] L. Hörmander. *The analysis of linear partial differential operators I: Distribution theory and Fourier analysis*. Classics in Mathematics. Springer, Berlin, 2nd edition, 2003.

[17] V. P. Krishnan and E. T. Quinto. *Microlocal analysis in tomography.* Handbook of Mathematical Methods in Imaging. Springer, 2015.

[18] G. Kutyniok and D. Labate. Resolution of the wavefront set using continuous shearlets. *Trans. Amer. Math. Soc.*, 361(5):2719–2754, 2009.

[19] G. Kutyniok and W.-Q. Lim. Compactly supported shearlets are optimally sparse. *J. Approx. Theory*, 163:1564–1589, 2011.

[20] G. Kutyniok, W.-Q. Lim, and R. Reisenhofer. ShearLab 3D: Faithful digital shearlet transforms based on compactly supported shearlets. *ACM Trans. Math. Softw.*, 42(1):5:1–5:42, 2016.

[21] G. Kutyniok, W.-Q. Lim, and X. Zhuang. Digital shearlet transforms. In *Shearlets*, pages 239–282. Springer, 2012.

[22] G. Kutyniok and P. Petersen. Classification of edges using compactly supported shearlets. *Appl. Comput. Harmon. Anal.*, 42(2):245–293, 2017.

[23] Y. LeCun, B. Léon, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 1998.

[24] Y. Liu, M.-M. Cheng, D.-P. Fan, L. Zhang, J.-W. Bian, and D. Tao. Semantic edge detection with diverse deep supervision. *IEEE transactions on pattern analysis and machine learning*, 2019.

[25] S. Mallat and W. Liang Hwang. Singularity detection and processing with wavelets. *IEEE Trans. on Information Theory*, 38(2):617–643, 1992.

[26] A. Markoe. *Analytic Tomography*, volume 106 of *Encyclopedia of mathematics and its applications*. Cambridge University Press, 2006.

[27] D. Marr and E. Hildreth. Theory of edge detection. *Proc. R. Soc. Lond. B*, 207(1167):187–217, 1980.

[28] J. M. Prewitt. Object enhancement and extraction. *Pict. Process. Psychopictorics*, 10(1):15–19, 1970.

[29] S. Ramalingam, S. Bouazis, P. Sturm, and M. Brand. Skyline2gps: Localization in urban canyons using omni-skylines. *IEEE-RSJ Int. Conf. Intell. Robot Syst.*, pages 3816–3823, 2010.

[30] R. Reisenhofer, J. Kiefer, and E. J. King. Shearlet-based detection of flame fronts. *Exp. Fluids*, 57, 11 2015.

[31] R. Reisenhofer and E. J. King. Edge, ridge and blob detection with symmetric molecules. *To appear in SIAM Journal of Imaging Sciences*, 2019.

[32] L. G. Roberts. *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology, 1963.

[33] I. Sobel. An isotropic 3x3 image gradient operator. *Presentation at Stanford A.I. Project 1968*, 2014.

[34] V. Torre and T. A. Poggio. On edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(2):147–163, 1986.

[35] F. Trèves. *Introduction to Pseudodifferential and Fourier Integral Operators*, volume 2 of *The University Series in Mathematics*. Plenum Press, New York, 1980.

[36] S. Yi, D. Labate, G. R. Easley, and H. Krim. A shearlet approach to edge analysis and detection. *IEEE Trans. Image Process.*, 18(5):929–941, 2009.

[37] S. Yi, D. Labate, G. R. Easley, and H. Krim. A shearlet approach to edge analysis and detection. *IEEE Trans. Image Process.*, 18(5):929–941, 2009.

[38] Z. Yu, C. Feng, M.-Y. Liu, and S. Ramalingam. Casenet: Deep category-aware semantic edge detection. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA*, pages 21–26, 2017.

[39] Z. Yu, W. Liu, Y. Zou, C. Feng, S. Ramalingam, B. V. Kumar, and J. Kautz. Simultaneous edge alignment and learning. *arXiv preprint arXiv:1808.01992*, 2018.
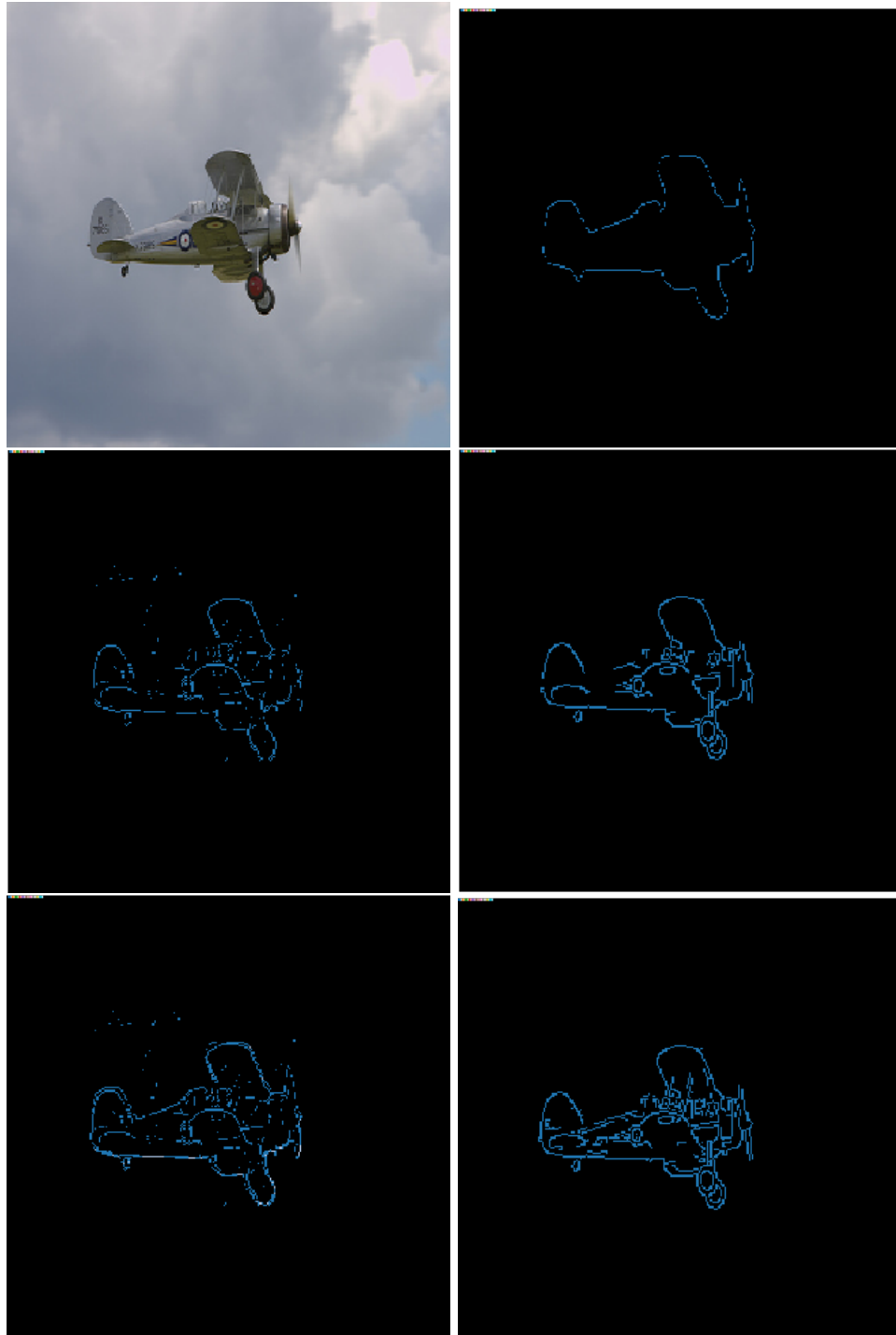
Figure 9: Computed semantic edges of an example of the semantic boundaries dataset (SBD). The color blue represents the category of airplane. Top-left: Input image. Top-right: Semantic edges, human annotation. Middle-left: Semantic edges predicted by the classical CASENet architectures. Middle-right: Semantic edges predicted by the classical DDS architecture. Bottom-left: Semantic edges predicted by the Shear-CASENet architecture. Bottom-right: Semantic edges predicted by the Shear-DDS architecture.
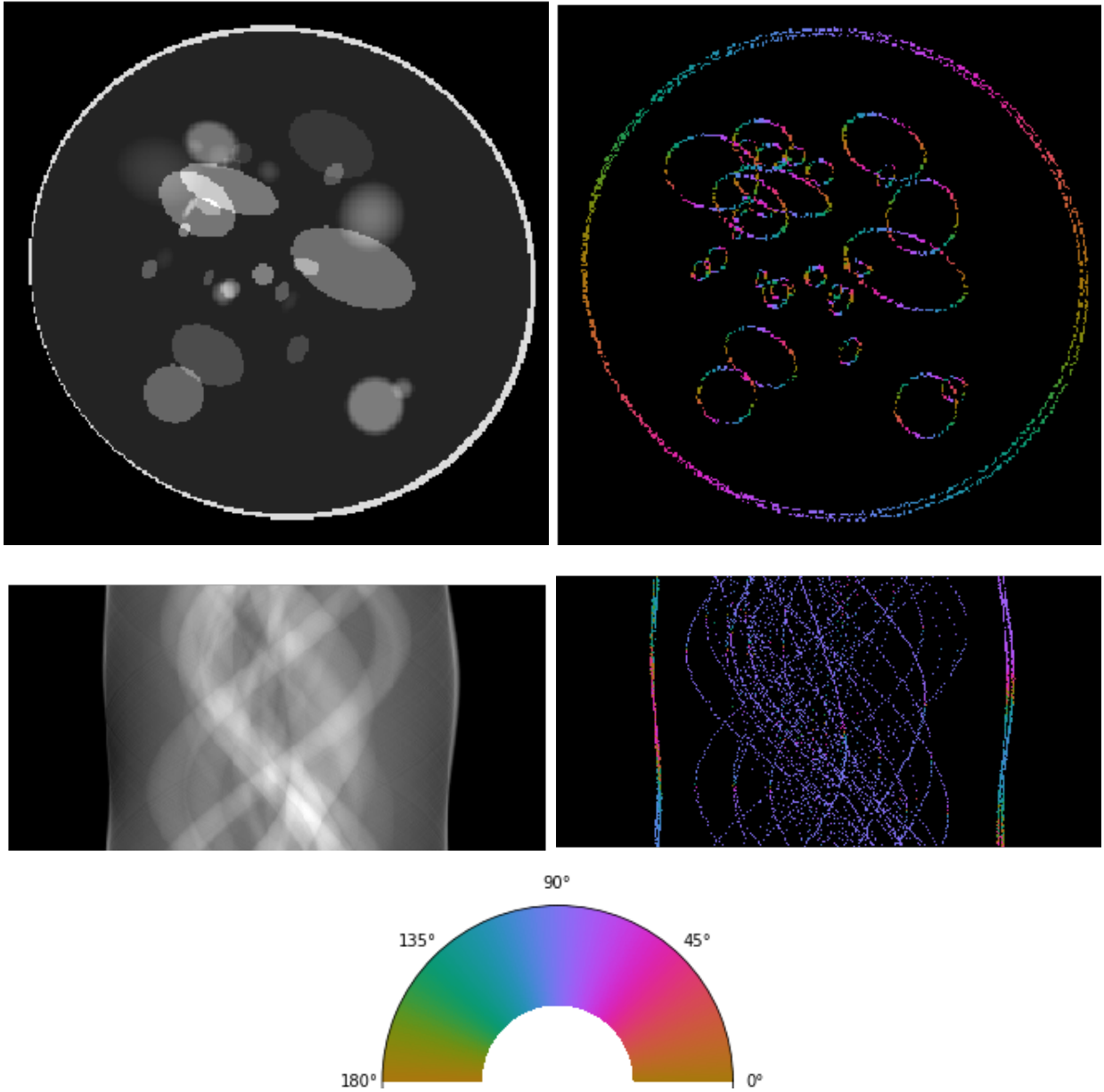
Figure 10: Top-left: Phantom made from ellipses. Top-right: Associated wavefront set extracted by DeNSE. Middle-left: Radon transform of the phantom. Middle-right: Associated wavefront set computed through digital canonical relations. Bottom: Color-code for normal directions.
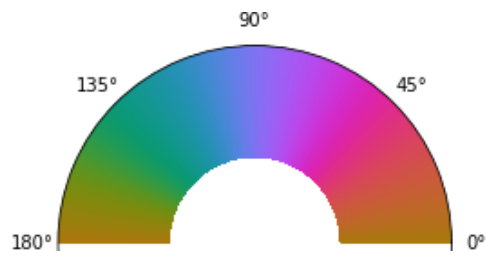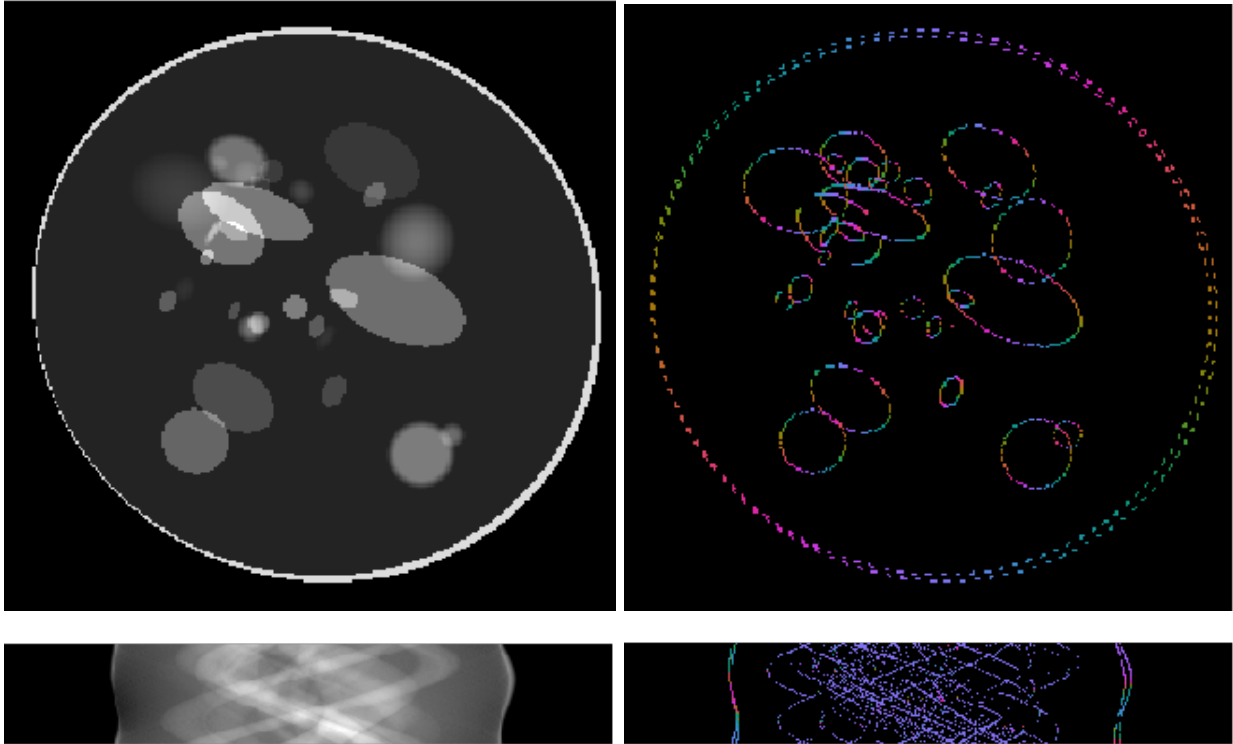
28

Figure 11: Top-left: Phantom made from ellipses. Top-right: Associated wavefront set obtained by the inverse canonical relation on the wavefront set of the low-dose sinogram . Middle-left: Low-sinogram, with every six angles measured. Middle-right: Associated wavefront set obtained by DeNSE. Bottom: Color-code for normal directions.
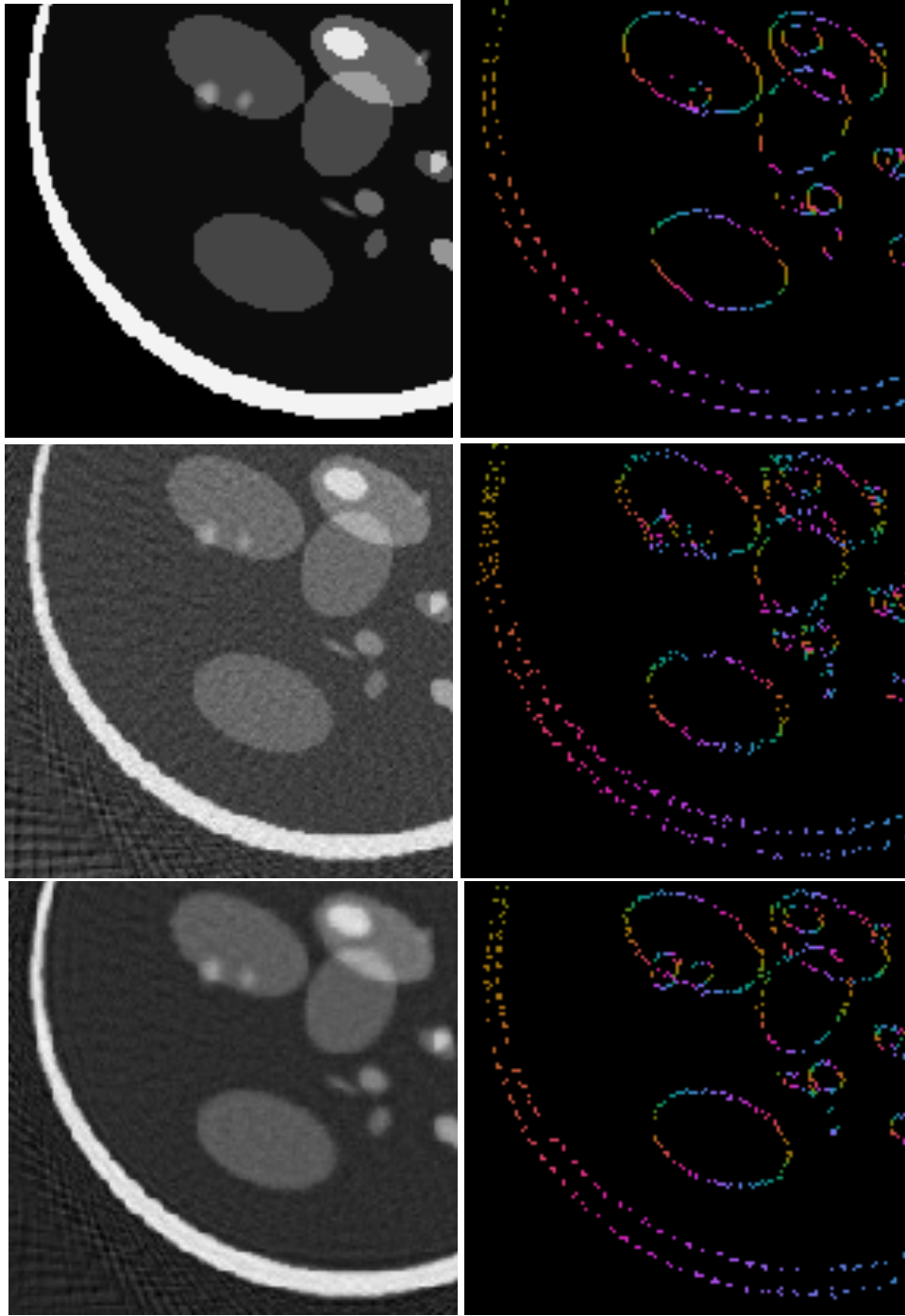
Figure 12: Top-left: Phantom made from ellipses. Top-right: wavefront set of the phantom data computed by the inverse canonical relation on the low dose sinogram wavefront set extracted by DeNSE. Middle-left: Filtered back projection reconstruction. Middle-right: wavefront set of the filtered backprojection reconstruction extracted by DeNSE. Bottom-left: Tikhonov reconstruction. Bottom-right: wavefront set of the Tikhonov reconstruction extracted by DeNSE.