

A Mathematical Theory of Deep Convolutional Neural Networks for Feature Extraction

Thomas Wiatowski¹ and Helmut Bölcskei, *Fellow, IEEE*

Abstract—Deep convolutional neural networks (DCNNs) have led to breakthrough results in numerous practical machine learning tasks, such as classification of images in the ImageNet data set, control-policy-learning to play Atari games or the board game Go, and image captioning. Many of these applications first perform feature extraction and then feed the results thereof into a classifier. The mathematical analysis of DCNNs for feature extraction was initiated by Mallat, 2012. Specifically, Mallat considered so-called scattering networks based on a wavelet transform followed by the modulus non-linearity in each network layer, and proved translation invariance (asymptotically in the wavelet scale parameter) and deformation stability of the corresponding feature extractor. This paper complements Mallat’s results by developing a theory that encompasses general convolutional transforms, or in more technical parlance, general semi-discrete frames (including Weyl–Heisenberg filters, curvelets, shearlets, ridgelets, wavelets, and learned filters), general Lipschitz-continuous non-linearities (e.g., rectified linear units, shifted logistic sigmoids, hyperbolic tangents, and modulus functions), and general Lipschitz-continuous pooling operators emulating, e.g., sub-sampling and averaging. In addition, all of these elements can be different in different network layers. For the resulting feature extractor, we prove a translation invariance result of vertical nature in the sense of the features becoming progressively more translation-invariant with increasing network depth, and we establish deformation sensitivity bounds that apply to signal classes such as, e.g., band-limited functions, cartoon functions, and Lipschitz functions.

Index Terms—Machine learning, deep convolutional neural networks, scattering networks, feature extraction, frame theory.

I. INTRODUCTION

A CENTRAL task in machine learning is feature extraction [2]–[4] as, e.g., in the context of handwritten digit classification [5]. The features to be extracted in this case correspond, for example, to the edges of the digits. The idea behind feature extraction is that feeding characteristic features of the signals—rather than the signals themselves—to a classifier (such as, e.g., a support vector machine (SVM) [6]) improves classification performance. Specifically, non-linear feature extractors can map input signal space dichotomies that are not linearly separable into linearly separable feature space dichotomies [3]. Sticking to the example of handwritten digit

classification, we would, moreover, want the feature extractor to be invariant to the digits’ spatial location within the image, which leads to the requirement of translation invariance. In addition, it is desirable that the feature extractor be robust with respect to (w.r.t.) handwriting styles. This can be accomplished by demanding limited sensitivity of the features to certain non-linear deformations of the signals to be classified.

Spectacular success in practical machine learning tasks has been reported for feature extractors generated by so-called deep convolutional neural networks (DCNNs) [2], [7]–[11], [13], [14]. These networks are composed of multiple layers, each of which computes convolutional transforms, followed by non-linearities and pooling¹ operators. While DCNNs can be used to perform classification (or other machine learning tasks such as regression) directly [2], [7], [9]–[11], typically based on the output of the last network layer, they can also act as stand-alone feature extractors [15]–[21] with the resulting features fed into a classifier such as a SVM. The present paper pertains to the latter philosophy.

The mathematical analysis of feature extractors generated by DCNNs was pioneered by Mallat in [22]. Mallat’s theory applies to so-called scattering networks, where signals are propagated through layers that compute a semi-discrete wavelet transform (i.e., convolutions with filters that are obtained from a mother wavelet through scaling and rotation operations), followed by the modulus non-linearity, without subsequent pooling. The resulting feature extractor is shown to be translation-invariant (asymptotically in the scale parameter of the underlying wavelet transform) and stable w.r.t. certain non-linear deformations. Moreover, Mallat’s scattering networks lead to state-of-the-art results in various classification tasks [23]–[25].

Contributions: DCNN-based feature extractors that were found to work well in practice employ a wide range of i) filters, namely pre-specified structured filters such as wavelets [16], [19]–[21], pre-specified unstructured filters such as random filters [16], [17], and filters that are learned in a supervised [15], [16] or an unsupervised [16]–[18] fashion, ii) non-linearities beyond the modulus function [16], [21], [22], namely hyperbolic tangents [15]–[17], rectified linear units [26], [27], and logistic sigmoids [28], [29], and iii) pooling operators, namely sub-sampling [19], average pooling [15], [16], and max-pooling [16], [17], [20], [21].

¹In the literature “pooling” broadly refers to some form of combining “nearby” values of a signal (e.g., through averaging) or picking one representative value (e.g., through maximization or sub-sampling).

Manuscript received December 18, 2015; revised August 25, 2016 and March 23, 2017; accepted September 24, 2017. Date of publication November 12, 2017; date of current version February 15, 2018. This paper was presented in part at the 2015 IEEE International Symposium on Information Theory, Hong Kong, China.

The authors are with the Department of Information Technology and Electrical Engineering, ETH Zürich, 8092 Zürich, Switzerland (e-mail: withomas@nari.ee.ethz.ch; boelcskei@nari.ee.ethz.ch).

Communicated by A. H. Sayed, Associate Editor for Signal Processing. Digital Object Identifier 10.1109/TIT.2017.2776228

In addition, the filters, non-linearities, and pooling operators can be different in different network layers [14]. The goal of this paper is to develop a mathematical theory that encompasses all these elements (apart from max-pooling) in full generality.

Convolutional transforms as employed in DCNNs can be interpreted as semi-discrete signal transforms [30]–[37] (i.e., convolutional transforms with filters that are countably parametrized). Corresponding prominent representatives are curvelet [34], [35], [38] and shearlet [36], [39] transforms, both of which are known to be highly effective in extracting features characterized by curved edges in images. Our theory allows for general semi-discrete signal transforms, general Lipschitz-continuous non-linearities (e.g., rectified linear units, shifted logistic sigmoids, hyperbolic tangents, and modulus functions), and incorporates continuous-time Lipschitz pooling operators that emulate discrete-time sub-sampling and averaging. Finally, different network layers may be equipped with different convolutional transforms, different (Lipschitz-continuous) non-linearities, and different (Lipschitz-continuous) pooling operators.

Regarding translation invariance, it was argued, e.g., in [15]–[17], [20], and [21], that in practice invariance of the features is crucially governed by network depth and by the presence of pooling operators (such as, e.g., sub-sampling [19], average-pooling [15], [16], or max-pooling [16], [17], [20], [21]). We show that the general feature extractor considered in this paper, indeed, exhibits such a *vertical* translation invariance and that pooling plays a crucial role in achieving it. Specifically, we prove that the depth of the network determines the extent to which the extracted features are translation-invariant. We also show that pooling is necessary to obtain vertical translation invariance as otherwise the features remain fully translation-covariant irrespective of network depth. We furthermore establish a deformation sensitivity bound valid for signal classes such as, e.g., band-limited functions, cartoon functions [40], and Lipschitz functions [40]. This bound shows that small non-linear deformations of the input signal lead to small changes in the corresponding feature vector.

In terms of mathematical techniques, we draw heavily from continuous frame theory [41], [42]. We develop a proof machinery that is completely detached from the structures² of the semi-discrete transforms and the specific form of the Lipschitz non-linearities and Lipschitz pooling operators. The proof of our deformation sensitivity bound is based on two key elements, namely Lipschitz continuity of the feature extractor and a deformation sensitivity bound for the signal class under consideration, namely band-limited functions (as established in the present paper) or cartoon functions and Lipschitz functions as shown in [40]. This “decoupling” approach has important practical ramifications as it shows that whenever we have deformation sensitivity bounds for a signal class, we automatically get deformation sensitivity bounds for the

DCNN feature extractor operating on that signal class. Our results hence establish that vertical translation invariance and limited sensitivity to deformations—for signal classes with inherent deformation insensitivity—are guaranteed by the network structure per se rather than the specific convolution kernels, non-linearities, and pooling operators.

Notation: The complex conjugate of $z \in \mathbb{C}$ is denoted by \bar{z} . We write $\text{Re}(z)$ for the real, and $\text{Im}(z)$ for the imaginary part of $z \in \mathbb{C}$. The Euclidean inner product of $x, y \in \mathbb{C}^d$ is $\langle x, y \rangle := \sum_{i=1}^d x_i \bar{y}_i$, with associated norm $|x| := \sqrt{\langle x, x \rangle}$. We denote the identity matrix by $E \in \mathbb{R}^{d \times d}$. For the matrix $M \in \mathbb{R}^{d \times d}$, $M_{i,j}$ designates the entry in its i -th row and j -th column, and for a tensor $T \in \mathbb{R}^{d \times d \times d}$, $T_{i,j,k}$ refers to its (i, j, k) -th component. The supremum norm of the matrix $M \in \mathbb{R}^{d \times d}$ is defined as $|M|_\infty := \sup_{i,j} |M_{i,j}|$, and the supremum norm of the tensor $T \in \mathbb{R}^{d \times d \times d}$ is $|T|_\infty := \sup_{i,j,k} |T_{i,j,k}|$. We write $B_r(x) \subseteq \mathbb{R}^d$ for the open ball of radius $r > 0$ centered at $x \in \mathbb{R}^d$. $O(d)$ stands for the orthogonal group of dimension $d \in \mathbb{N}$, and $SO(d)$ for the special orthogonal group.

For a Lebesgue-measurable function $f : \mathbb{R}^d \rightarrow \mathbb{C}$, we write $\int_{\mathbb{R}^d} f(x) dx$ for the integral of f w.r.t. Lebesgue measure μ_L . For $p \in [1, \infty)$, $L^p(\mathbb{R}^d)$ stands for the space of Lebesgue-measurable functions $f : \mathbb{R}^d \rightarrow \mathbb{C}$ satisfying $\|f\|_p := (\int_{\mathbb{R}^d} |f(x)|^p dx)^{1/p} < \infty$. $L^\infty(\mathbb{R}^d)$ denotes the space of Lebesgue-measurable functions $f : \mathbb{R}^d \rightarrow \mathbb{C}$ such that $\|f\|_\infty := \inf\{\alpha > 0 \mid |f(x)| \leq \alpha \text{ for a.e. }^3 x \in \mathbb{R}^d\} < \infty$. For $f, g \in L^2(\mathbb{R}^d)$ we set $\langle f, g \rangle := \int_{\mathbb{R}^d} f(x) \bar{g}(x) dx$. For $R > 0$, the space of R -band-limited functions is denoted as $L_R^2(\mathbb{R}^d) := \{f \in L^2(\mathbb{R}^d) \mid \text{supp}(\hat{f}) \subseteq B_R(0)\}$. For a countable set \mathcal{Q} , $(L^2(\mathbb{R}^d))^{\mathcal{Q}}$ stands for the space of sets $s := \{s_q\}_{q \in \mathcal{Q}}$, $s_q \in L^2(\mathbb{R}^d)$, for all $q \in \mathcal{Q}$, satisfying $\|s\| := (\sum_{q \in \mathcal{Q}} \|s_q\|_2^2)^{1/2} < \infty$.

$\text{Id} : L^p(\mathbb{R}^d) \rightarrow L^p(\mathbb{R}^d)$ denotes the identity operator on $L^p(\mathbb{R}^d)$. The tensor product of functions $f, g : \mathbb{R}^d \rightarrow \mathbb{C}$ is $(f \otimes g)(x, y) := f(x)g(y)$, $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$. The operator norm of the bounded linear operator $A : L^p(\mathbb{R}^d) \rightarrow L^q(\mathbb{R}^d)$ is defined as $\|A\|_{p,q} := \sup_{\|f\|_p=1} \|Af\|_q$. We denote the Fourier transform of $f \in L^1(\mathbb{R}^d)$ by $\hat{f}(\omega) := \int_{\mathbb{R}^d} f(x) e^{-2\pi i \langle x, \omega \rangle} dx$ and extend it in the usual way to $L^2(\mathbb{R}^d)$ [43, Th. 7.9]. The convolution of $f \in L^2(\mathbb{R}^d)$ and $g \in L^1(\mathbb{R}^d)$ is $(f * g)(y) := \int_{\mathbb{R}^d} f(x) g(y - x) dx$. We write $(T_t f)(x) := f(x - t)$, $t \in \mathbb{R}^d$, for the translation operator, and $(M_\omega f)(x) := e^{2\pi i \langle x, \omega \rangle} f(x)$, $\omega \in \mathbb{R}^d$, for the modulation operator. Involution is defined by $(If)(x) := \overline{f(-x)}$.

A multi-index $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ is an ordered d -tuple of non-negative integers $\alpha_i \in \mathbb{N}_0$. For a multi-index $\alpha \in \mathbb{N}_0^d$, D^α denotes the differential operator $D^\alpha := (\partial/\partial x_1)^{\alpha_1} \dots (\partial/\partial x_d)^{\alpha_d}$, with order $|\alpha| := \sum_{i=1}^d \alpha_i$. If $|\alpha|=0$, $D^\alpha f := f$, for $f : \mathbb{R}^d \rightarrow \mathbb{C}$. The space of functions $f : \mathbb{R}^d \rightarrow \mathbb{C}$ whose derivatives $D^\alpha f$ of order at most $N \in \mathbb{N}_0$ are continuous is designated by $C^N(\mathbb{R}^d, \mathbb{C})$, and the space of infinitely differentiable functions is $C^\infty(\mathbb{R}^d, \mathbb{C})$. $S(\mathbb{R}^d, \mathbb{C})$ stands for the Schwartz space, i.e., the space of

²Structure here refers to the structural relationship between the convolution kernels in a given layer, e.g., scaling and rotation operations in the case of the wavelet transform.

³Throughout “a.e.” is w.r.t. Lebesgue measure.

functions $f \in C^\infty(\mathbb{R}^d, \mathbb{C})$ whose derivatives $D^\alpha f$ along with the function itself are rapidly decaying [43, Sec. 7.3] in the sense of $\sup_{|\alpha| \leq N} \sup_{x \in \mathbb{R}^d} (1 + |x|^2)^N |(D^\alpha f)(x)| < \infty$, for all $N \in \mathbb{N}_0$. We denote the gradient of a function $f : \mathbb{R}^d \rightarrow \mathbb{C}$ as ∇f . The space of continuous mappings $v : \mathbb{R}^p \rightarrow \mathbb{R}^q$ is $C(\mathbb{R}^p, \mathbb{R}^q)$, and for $k, p, q \in \mathbb{N}$, the space of k -times continuously differentiable mappings $v : \mathbb{R}^p \rightarrow \mathbb{R}^q$ is written as $C^k(\mathbb{R}^p, \mathbb{R}^q)$. For a mapping $v : \mathbb{R}^d \rightarrow \mathbb{R}^d$, we let Dv be its Jacobian matrix, and D^2v its Jacobian tensor, with associated norms $\|v\|_\infty := \sup_{x \in \mathbb{R}^d} |v(x)|$, $\|Dv\|_\infty := \sup_{x \in \mathbb{R}^d} |(Dv)(x)|_\infty$, and $\|D^2v\|_\infty := \sup_{x \in \mathbb{R}^d} |(D^2v)(x)|_\infty$.

II. SCATTERING NETWORKS

We set the stage by reviewing scattering networks as introduced in [22], the basis of which is a multi-layer architecture that involves a wavelet transform followed by the modulus non-linearity, without subsequent pooling. Specifically, [22, Def. 2.4] defines the feature vector $\Phi_W(f)$ of the signal $f \in L^2(\mathbb{R}^d)$ as the set⁴

$$\Phi_W(f) := \bigcup_{n=0}^{\infty} \Phi_W^n(f), \quad (1)$$

where $\Phi_W^0(f) := \{f * \psi_{(-J,0)}\}$, and

$$\Phi_W^n(f) := \left\{ \left(U \left[\underbrace{\lambda^{(j)}, \dots, \lambda^{(p)}}_{n \text{ indices}} \right] f \right) * \psi_{(-J,0)} \right\}_{\lambda^{(j)}, \dots, \lambda^{(p)} \in \Lambda_W \setminus \{(-J,0)\}},$$

for all $n \in \mathbb{N}$, with

$$U[\lambda^{(j)}, \dots, \lambda^{(p)}]f := \underbrace{|\dots| |f * \psi_{\lambda^{(j)}}| * \psi_{\lambda^{(k)}}| \dots * \psi_{\lambda^{(p)}}|}_{n\text{-fold convolution followed by modulus}}.$$

Here, the index set $\Lambda_W := \{(-J,0)\} \cup \{(j,k) \mid j \in \mathbb{Z} \text{ with } j > -J, k \in \{0, \dots, K-1\}\}$ contains pairs of scales j and directions k (in fact, k is the index of the direction described by the rotation matrix r_k), and

$$\psi_\lambda(x) := 2^{dj} \psi(2^j r_k^{-1} x), \quad (2)$$

where $\lambda = (j,k) \in \Lambda_W \setminus \{(-J,0)\}$ are directional wavelets [30], [44], [45] with (complex-valued) mother wavelet $\psi \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$. The r_k , $k \in \{0, \dots, K-1\}$, are elements of a finite rotation group G (if d is even, G is a subgroup of $SO(d)$; if d is odd, G is a subgroup of $O(d)$). The index $(-J,0) \in \Lambda_W$ is associated with the low-pass filter $\psi_{(-J,0)} \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$, and $J \in \mathbb{Z}$ corresponds to the coarsest scale resolved by the directional wavelets (2).

The family of functions $\{\psi_\lambda\}_{\lambda \in \Lambda_W}$ is taken to form a semi-discrete Parseval frame

$$\Psi_{\Lambda_W} := \{T_b I \psi_\lambda\}_{b \in \mathbb{R}^d, \lambda \in \Lambda_W},$$

for $L^2(\mathbb{R}^d)$ [30], [41], [42] and hence satisfies

$$\sum_{\lambda \in \Lambda_W} \int_{\mathbb{R}^d} |\langle f, T_b I \psi_\lambda \rangle|^2 db = \sum_{\lambda \in \Lambda_W} \|f * \psi_\lambda\|_2^2 = \|f\|_2^2,$$

⁴We emphasize that the feature vector $\Phi_W(f)$ is a union of the sets of feature vectors $\Phi_W^n(f)$.

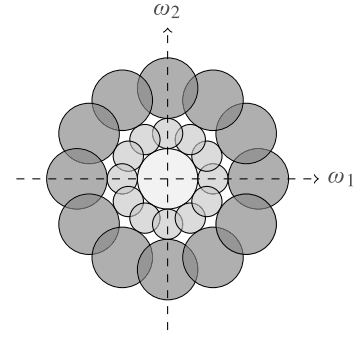


Fig. 1. Partitioning of the frequency plane \mathbb{R}^2 induced by a semi-discrete directional wavelet frame with $K = 12$ directions.

for all $f \in L^2(\mathbb{R}^d)$, where $\langle f, T_b I \psi_\lambda \rangle = (f * \psi_\lambda)(b)$, $(\lambda, b) \in \Lambda_W \times \mathbb{R}^d$, are the underlying frame coefficients. Note that for given $\lambda \in \Lambda_W$, we actually have a continuum of frame coefficients as the translation parameter $b \in \mathbb{R}^d$ is left unsampled. We refer to Figure 1 for a frequency-domain illustration of a semi-discrete directional wavelet frame. In Appendix A, we give a brief review of the general theory of semi-discrete frames, and in Appendices B and C we collect structured example frames in 1-D and 2-D, respectively.

The architecture corresponding to the feature extractor Φ_W in (1), illustrated in Fig. 2, is known as *scattering network* [22], and employs the frame Ψ_{Λ_W} and the modulus non-linearity $|\cdot|$ in every network layer, but does not include pooling. For given $n \in \mathbb{N}$, the set $\Phi_W^n(f)$ in (1) corresponds to the features of the function f generated in the n -th network layer, see Fig. 2.

Remark 1: The function $|f * \psi_\lambda|$, $\lambda \in \Lambda_W \setminus \{(-J,0)\}$, can be thought of as indicating the locations of singularities of $f \in L^2(\mathbb{R}^d)$. Specifically, with the relation of $|f * \psi_\lambda|$ to the Canny edge detector [46] as described in [31], in dimension $d = 2$, we can think of $|f * \psi_\lambda| = |f * \psi_{(j,k)}|$, $\lambda = (j,k) \in \Lambda_W \setminus \{(-J,0)\}$, as an image at scale j specifying the locations of edges of the image f that are oriented in direction k . Furthermore, it was argued in [23], [25], and [47] that the feature vector $\Phi_W^1(f)$ generated in the first layer of the scattering network is very similar, in dimension $d = 1$, to mel frequency cepstral coefficients [48], and in dimension $d = 2$ to SIFT-descriptors [49], [50].

It is shown in [22, Th. 2.10] that the feature extractor Φ_W is translation-invariant in the sense of

$$\lim_{J \rightarrow \infty} \|\Phi_W(T_t f) - \Phi_W(f)\| = 0, \quad (3)$$

for all $f \in L^2(\mathbb{R}^d)$ and $t \in \mathbb{R}^d$. This invariance result is asymptotic in the scale parameter $J \in \mathbb{Z}$ and does not depend on the network depth, i.e., it guarantees full translation invariance in every network layer. Furthermore, [22, Th. 2.12] establishes that Φ_W is stable w.r.t. deformations of the form $(F_\tau f)(x) := f(x - \tau(x))$. More formally, for the function space $(H_W, \|\cdot\|_{H_W})$ defined in [22, Eq. (2.46)], it is shown in [22, Th. 2.12] that there exists a constant $C > 0$ such that for all $f \in H_W$, and

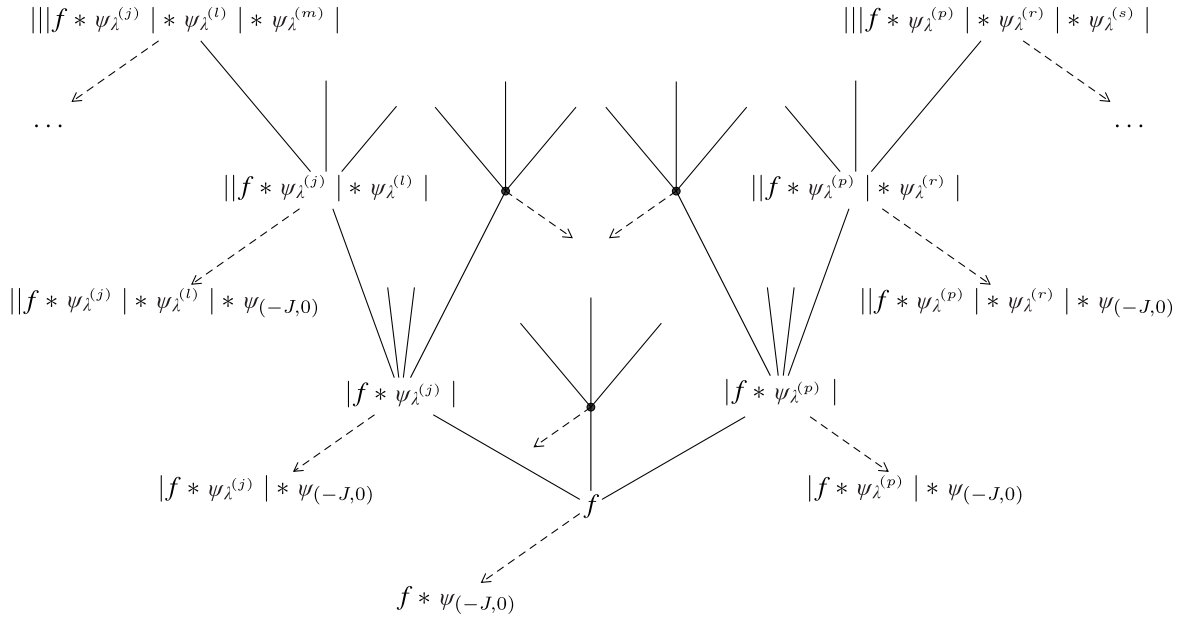


Fig. 2. Scattering network architecture based on wavelet filters and the modulus non-linearity. The elements of the feature vector $\Phi_W(f)$ in (1) are indicated at the tips of the arrows.

$\tau \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ with⁵ $\|D\tau\|_\infty \leq \frac{1}{2d}$, the deformation error satisfies the following deformation stability bound

$$\begin{aligned} & \| \Phi_W(F_\tau f) - \Phi_W(f) \| \\ & \leq C(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty)\|f\|_{H_W}. \end{aligned} \quad (4)$$

Note that this upper bound goes to infinity as translation invariance through $J \rightarrow \infty$ is induced. In practice signal classification based on scattering networks is performed as follows. First, the function f and the wavelet frame atoms $\{\psi_\lambda\}_{\lambda \in \Lambda_W}$ are discretized to finite-dimensional vectors. The resulting scattering network then computes the finite-dimensional feature vector $\Phi_W(f)$, whose dimension is typically reduced through an orthogonal least squares step [51], and then feeds the result into a classifier such as, e.g., a SVM. State-of-the-art results for scattering networks were reported for various classification tasks such as handwritten digit recognition [23], texture discrimination [23], [24], and musical genre classification [25].

III. GENERAL DEEP CONVOLUTIONAL FEATURE EXTRACTORS

As already mentioned, scattering networks follow the architecture of DCNNs [2], [7]–[11], [15]–[21] in the sense of cascading convolutions (with atoms $\{\psi_\lambda\}_{\lambda \in \Lambda_W}$ of the wavelet frame Ψ_{Λ_W}) and non-linearities, namely the modulus function, but without pooling. General DCNNs as used in the literature exhibit a number of additional features:

- a wide variety of filters are employed, namely pre-specified unstructured filters such as random filters [16], [17], and filters that are learned in a

supervised [15], [16] or an unsupervised [16]–[18] fashion.

- a wide variety of non-linearities are used such as, e.g., hyperbolic tangents [15]–[17], rectified linear units [26], [27], and logistic sigmoids [28], [29].
- convolution and the application of a non-linearity is typically followed by a pooling operator such as, e.g., sub-sampling [19], average-pooling [15], [16], or max-pooling [16], [17], [20], [21].
- the filters, non-linearities, and pooling operators are allowed to be different in different network layers [11], [14].

As already mentioned, the purpose of this paper is to develop a mathematical theory of DCNNs for feature extraction that encompasses all of the aspects above (apart from max-pooling) with the proviso that the pooling operators we analyze are continuous-time emulations of discrete-time pooling operators. Formally, compared to scattering networks, in the n -th network layer, we replace the wavelet-modulus operation $|f * \psi_\lambda|$ by a convolution with the atoms $g_{\lambda_n} \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ of a general semi-discrete frame $\Psi_n := \{T_b I g_{\lambda_n}\}_{b \in \mathbb{R}^d, \lambda_n \in \Lambda_n}$ for $L^2(\mathbb{R}^d)$ with countable index set Λ_n (see Appendix A for a brief review of the theory of semi-discrete frames), followed by a non-linearity $M_n : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$ that satisfies the Lipschitz property $\|M_n f - M_n h\|_2 \leq L_n \|f - h\|_2$, for all $f, h \in L^2(\mathbb{R}^d)$, and $M_n f = 0$ for $f = 0$. The output of this non-linearity, $M_n(f * g_{\lambda_n})$, is then pooled according to

$$f \mapsto S_n^{d/2} P_n(f)(S_n \cdot), \quad (5)$$

where $S_n \geq 1$ is the pooling factor and $P_n : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$ satisfies the Lipschitz property $\|P_n f - P_n h\|_2 \leq R_n \|f - h\|_2$, for all $f, h \in L^2(\mathbb{R}^d)$, and $P_n f = 0$ for $f = 0$. We next comment on the individual elements in our network architecture in more detail. The frame atoms g_{λ_n} are

⁵It is actually the assumption $\|D\tau\|_\infty \leq \frac{1}{2d}$, rather than $\|D\tau\|_\infty \leq \frac{1}{2}$ as stated in [22, Th. 2.12], that is needed in [22, p. 1390] to establish that $|\det(E - (D\tau)(x))| \geq 1 - d\|D\tau\|_\infty \geq 1/2$.

arbitrary and can, therefore, also be taken to be structured, e.g., Weyl-Heisenberg functions, curvelets, shearlets, ridgelets, or wavelets as considered in [22] (where the atoms g_{λ_n} are obtained from a mother wavelet through scaling and rotation operations, see Section II). The corresponding semi-discrete signal transforms,⁶ briefly reviewed in Appendices B and C, have been employed successfully in the literature in various feature extraction tasks [32], [54]–[61], but their use—apart from wavelets—in DCNNs appears to be new. We refer the reader to Appendix D for a detailed discussion of several relevant example non-linearities (e.g., rectified linear units, shifted logistic sigmoids, hyperbolic tangents, and, of course, the modulus function) that fit into our framework. We next explain how the continuous-time pooling operator (5) emulates discrete-time pooling by sub-sampling [19] or by averaging [15], [16]. Consider a one-dimensional discrete-time signal $f_d \in \ell^2(\mathbb{Z}) := \{f_d : \mathbb{Z} \rightarrow \mathbb{C} \mid \sum_{k \in \mathbb{Z}} |f_d[k]|^2 < \infty\}$. Sub-sampling by a factor of $S \in \mathbb{N}$ in discrete time is defined by [62, Sec. 4]

$$f_d \mapsto h_d := f_d[S \cdot]$$

and amounts to simply retaining every S -th sample of f_d . The discrete-time Fourier transform of h_d is given by a summation over translated and dilated copies of \hat{f}_d according to [62, Sec. 4]

$$\hat{h}_d(\theta) := \sum_{k \in \mathbb{Z}} h_d[k] e^{-2\pi i k \theta} = \frac{1}{S} \sum_{k=0}^{S-1} \hat{f}_d\left(\frac{\theta - k}{S}\right). \quad (6)$$

The translated copies of \hat{f}_d in (6) are a consequence of the 1-periodicity of the discrete-time Fourier transform. We therefore emulate the discrete-time sub-sampling operation in continuous time through the dilation operation

$$f \mapsto h := S^{d/2} f(S \cdot), \quad f \in L^2(\mathbb{R}^d), \quad (7)$$

which in the frequency domain amounts to dilation according to $\hat{h} = S^{-d/2} \hat{f}(S^{-1} \cdot)$. The scaling by $S^{d/2}$ in (7) ensures unitarity of the continuous-time sub-sampling operation. The overall operation in (7) fits into our general definition of pooling as it can be recovered from (5) simply by taking P to equal the identity mapping (which is, of course, Lipschitz-continuous with Lipschitz constant $R = 1$ and satisfies $\text{Id}f = 0$ for $f = 0$). Next, we consider average pooling. In discrete time average pooling is defined by

$$f_d \mapsto h_d := (f_d * \phi_d)[S \cdot] \quad (8)$$

for the (typically compactly supported) “averaging kernel” $\phi_d \in \ell^2(\mathbb{Z})$ and the averaging factor $S \in \mathbb{N}$. Taking ϕ_d to be a box function of length S amounts to computing local averages of S consecutive samples. Weighted averages are obtained by identifying the desired weights with the averaging

⁶Let $\{g_\lambda\}_{\lambda \in \Lambda} \subseteq L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ be a set of functions indexed by a countable set Λ . Then, the mapping $f \mapsto \{f * g_\lambda(b)\}_{b \in \mathbb{R}^d, \lambda \in \Lambda} = \{(f, T_b I g_\lambda)\}_{\lambda \in \Lambda, f \in L^2(\mathbb{R}^d)}$, is called a semi-discrete signal transform, as it depends on discrete indices $\lambda \in \Lambda$ and continuous variables $b \in \mathbb{R}^d$. We can think of this mapping as the analysis operator in frame theory [53], with the proviso that for given $\lambda \in \Lambda$, we actually have a continuum of frame coefficients as the translation parameter $b \in \mathbb{R}^d$ is left unsampled.

kernel ϕ_d . The operation (8) can be emulated in continuous time according to

$$f \mapsto S^{d/2} (f * \phi)(S \cdot), \quad f \in L^2(\mathbb{R}^d), \quad (9)$$

with the averaging window $\phi \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$. We note that (9) can be recovered from (5) by taking $P(f) = f * \phi$, $f \in L^2(\mathbb{R}^d)$, and noting that convolution with ϕ is Lipschitz-continuous with Lipschitz constant $R = \|\phi\|_1$ (thanks to Young’s inequality [63, Th. 1.2.12]) and trivially satisfies $Pf = 0$ for $f = 0$. In the remainder of the paper, we refer to the operation in (5) as *Lipschitz pooling through dilation* to indicate that (5) essentially amounts to the application of a Lipschitz-continuous mapping followed by a continuous-time dilation. Note, however, that the operation in (5) will not be unitary in general.

We next state definitions and collect preliminary results needed for the analysis of the general DCNN feature extractor considered. The basic building blocks of this network are the triplets (Ψ_n, M_n, P_n) associated with individual network layers n and referred to as *modules*.

Definition 1: For $n \in \mathbb{N}$, let $\Psi_n = \{T_b I g_{\lambda_n}\}_{b \in \mathbb{R}^d, \lambda_n \in \Lambda_n}$ be a semi-discrete frame for $L^2(\mathbb{R}^d)$ and let $M_n : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$ and $P_n : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$ be Lipschitz-continuous operators with $M_n f = 0$ and $P_n f = 0$ for $f = 0$, respectively. Then, the sequence of triplets

$$\Omega := ((\Psi_n, M_n, P_n))_{n \in \mathbb{N}}$$

is referred to as a *module-sequence*.

The following definition introduces the concept of paths on index sets, which will prove useful in formalizing the feature extraction network. The idea for this formalism is due to [22].

Definition 2: Let $\Omega = ((\Psi_n, M_n, P_n))_{n \in \mathbb{N}}$ be a module-sequence, let $\{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$ be the atoms of the frame Ψ_n , and let $S_n \geq 1$ be the pooling factor (according to (5)) associated with the n -th network layer. Define the operator U_n associated with the n -th layer of the network as $U_n : \Lambda_n \times L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$,

$$U_n(\lambda_n, f) := U_n[\lambda_n]f := S_n^{d/2} P_n(M_n(f * g_{\lambda_n}))(S_n \cdot). \quad (10)$$

For $n \in \mathbb{N}$, define the set $\Lambda_1^n := \Lambda_1 \times \Lambda_2 \times \dots \times \Lambda_n$. An ordered sequence $q = (\lambda_1, \lambda_2, \dots, \lambda_n) \in \Lambda_1^n$ is called a *path*. For the empty path $e := \emptyset$ we set $\Lambda_1^0 := \{e\}$ and $U_0[e]f := f$, for all $f \in L^2(\mathbb{R}^d)$.

The operator U_n is well-defined, i.e., $U_n[\lambda_n]f \in L^2(\mathbb{R}^d)$, for all $(\lambda_n, f) \in \Lambda_n \times L^2(\mathbb{R}^d)$, thanks to

$$\begin{aligned} & \|U_n[\lambda_n]f\|_2^2 \\ &= S_n^d \int_{\mathbb{R}^d} |P_n(M_n(f * g_{\lambda_n}))(S_n x)|^2 dx \\ &= \int_{\mathbb{R}^d} |P_n(M_n(f * g_{\lambda_n}))(y)|^2 dy \\ &= \|P_n(M_n(f * g_{\lambda_n}))\|_2^2 \leq R_n^2 \|M_n(f * g_{\lambda_n})\|_2^2 \\ &\leq L_n^2 R_n^2 \|f * g_{\lambda_n}\|_2^2 \leq B_n L_n^2 R_n^2 \|f\|_2^2. \end{aligned} \quad (11)$$

For the inequality in (11) we used the Lipschitz continuity of P_n according to $\|P_n f - P_n h\|_2^2 \leq R_n^2 \|f - h\|_2^2$, together with $P_n h = 0$ for $h = 0$ to get $\|P_n f\|_2^2 \leq R_n^2 \|f\|_2^2$.

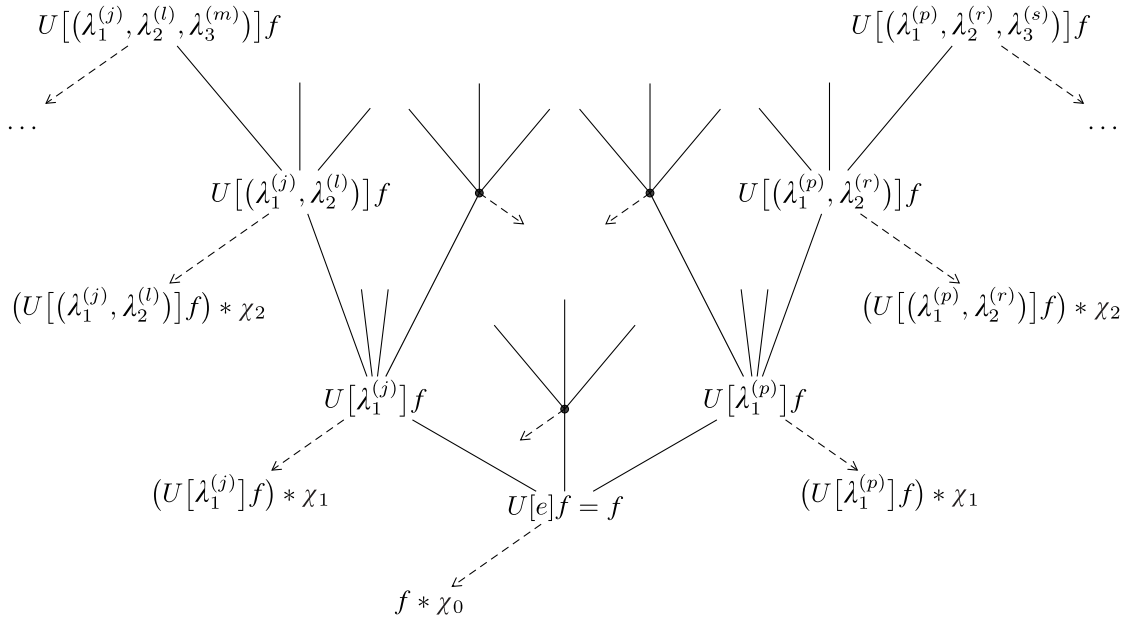


Fig. 3. Network architecture underlying the general DCNN feature extractor. The index $\lambda_n^{(k)}$ corresponds to the k -th atom $g_{\lambda_n^{(k)}}$ of the frame Ψ_n associated with the n -th network layer. The function χ_n is the output-generating atom of the n -th layer.

Similar arguments lead to the first inequality in (12). The last step in (12) is thanks to

$$\|f * g_{\lambda_n}\|_2^2 \leq \sum_{\lambda'_n \in \Lambda_n} \|f * g_{\lambda'_n}\|_2^2 \leq B_n \|f\|_2^2,$$

which follows from the frame condition (30) on Ψ_n . We will also need the extension of the operator U_n to paths $q \in \Lambda_1^n$ according to

$$\begin{aligned} U[q]f &= U[(\lambda_1, \lambda_2, \dots, \lambda_n)]f \\ &:= U_n[\lambda_n] \cdots U_2[\lambda_2] U_1[\lambda_1]f, \end{aligned} \quad (13)$$

with $U[e]f := f$. Note that the multi-stage operation (13) is again well-defined thanks to

$$\|U[q]f\|_2^2 \leq \left(\prod_{k=1}^n B_k L_k^2 R_k^2 \right) \|f\|_2^2, \quad (14)$$

for $q \in \Lambda_1^n$ and $f \in L^2(\mathbb{R}^d)$, which follows by repeated application of (12).

In scattering networks one atom ψ_λ , $\lambda \in \Lambda_W$, in the wavelet frame Ψ_{Λ_W} , namely the low-pass filter $\psi_{(-J,0)}$, is singled out to generate the extracted features according to (1), see also Fig. 2. We follow this construction and designate one of the atoms in each frame in the module-sequence $\Omega = ((\Psi_n, M_n, P_n))_{n \in \mathbb{N}}$ as the output-generating atom $\chi_{n-1} := g_{\lambda_n^*}$, $\lambda_n^* \in \Lambda_n$, of the $(n-1)$ -th layer. The atoms $\{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n \setminus \{\lambda_n^*\}} \cup \{\chi_{n-1}\}$ in Ψ_n are thus used across two consecutive layers in the sense of $\chi_{n-1} = g_{\lambda_n^*}$ generating the output in the $(n-1)$ -th layer, and the $\{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n \setminus \{\lambda_n^*\}}$ propagating signals from the $(n-1)$ -th layer to the n -th layer according to (10), see Fig. 3. Note, however, that our theory does not require the output-generating atoms to be low-pass filters.⁷ From now on, with slight abuse of notation,

⁷It is evident, though, that the actual choices of the output-generating atoms will have an impact on practical performance.

we shall write Λ_n for $\Lambda_n \setminus \{\lambda_n^*\}$ as well. Finally, we note that extracting features in every network layer via an output-generating atom can be regarded as employing skip-layer connections [13], which skip network layers further down and feed the propagated signals into the feature vector.

We are now ready to define the feature extractor Φ_Ω based on the module-sequence Ω .

Definition 3: Let $\Omega = ((\Psi_n, M_n, P_n))_{n \in \mathbb{N}}$ be a module-sequence. The feature extractor Φ_Ω based on Ω maps $L^2(\mathbb{R}^d)$ to its feature vector

$$\Phi_\Omega(f) := \bigcup_{n=0}^{\infty} \Phi_\Omega^n(f), \quad (15)$$

where $\Phi_\Omega^n(f) := \{(U[q]f) * \chi_n\}_{q \in \Lambda_1^n}$, for all $n \in \mathbb{N}$.

The set $\Phi_\Omega^n(f)$ in (15) corresponds to the features of the function f generated in the n -th network layer, see Fig. 3, where $n = 0$ corresponds to the root of the network. The feature extractor $\Phi_\Omega : L^2(\mathbb{R}^d) \rightarrow (L^2(\mathbb{R}^d))^{\mathcal{Q}}$, with $\mathcal{Q} := \bigcup_{n=0}^{\infty} \Lambda_1^n$, is well-defined, i.e., $\Phi_\Omega(f) \in (L^2(\mathbb{R}^d))^{\mathcal{Q}}$, for all $f \in L^2(\mathbb{R}^d)$, under a technical condition on the module-sequence Ω formalized as follows.

Proposition 1: Let $\Omega = ((\Psi_n, M_n, P_n))_{n \in \mathbb{N}}$ be a module-sequence. Denote the frame upper bounds of Ψ_n by $B_n > 0$ and the Lipschitz constants of the operators M_n and P_n by $L_n > 0$ and $R_n > 0$, respectively. If

$$\max\{B_n, B_n L_n^2 R_n^2\} \leq 1, \quad \forall n \in \mathbb{N}, \quad (16)$$

then the feature extractor $\Phi_\Omega : L^2(\mathbb{R}^d) \rightarrow (L^2(\mathbb{R}^d))^{\mathcal{Q}}$ is well-defined, i.e., $\Phi_\Omega(f) \in (L^2(\mathbb{R}^d))^{\mathcal{Q}}$, for all $f \in L^2(\mathbb{R}^d)$.

Proof: The proof is given in Appendix E. \square

As condition (16) is of central importance, we formalize it as follows.

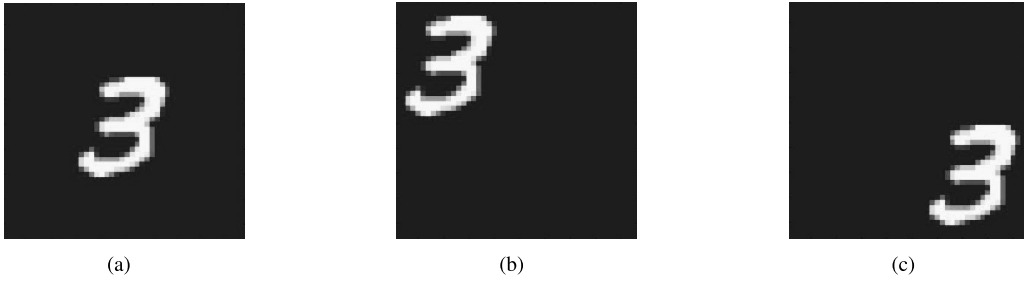


Fig. 4. Handwritten digits from the MNIST data set [5]. For practical machine learning tasks (e.g., signal classification), we often want the feature vector $\Phi_\Omega(f)$ to be invariant to the digits' spatial location within the image f . Theorem 1 establishes that the features $\Phi_\Omega^n(f)$ become more translation-invariant with increasing layer index n .

Definition 4: Let $\Omega = ((\Psi_n, M_n, P_n))_{n \in \mathbb{N}}$ be a module-sequence with frame upper bounds $B_n > 0$ and Lipschitz constants $L_n, R_n > 0$ of the operators M_n and P_n , respectively. The condition

$$\max\{B_n, B_n L_n^2 R_n^2\} \leq 1, \quad \forall n \in \mathbb{N}, \quad (17)$$

is referred to as *admissibility condition*. Module-sequences that satisfy (17) are called *admissible*.

We emphasize that condition (17) is easily met in practice. To see this, first note that B_n is determined through the frame Ψ_n (e.g., the directional wavelet frame introduced in Section II has $B = 1$), L_n is set through the non-linearity M_n (e.g., the modulus function $M = |\cdot|$ has $L = 1$, see Appendix D), and R_n depends on the operator P_n in (5) (e.g., pooling by sub-sampling amounts to $P = \text{Id}$ and has $R = 1$). Obviously, condition (17) is met if

$$B_n \leq \min\{1, L_n^{-2} R_n^{-2}\}, \quad \forall n \in \mathbb{N},$$

which can be satisfied by simply normalizing the frame elements of Ψ_n accordingly. We refer to Proposition 3 in Appendix A for corresponding normalization techniques, which, as explained in Section IV, affect neither our translation invariance result nor our deformation sensitivity bounds.

IV. PROPERTIES OF THE FEATURE EXTRACTOR Φ_Ω

A. Vertical Translation Invariance

The following theorem states that under very mild decay conditions on the Fourier transforms $\widehat{\chi}_n$ of the output-generating atoms χ_n , the feature extractor Φ_Ω exhibits vertical translation invariance in the sense of the features becoming more translation-invariant with increasing network depth. This result is in line with observations made in the deep learning literature, e.g., in [15]–[17], [20], and [21], where it is informally argued that the network outputs generated at deeper layers tend to be more translation-invariant.

Theorem 1: Let $\Omega = ((\Psi_n, M_n, P_n))_{n \in \mathbb{N}}$ be an admissible module-sequence, let $S_n \geq 1$, $n \in \mathbb{N}$, be the pooling factors in (10), and assume that the operators $M_n : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$ and $P_n : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$ commute with the translation operator T_t , i.e.,

$$M_n T_t f = T_t M_n f, \quad P_n T_t f = T_t P_n f, \quad (18)$$

for all $f \in L^2(\mathbb{R}^d)$, $t \in \mathbb{R}^d$, and $n \in \mathbb{N}$.

- i) The features $\Phi_\Omega^n(f)$ generated in the n -th network layer satisfy

$$\Phi_\Omega^n(T_t f) = T_{t/(S_1 \dots S_n)} \Phi_\Omega^n(f), \quad (19)$$

for all $f \in L^2(\mathbb{R}^d)$, $t \in \mathbb{R}^d$, and $n \in \mathbb{N}$, where $T_t \Phi_\Omega^n(f)$ refers to element-wise application of T_t , i.e., $T_t \Phi_\Omega^n(f) := \{T_t h \mid h \in \Phi_\Omega^n(f)\}$.

- ii) If, in addition, there exists a constant $K > 0$ (that does not depend on n) such that the Fourier transforms $\widehat{\chi}_n$ of the output-generating atoms χ_n satisfy the decay condition

$$|\widehat{\chi}_n(\omega)| |\omega| \leq K, \quad \text{a.e. } \omega \in \mathbb{R}^d, \quad \forall n \in \mathbb{N}_0, \quad (20)$$

then

$$\|\Phi_\Omega^n(T_t f) - \Phi_\Omega^n(f)\| \leq \frac{2\pi |t| K}{S_1 \dots S_n} \|f\|_2, \quad (21)$$

for all $f \in L^2(\mathbb{R}^d)$ and $t \in \mathbb{R}^d$.

Proof: The proof is given in Appendix F. \square

We start by noting that all pointwise (also referred to as memoryless in the signal processing literature) non-linearities $M_n : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$ satisfy the commutation relation in (18). A large class of non-linearities widely used in the deep learning literature, such as rectified linear units, hyperbolic tangents, shifted logistic sigmoids, and the modulus function as employed in [22], are, indeed, pointwise and hence covered by Theorem 1. Moreover, $P = \text{Id}$ as in pooling by sub-sampling trivially satisfies (18). Pooling by averaging $Pf = f * \phi$, with $\phi \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$, satisfies (18) as a consequence of the convolution operator commuting with the translation operator T_t .

Note that (20) can easily be met by taking the output-generating atoms $\{\chi_n\}_{n \in \mathbb{N}_0}$ either to satisfy

$$\sup_{n \in \mathbb{N}_0} \{\|\chi_n\|_1 + \|\nabla \chi_n\|_1\} < \infty,$$

see, e.g., [43, Ch. 7], or to be uniformly band-limited in the sense of $\text{supp}(\widehat{\chi}_n) \subseteq B_r(0)$, for all $n \in \mathbb{N}_0$, with an r that is independent of n (see, e.g., [30, Ch. 2.3]). The bound in (21) shows that we can explicitly control the amount of translation invariance via the pooling factors S_n . This result is in line with observations made in the deep learning literature, e.g., in [15]–[17], [20], and [21], where it is informally argued that pooling is crucial to get translation invariance of the

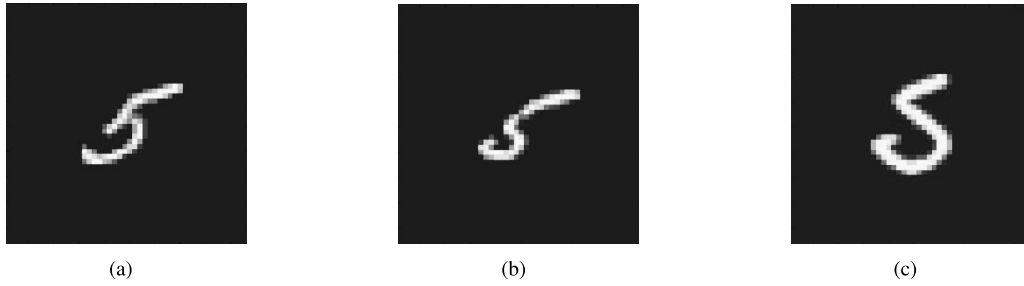


Fig. 5. Handwritten digits from the MNIST data set [5]. If f denotes the image of the handwritten digit “5” in (a), then—for appropriately chosen τ —the function $F_\tau f = f(\cdot - \tau(\cdot))$ models images of “5” based on different handwriting styles as in (b) and (c).

extracted features. Furthermore, the condition $\lim_{n \rightarrow \infty} S_1 \cdot S_2 \cdot \dots \cdot S_n = \infty$ (easily met by taking $S_n > 1$, for all $n \in \mathbb{N}$) guarantees, thanks to (21), asymptotically full translation invariance according to

$$\lim_{n \rightarrow \infty} |||\Phi_\Omega^n(T_t f) - \Phi_\Omega^n(f)||| = 0, \quad (22)$$

for all $f \in L^2(\mathbb{R}^d)$ and $t \in \mathbb{R}^d$. This means that the features $\Phi_\Omega^n(T_t f)$ corresponding to the shifted versions $T_t f$ of the handwritten digit “3” in Figs. 4 (b) and (c) with increasing network depth increasingly “look like” the features $\Phi_\Omega^n(f)$ corresponding to the unshifted handwritten digit in Fig. 4 (a). Casually speaking, the shift operator T_t is increasingly absorbed by Φ_Ω^n as $n \rightarrow \infty$, with the upper bound (21) quantifying this absorption.

In contrast, the translation invariance result (3) in [22] is asymptotic in the wavelet scale parameter J , and does not depend on the network depth, i.e., it guarantees full translation invariance in every network layer. We honor this difference by referring to (3) as *horizontal* translation invariance and to (22) as *vertical* translation invariance.

We emphasize that vertical translation invariance is a structural property. Specifically, if P_n is unitary (such as, e.g., in the case of pooling by sub-sampling where P_n simply equals the identity mapping), then so is the pooling operation in (5) owing to

$$\begin{aligned} & \|S_n^{d/2} P_n(f)(S_n \cdot)\|_2^2 \\ &= S_n^d \int_{\mathbb{R}^d} |P_n(f)(S_n x)|^2 dx \\ &= \int_{\mathbb{R}^d} |P_n(f)(x)|^2 dx = \|P_n(f)\|_2^2 = \|f\|_2^2, \end{aligned}$$

where we employed the change of variables $y = S_n x$, $\frac{dy}{dx} = S_n^d$. Regarding average pooling, as already mentioned, the operators $P_n(f) = f * \phi_n$, $f \in L^2(\mathbb{R}^d)$, $n \in \mathbb{N}$, are, in general, not unitary, but we still get translation invariance as a consequence of structural properties, namely translation covariance of the convolution operator combined with unitary dilation according to (7).

Finally, we note that in practice in certain applications it is actually translation *covariance* in the sense of $\Phi_\Omega^n(T_t f) = T_t \Phi_\Omega^n(f)$, for all $f \in L^2(\mathbb{R}^d)$ and $t \in \mathbb{R}^d$, that is desirable, for example, in facial landmark detection where the goal is to estimate the absolute position of facial landmarks in images. In such applications features in the layers closer to the root

of the network are more relevant as they are less translation-invariant and more translation-covariant. The reader is referred to [64] where corresponding numerical evidence is provided. We proceed to the formal statement of our translation covariance result.

Corollary 1: Let $\Omega = ((\Psi_n, M_n, P_n))_{n \in \mathbb{N}}$ be an admissible module-sequence, let $S_n \geq 1$, $n \in \mathbb{N}$, be the pooling factors in (10), and assume that the operators $M_n : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$ and $P_n : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$ commute with the translation operator T_t in the sense of (18). If, in addition, there exists a constant $K > 0$ (that does not depend on n) such that the Fourier transforms $\hat{\chi}_n$ of the output-generating atoms χ_n satisfy the decay condition (20), then

$$|||\Phi_\Omega^n(T_t f) - T_t \Phi_\Omega^n(f)||| \leq 2\pi |t| K |1/(S_1 \dots S_n) - 1| \|f\|_2,$$

for all $f \in L^2(\mathbb{R}^d)$ and $t \in \mathbb{R}^d$.

Proof: The proof is given in Appendix G. \square

Corollary 1 shows that in the absence of pooling, i.e., taking $S_n = 1$, for all $n \in \mathbb{N}$, leads to full translation covariance in every network layer. This proves that pooling is necessary to get vertical translation invariance as otherwise the features remain fully translation-covariant irrespective of the network depth. Finally, we note that scattering networks [22] (which do not employ pooling operators, see Section II) are rendered horizontally translation-invariant by letting the wavelet scale parameter $J \rightarrow \infty$.

B. Deformation Sensitivity Bound

The next result provides a bound—for band-limited signals $f \in L_R^2(\mathbb{R}^d)$ —on the sensitivity of the feature extractor Φ_Ω w.r.t. time-frequency deformations of the form

$$(F_{\tau, \omega} f)(x) := e^{2\pi i \omega(x)} f(x - \tau(x)).$$

This class of deformations encompasses non-linear distortions $f(x - \tau(x))$ as illustrated in Fig. 5, and modulation-like deformations $e^{2\pi i \omega(x)} f(x)$ which occur, e.g., if the signal f is subject to an undesired modulation and we therefore have access to a bandpass version of f only.

The deformation sensitivity bound we derive is signal-class specific in the sense of applying to input signals belonging to a particular class, here band-limited functions. The proof technique we develop applies, however, to all signal classes that exhibit “inherent” deformation insensitivity in the following sense.

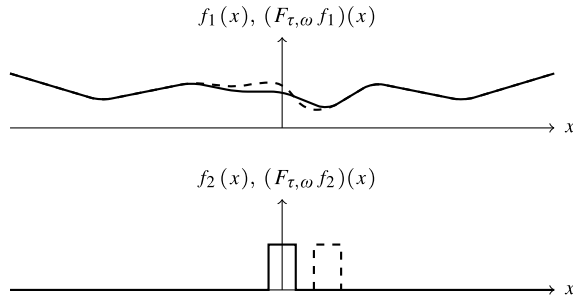


Fig. 6. Impact of the deformation $F_{\tau, \omega}$, with $\tau(x) = \frac{1}{2}e^{-x^2}$ and $\omega = 0$, on the functions $f_1 \in C_1 \subseteq L^2(\mathbb{R})$ and $f_2 \in C_2 \subseteq L^2(\mathbb{R})$. The signal class C_1 consists of smooth, slowly varying functions (e.g., band-limited functions), and C_2 consists of compactly supported functions that exhibit discontinuities (e.g., cartoon functions [65]). We observe that f_1 , unlike f_2 , is affected only mildly by $F_{\tau, \omega}$. The amount of deformation induced therefore depends drastically on the specific $f \in L^2(\mathbb{R})$.

Definition 5: A signal class $\mathcal{C} \subseteq L^2(\mathbb{R}^d)$ is called *deformation-insensitive* if there exist $\alpha, \beta, C > 0$ such that for all $f \in \mathcal{C}$, $\omega \in C(\mathbb{R}^d, \mathbb{R})$, and (possibly non-linear) $\tau \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ with $\|D\tau\|_\infty \leq \frac{1}{2d}$, it holds that

$$\|f - F_{\tau, \omega} f\|_2 \leq C(\|\tau\|_\infty^\alpha + \|\omega\|_\infty^\beta). \quad (23)$$

The constant $C > 0$ and the exponents $\alpha, \beta > 0$ in (23) depend on the particular signal class \mathcal{C} . Examples of deformation-insensitive signal classes are the class of R -band-limited functions (see Proposition 5 in Appendix J), the class of cartoon functions [40, Prop. 1], and the class of Lipschitz functions [40, Lem. 1]. While a deformation sensitivity bound that applies to all $f \in L^2(\mathbb{R}^d)$ would be desirable, the example in Fig. 6 illustrates the difficulty underlying this desideratum. Specifically, we can see in Fig. 6 that for given $\tau(x)$ and $\omega(x)$ the impact of the deformation induced by $e^{2\pi i \omega(x)} f(x - \tau(x))$ can depend drastically on the function $f \in L^2(\mathbb{R}^d)$ itself. The deformation stability bound (4) for scattering networks reported in [22, Th. 2.12] applies to a signal class as well, characterized, albeit implicitly, through [22, Eq. (2.46)] and depending on the mother wavelet and the (modulus) non-linearity.

Our signal-class-specific deformation sensitivity bound is based on the following two ingredients. First, we establish—in Proposition 4 in Appendix I—that the feature extractor Φ_Ω is Lipschitz-continuous with Lipschitz constant $L_\Omega = 1$, i.e.,

$$\|\Phi_\Omega(f) - \Phi_\Omega(h)\| \leq \|f - h\|_2, \quad \forall f, h \in L^2(\mathbb{R}^d), \quad (24)$$

where, thanks to the admissibility condition (17), the Lipschitz constant $L_\Omega = 1$ in (24) is completely independent of the frame upper bounds B_n and the Lipschitz-constants L_n and R_n of M_n and P_n , respectively. Second, we derive—in Proposition 5 in Appendix J—an upper bound on the deformation error $\|f - F_{\tau, \omega} f\|_2$ for R -band-limited functions, i.e., $f \in L_R^2(\mathbb{R}^d)$, according to

$$\|f - F_{\tau, \omega} f\|_2 \leq C(R\|\tau\|_\infty + \|\omega\|_\infty)\|f\|_2. \quad (25)$$

The deformation sensitivity bound for the feature extractor is then obtained by setting $h = F_{\tau, \omega} f$ in (24) and using (25) (see Appendix H for the corresponding technical details). This

“decoupling” into Lipschitz continuity of Φ_Ω and a deformation sensitivity bound for the signal class under consideration (here, band-limited functions) has important practical ramifications as it shows that whenever we have a deformation sensitivity bound for the signal class, we automatically get a deformation sensitivity bound for the feature extractor thanks to its Lipschitz continuity. The same approach was used in [40] to derive deformation sensitivity bounds for cartoon functions and for Lipschitz functions.

Lipschitz continuity of Φ_Ω according to (24) also guarantees that pairwise distances in the input signal space do not increase through feature extraction. An immediate consequence is robustness of the feature extractor w.r.t. additive noise $\eta \in L^2(\mathbb{R}^d)$ in the sense of

$$\|\Phi_\Omega(f + \eta) - \Phi_\Omega(f)\| \leq \|\eta\|_2, \quad \forall f \in L^2(\mathbb{R}^d).$$

We proceed to the formal statement of our deformation sensitivity result.

Theorem 2: Let $\Omega = ((\Psi_n, M_n, P_n))_{n \in \mathbb{N}}$ be an admissible module-sequence. There exists a constant $C > 0$ (that does not depend on Ω) such that for all $f \in L_R^2(\mathbb{R}^d)$, $\omega \in C(\mathbb{R}^d, \mathbb{R})$, and $\tau \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ with $\|D\tau\|_\infty \leq \frac{1}{2d}$, the feature extractor Φ_Ω satisfies

$$\|\Phi_\Omega(F_{\tau, \omega} f) - \Phi_\Omega(f)\| \leq C(R\|\tau\|_\infty + \|\omega\|_\infty)\|f\|_2. \quad (26)$$

Proof: The proof is given in Appendix H. \square

First, we note that the bound in (26) holds for τ with sufficiently “small” Jacobian matrix, i.e., as long as $\|D\tau\|_\infty \leq \frac{1}{2d}$. We can think of this condition on the Jacobian matrix as follows⁸: Let f be an image of the handwritten digit “5” (see Fig. 5 (a)). Then, $\{F_{\tau, \omega} f \mid \|D\tau\|_\infty \leq \frac{1}{2d}\}$ is a collection of images of the handwritten digit “5”, where each $F_{\tau, \omega} f$ models an image that may be generated, e.g., based on a different handwriting style (see Figs. 5 (b) and (c)). The condition $\|D\tau\|_\infty \leq \frac{1}{2d}$ now imposes a quantitative limit on the amount of deformation tolerated. The deformation sensitivity bound (26) provides a limit on how much the features corresponding to the images in the set $\{F_{\tau, \omega} f \mid \|D\tau\|_\infty \leq \frac{1}{2d}\}$ can differ. The strength of Theorem 2 derives itself from the fact that the only condition on the underlying module-sequence Ω needed is admissibility according to (17), which as outlined in Section III, can easily be obtained by normalizing the frame elements of Ψ_n , for all $n \in \mathbb{N}$, appropriately. This normalization does not have an impact on the constant C in (26). More specifically, C is shown in (115) to be completely independent of Ω . All this is thanks to the decoupling technique used to prove Theorem 2 being completely independent of the structures of the frames Ψ_n and of the specific forms of the Lipschitz-continuous operators M_n and P_n . The deformation sensitivity bound (26) is very general in the sense of applying to all Lipschitz-continuous (linear or non-linear) mappings Φ , not only those generated by DCNNs.

The bound (4) for scattering networks reported in [22, Th. 2.12] depends upon first-order ($D\tau$) and second-order ($D^2\tau$) derivatives of τ . In contrast, our bound (26) depends on ($D\tau$) implicitly only as we need to impose the

⁸The ensuing argument is taken from [40].

condition $\|D\tau\|_\infty \leq \frac{1}{2d}$ for the bound to hold.⁹ We honor this difference by referring to (4) as deformation *stability* bound and to our bound (26) as deformation *sensitivity* bound.

The dependence of the upper bound in (26) on the bandwidth R reflects the intuition that the deformation sensitivity bound should depend on the input signal class “description complexity”. Many signals of practical significance (e.g., natural images) are, however, either not band-limited due to the presence of sharp (and possibly curved) edges or exhibit large bandwidths. In the latter case, the bound (26) is effectively rendered void owing to its linear dependence on R . We refer the reader to [40] where deformation sensitivity bounds for non-smooth signals were established. Specifically, the main contributions in [40] are deformation sensitivity bounds—again obtained through decoupling—for non-linear deformations $(F_\tau f)(x) = f(x - \tau(x))$ according to

$$\|f - F_\tau f\|_2 \leq C\|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d), \quad (27)$$

for the signal classes $\mathcal{C} \subseteq L^2(\mathbb{R}^d)$ of cartoon functions [65] and for Lipschitz-continuous functions. The constant $C > 0$ and the exponent $\alpha > 0$ in (27) depend on the particular signal class \mathcal{C} and are specified in [40]. As the vertical translation invariance result in Theorem 1 applies to all $f \in L^2(\mathbb{R}^d)$, the results established in the present paper and in [40] taken together show that vertical translation invariance and limited sensitivity to deformations—for signal classes with inherent deformation insensitivity—are guaranteed by the feature extraction network structure per se rather than the specific convolution kernels, non-linearities, and pooling operators.

Finally, the deformation stability bound (4) for scattering networks reported in [22, Th. 2.12] applies to the space

$$H_W := \left\{ f \in L^2(\mathbb{R}^d) \mid \|f\|_{H_W} < \infty \right\}, \quad (28)$$

where

$$\|f\|_{H_W} := \sum_{n=0}^{\infty} \left(\sum_{q \in (\Lambda_W)_1^n} \|U[q]f\|_2^2 \right)^{1/2}$$

and $(\Lambda_W)_1^n$ denotes the set of paths $q = (\lambda^{(i)}, \dots, \lambda^{(p)})$ of length n with $\lambda^{(i)}, \dots, \lambda^{(p)} \in \Lambda_W$. While [22, p. 1350] cites numerical evidence on the series $\sum_{q \in (\Lambda_W)_1^n} \|U[q]f\|_2^2$ being finite for a large class of signals $f \in L^2(\mathbb{R}^d)$, it seems difficult to establish this analytically, let alone to show that

$$\sum_{n=0}^{\infty} \left(\sum_{q \in (\Lambda_W)_1^n} \|U[q]f\|_2^2 \right)^{1/2} < \infty.$$

In contrast, the deformation sensitivity bound (26) applies *provably* to the space of R -band-limited functions $L_R^2(\mathbb{R}^d)$. Finally, the space H_W in (28) depends on the wavelet frame atoms $\{\psi_\lambda\}_{\lambda \in \Lambda_W}$ and the (modulus) non-linearity, and thereby on the underlying signal transform, whereas $L_R^2(\mathbb{R}^d)$ is, trivially, independent of the module-sequence Ω .

⁹We note that the condition $\|D\tau\|_\infty \leq \frac{1}{2d}$ is needed for the bound (4) to hold as well.

V. FINAL REMARKS AND OUTLOOK

It is interesting to note that the frame lower bounds $A_n > 0$ of the semi-discrete frames Ψ_n affect neither the vertical translation invariance result in Theorem 1 nor the deformation sensitivity bound in Theorem 2. In fact, the entire theory in this paper carries through as long as the collections $\Psi_n = \{T_b I g_{\lambda_n}\}_{b \in \mathbb{R}^d, \lambda_n \in \Lambda_n}$, for all $n \in \mathbb{N}$, satisfy the Bessel property

$$\sum_{\lambda_n \in \Lambda_n} \int_{\mathbb{R}^d} |\langle f, T_b I g_{\lambda_n} \rangle|^2 db = \sum_{\lambda_n \in \Lambda_n} \|f * g_{\lambda_n}\|_2^2 \leq B_n \|f\|_2^2,$$

for all $f \in L^2(\mathbb{R}^d)$ for some $B_n > 0$, which, by Proposition 2, is equivalent to

$$\sum_{\lambda_n \in \Lambda_n} |\widehat{g_{\lambda_n}}(\omega)|^2 \leq B_n, \quad a.e. \omega \in \mathbb{R}^d. \quad (29)$$

Pre-specified unstructured filters [16], [17] and learned filters [15]–[18] are therefore covered by our theory as long as (29) is satisfied. In classical frame theory $A_n > 0$ guarantees completeness of the set $\Psi_n = \{T_b I g_{\lambda_n}\}_{b \in \mathbb{R}^d, \lambda_n \in \Lambda_n}$ for the signal space under consideration, here $L^2(\mathbb{R}^d)$. The absence of a frame lower bound $A_n > 0$ therefore translates into a lack of completeness of Ψ_n , which may result in the frame coefficients $\langle f, T_b I g_{\lambda_n} \rangle = (f * g_{\lambda_n})(b)$, $(\lambda_n, b) \in \Lambda_n \times \mathbb{R}^d$, not containing all essential features of the signal f . This will, in general, have a (possibly significant) impact on practical feature extraction performance which is why ensuring the entire frame property (30) is prudent. Interestingly, satisfying the frame property (30) for all Ψ_n , $n \in \mathbb{Z}$, does, however, not guarantee that the feature extractor Φ_Ω has a trivial null-set, i.e., $\Phi_\Omega(f) = 0$ if and only if $f = 0$. We refer the reader to [66, App. A] for an example of a feature extractor with non-trivial null-set.

APPENDIX A SEMI-DISCRETE FRAMES

This appendix gives a brief review of the theory of semi-discrete frames. A list of structured example frames of interest in the context of this paper is provided in Appendix B for the 1-D case, and in Appendix C for the 2-D case. Semi-discrete frames are instances of *continuous* frames [41], [42], and appear in the literature, e.g., in the context of translation-covariant signal decompositions [31]–[33], and as an intermediate step in the construction of various *fully-discrete* frames [34], [35], [37], [52]. We first collect some basic results on semi-discrete frames.

Definition 6: Let $\{g_\lambda\}_{\lambda \in \Lambda} \subseteq L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ be a set of functions indexed by a countable set Λ . The collection

$$\Psi_\Lambda := \{T_b I g_\lambda\}_{(\lambda, b) \in \Lambda \times \mathbb{R}^d}$$

is a *semi-discrete frame* for $L^2(\mathbb{R}^d)$ if there exist constants $A, B > 0$ such that

$$\begin{aligned} A \|f\|_2^2 &\leq \sum_{\lambda \in \Lambda} \int_{\mathbb{R}^d} |\langle f, T_b I g_\lambda \rangle|^2 db \\ &= \sum_{\lambda \in \Lambda} \|f * g_\lambda\|_2^2 \leq B \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d). \end{aligned} \quad (30)$$

The functions $\{g_\lambda\}_{\lambda \in \Lambda}$ are called the atoms of the frame Ψ_Λ . When $A = B$ the frame is said to be tight. A tight frame with frame bound $A = 1$ is called a Parseval frame.

The frame operator associated with the semi-discrete frame Ψ_Λ is defined in the weak sense as $S_\Lambda : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$,

$$\begin{aligned} S_\Lambda f &:= \sum_{\lambda \in \Lambda} \int_{\mathbb{R}^d} \langle f, T_b I g_\lambda \rangle (T_b I g_\lambda) db \\ &= \left(\sum_{\lambda \in \Lambda} g_\lambda * I g_\lambda \right) * f, \end{aligned} \quad (31)$$

where $\langle f, T_b I g_\lambda \rangle = (f * g_\lambda)(b)$, $(\lambda, b) \in \Lambda \times \mathbb{R}^d$, are called the frame coefficients. S_Λ is a bounded, positive, and boundedly invertible operator [41].

The reader might want to think of semi-discrete frames as shift-invariant frames [67], [68] with a continuous translation parameter, and of the countable index set Λ as labeling a collection of scales, directions, or frequency-shifts, hence the terminology *semi-discrete*. For instance, scattering networks are based on a (single) semi-discrete wavelet frame, where the atoms $\{g_\lambda\}_{\lambda \in \Lambda_W}$ are indexed by the set $\Lambda_W := \{(-J, 0)\} \cup \{(j, k) \mid j \in \mathbb{Z} \text{ with } j > -J, k \in \{0, \dots, K-1\}\}$ labeling a collection of scales j and directions k .

The following result gives a so-called Littlewood-Paley condition [53], [69] for the collection $\Psi_\Lambda = \{T_b I g_\lambda\}_{(\lambda, b) \in \Lambda \times \mathbb{R}^d}$ to form a semi-discrete frame.

Proposition 2: Let Λ be a countable set. The collection $\Psi_\Lambda = \{T_b I g_\lambda\}_{(\lambda, b) \in \Lambda \times \mathbb{R}^d}$ with atoms $\{g_\lambda\}_{\lambda \in \Lambda} \subseteq L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ is a semi-discrete frame for $L^2(\mathbb{R}^d)$ with frame bounds $A, B > 0$ if and only if

$$A \leq \sum_{\lambda \in \Lambda} |\widehat{g}_\lambda(\omega)|^2 \leq B, \quad a.e. \omega \in \mathbb{R}^d. \quad (32)$$

Proof: The proof is standard and can be found, e.g., in [30, Th. 5.11]. \square

Remark 2: What is behind Proposition 2 is a result on the unitary equivalence between operators [70, Def. 5.19.3]. Specifically, Proposition 2 follows from the fact that the multiplier $\sum_{\lambda \in \Lambda} |\widehat{g}_\lambda|^2$ is unitarily equivalent to the frame operator S_Λ in (31) according to

$$\mathcal{F} S_\Lambda \mathcal{F}^{-1} = \sum_{\lambda \in \Lambda} |\widehat{g}_\lambda|^2,$$

where $\mathcal{F} : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$ denotes the Fourier transform. We refer the interested reader to [71] where the framework of unitary equivalence was formalized in the context of shift-invariant frames for $\ell^2(\mathbb{Z})$.

The following proposition states normalization results for semi-discrete frames that come in handy in satisfying the admissibility condition (17) as discussed in Section III.

Proposition 3: Let $\Psi_\Lambda = \{T_b I g_\lambda\}_{(\lambda, b) \in \Lambda \times \mathbb{R}^d}$ be a semi-discrete frame for $L^2(\mathbb{R}^d)$ with frame bounds A, B .

- i) For $C > 0$, the family of functions $\tilde{\Psi}_\Lambda := \{T_b I \tilde{g}_\lambda\}_{(\lambda, b) \in \Lambda \times \mathbb{R}^d}$, $\tilde{g}_\lambda := C^{-1/2} g_\lambda$, $\forall \lambda \in \Lambda$, is a semi-discrete frame for $L^2(\mathbb{R}^d)$ with frame bounds $\tilde{A} := \frac{A}{C}$ and $\tilde{B} := \frac{B}{C}$.

- ii) The family of functions $\Psi_\Lambda^\natural := \{T_b I g_\lambda^\natural\}_{(\lambda, b) \in \Lambda \times \mathbb{R}^d}$,

$$g_\lambda^\natural := \mathcal{F}^{-1} \left(\widehat{g}_\lambda \left(\sum_{\lambda' \in \Lambda} |\widehat{g}_{\lambda'}|^2 \right)^{-1/2} \right), \quad \forall \lambda \in \Lambda,$$

is a semi-discrete Parseval frame for $L^2(\mathbb{R}^d)$, i.e., the frame bounds satisfy $A^\natural = B^\natural = 1$.

Proof: We start by proving statement i). As Ψ_Λ is a frame for $L^2(\mathbb{R}^d)$, we have

$$A \|f\|_2^2 \leq \sum_{\lambda \in \Lambda} \|f * g_\lambda\|_2^2 \leq B \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d). \quad (33)$$

With $g_\lambda = \sqrt{C} \tilde{g}_\lambda$, for all $\lambda \in \Lambda$, in (33) we get $A \|f\|_2^2 \leq \sum_{\lambda \in \Lambda} \|f * \sqrt{C} \tilde{g}_\lambda\|_2^2 \leq B \|f\|_2^2$, for all $f \in L^2(\mathbb{R}^d)$, which is equivalent to $\frac{A}{C} \|f\|_2^2 \leq \sum_{\lambda \in \Lambda} \|f * \tilde{g}_\lambda\|_2^2 \leq \frac{B}{C} \|f\|_2^2$, for all $f \in L^2(\mathbb{R}^d)$, and hence establishes i). To prove statement ii), we first note that $\mathcal{F} g_\lambda^\natural = \widehat{g}_\lambda \left(\sum_{\lambda' \in \Lambda} |\widehat{g}_{\lambda'}|^2 \right)^{-1/2}$, for all $\lambda \in \Lambda$, and thus $\sum_{\lambda \in \Lambda} |\mathcal{F} g_\lambda^\natural(\omega)|^2 = \sum_{\lambda \in \Lambda} |\widehat{g}_\lambda(\omega)|^2 \left(\sum_{\lambda' \in \Lambda} |\widehat{g}_{\lambda'}(\omega)|^2 \right)^{-1} = 1$, a.e. $\omega \in \mathbb{R}^d$. Application of Proposition 2 then establishes that Ψ_Λ^\natural is a semi-discrete Parseval frame for $L^2(\mathbb{R}^d)$, i.e., the frame bounds satisfy $A^\natural = B^\natural = 1$. \square

APPENDIX B

EXAMPLES OF SEMI-DISCRETE FRAMES IN 1-D

General 1-D semi-discrete frames are given by collections

$$\Psi = \{T_b I g_k\}_{(k, b) \in \mathbb{Z} \times \mathbb{R}} \quad (34)$$

with atoms $g_k \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, indexed by the integers $\Lambda = \mathbb{Z}$, and satisfying the Littlewood-Paley condition

$$A \leq \sum_{k \in \mathbb{Z}} |\widehat{g}_k(\omega)|^2 \leq B, \quad a.e. \omega \in \mathbb{R}. \quad (35)$$

The structural example frames we consider are Weyl-Heisenberg (Gabor) frames where the g_k are obtained through modulation from a prototype function, and wavelet frames where the g_k are obtained through scaling from a mother wavelet.

Semi-discrete Weyl-Heisenberg (Gabor) frames: Weyl-Heisenberg frames [72]–[75] are well-suited to the extraction of sinusoidal features [76], and have been applied successfully in various practical feature extraction tasks [54], [77]. A semi-discrete Weyl-Heisenberg frame for $L^2(\mathbb{R})$ is a collection of functions according to (34), where $g_k(x) := e^{2\pi i k x} g(x)$, $k \in \mathbb{Z}$, with the prototype function $g \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. The atoms $\{g_k\}_{k \in \mathbb{Z}}$ satisfy the Littlewood-Paley condition (35) according to

$$A \leq \sum_{k \in \mathbb{Z}} |\widehat{g}(\omega - k)|^2 \leq B, \quad a.e. \omega \in \mathbb{R}. \quad (36)$$

A popular function $g \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ satisfying (36) is the Gaussian function [74].

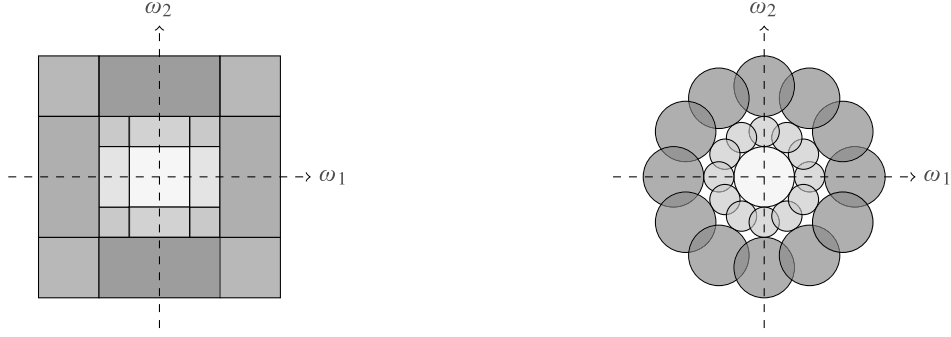


Fig. 7. Partitioning of the frequency plane \mathbb{R}^2 induced by (left) a semi-discrete tensor wavelet frame, and (right) a semi-discrete directional wavelet frame.

Semi-discrete wavelet frames: Wavelets are well-suited to the extraction of signal features characterized by singularities [31], [53], and have been applied successfully in various practical feature extraction tasks [55], [56]. A semi-discrete wavelet frame for $L^2(\mathbb{R})$ is a collection of functions according to (34), where $g_k(x) := 2^k \psi(2^k x)$, $k \in \mathbb{Z}$, with the mother wavelet $\psi \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. The atoms $\{g_k\}_{k \in \mathbb{Z}}$ satisfy the Littlewood-Paley condition (35) according to

$$A \leq \sum_{k \in \mathbb{Z}} |\widehat{\psi}(2^{-k} \omega)|^2 \leq B, \quad a.e. \omega \in \mathbb{R}. \quad (37)$$

A large class of functions ψ satisfying (37) can be obtained through a multi-resolution analysis in $L^2(\mathbb{R})$ [30, Def. 7.1].

APPENDIX C

EXAMPLES OF SEMI-DISCRETE FRAMES IN 2-D

Semi-discrete wavelet frames: Two-dimensional wavelets are well-suited to the extraction of signal features characterized by point singularities (such as, e.g., stars in astronomical images [78]), and have been applied successfully in various practical feature extraction tasks, e.g., in [19]–[21] and [32]. Prominent families of two-dimensional wavelet frames are tensor wavelet frames and directional wavelet frames:

- i) *Semi-discrete tensor wavelet frames:* A semi-discrete tensor wavelet frame for $L^2(\mathbb{R}^2)$ is a collection of functions according to $\Psi_{\Lambda_{TW}} := \{T_b I g(e, j)\}_{(e, j) \in \Lambda_{TW}, b \in \mathbb{R}^2}$, $g(e, j)(x) := 2^{2j} \psi^e(2^j x)$, where $\Lambda_{TW} := \{(0, 0), 0\} \cup \{(e, j) \mid e \in E \setminus \{(0, 0)\}, j \geq 0\}$, and $E := \{0, 1\}^2$. Here, the functions $\psi^e \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$ are tensor products of a coarse-scale function $\phi \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ and a fine-scale function $\psi \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ according to $\psi^{(0,0)} := \phi \otimes \phi$, $\psi^{(1,0)} := \psi \otimes \phi$, $\psi^{(0,1)} := \phi \otimes \psi$, and $\psi^{(1,1)} := \psi \otimes \psi$. The corresponding Littlewood-Paley condition (32) reads

$$A \leq |\widehat{\psi^{(0,0)}}(\omega)|^2 + \sum_{j \geq 0} \sum_{e \in E \setminus \{(0,0)\}} |\widehat{\psi^e}(2^{-j} \omega)|^2 \leq B, \quad (38)$$

a.e. $\omega \in \mathbb{R}^2$. A large class of functions ϕ, ψ satisfying (38) can be obtained through a multi-resolution analysis in $L^2(\mathbb{R})$ [30, Def. 7.1].

- ii) *Semi-discrete directional wavelet frames:* A semi-discrete directional wavelet frame for $L^2(\mathbb{R}^2)$ is a collection of functions according to

$$\Psi_{\Lambda_{DW}} := \{T_b I g(j, k)\}_{(j, k) \in \Lambda_{DW}, b \in \mathbb{R}^2},$$

with $g_{(-J,0)}(x) := 2^{-2J} \phi(2^{-J} x)$, $g_{(j,k)}(x) := 2^{2j} \psi(2^j R_{\theta_k} x)$, where $\Lambda_{DW} := \{(-J, 0)\} \cup \{(j, k) \mid j \in \mathbb{Z} \text{ with } j > -J, k \in \{0, \dots, K-1\}\}$, R_{θ} is a 2×2 rotation matrix defined as

$$R_{\theta} := \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}, \quad \theta \in [0, 2\pi), \quad (39)$$

and $\theta_k := \frac{2\pi k}{K}$, with $k = 0, \dots, K-1$, for a fixed $K \in \mathbb{N}$, are rotation angles. The functions $\phi \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$ and $\psi \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$ are referred to in the literature as coarse-scale wavelet and fine-scale wavelet, respectively. The integer $J \in \mathbb{Z}$ corresponds to the coarsest scale resolved and the atoms $\{g_{(j,k)}\}_{(j,k) \in \Lambda_{DW}}$ satisfy the Littlewood-Paley condition (32) according to

$$A \leq |\widehat{\phi}(2^J \omega)|^2 + \sum_{j > -J} \sum_{k=0}^{K-1} |\widehat{\psi}(2^{-j} R_{\theta_k} \omega)|^2 \leq B, \quad (40)$$

a.e. $\omega \in \mathbb{R}^2$. Prominent examples of functions ϕ, ψ satisfying (40) are the Gaussian function for ϕ and a modulated Gaussian function for ψ [30].

Semi-discrete curvelet frames: Curvelets, introduced in [34] and [38], are well-suited to the extraction of signal features characterized by curve-like singularities (such as, e.g., curved edges in images), and have been applied successfully in various practical feature extraction tasks [60], [61].

A semi-discrete curvelet frame for $L^2(\mathbb{R}^2)$ is a collection of functions according to $\Psi_{\Lambda_C} := \{T_b I g(j, l)\}_{(j, l) \in \Lambda_C, b \in \mathbb{R}^2}$, with $g_{(-1,0)}(x) := \phi(x)$, $g_{(j,l)}(x) := \psi_j(R_{\theta_{j,l}} x)$, where $\Lambda_C := \{(-1, 0)\} \cup \{(j, l) \mid j \geq 0, l = 0, \dots, L_j - 1\}$, $R_{\theta} \in \mathbb{R}^{2 \times 2}$ is the rotation matrix defined in (39), and $\theta_{j,l} := \pi l 2^{-\lceil j/2 \rceil - 1}$, for $j \geq 0$, and $0 \leq l < L_j := 2^{\lceil j/2 \rceil + 2}$, are scale-dependent rotation angles. The functions $\phi \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$ and $\psi_j \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$ satisfy the Littlewood-Paley condition (32) according to

$$A \leq |\widehat{\phi}(\omega)|^2 + \sum_{j=0}^{\infty} \sum_{l=0}^{L_j-1} |\widehat{\psi_j}(R_{\theta_{j,l}} \omega)|^2 \leq B, \quad (41)$$

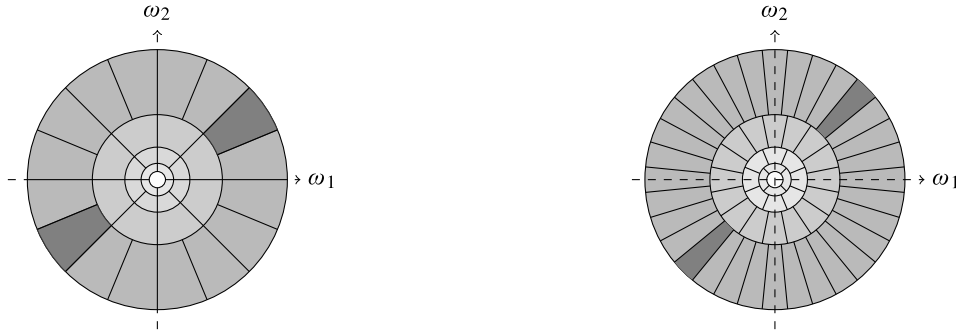


Fig. 8. Partitioning of the frequency plane \mathbb{R}^2 induced by (left) a semi-discrete curvelet frame, and (right) a semi-discrete ridgelet frame.

a.e. $\omega \in \mathbb{R}^2$. The ψ_j , $j \geq 0$, are designed to have their Fourier transforms $\widehat{\psi}_j$ supported on a pair of opposite wedges of size $2^{-j/2} \times 2^j$ in the dyadic corona $\{\omega \in \mathbb{R}^2 \mid 2^j \leq |\omega| \leq 2^{j+1}\}$, see Fig. 8 (left). We refer the reader to [34, Th. 4.1] for constructions of functions ϕ, ψ_j satisfying (41) with $A = B = 1$.

Semi-discrete ridgelet frames: Ridgelets, introduced in [79] and [80], are well-suited to the extraction of signal features characterized by straight-line singularities (such as, e.g., straight edges in images), and have been applied successfully in various practical feature extraction tasks [57]–[59], [61].

A semi-discrete ridgelet frame for $L^2(\mathbb{R}^2)$ is a collection of functions according to $\Psi_{\Lambda_R} := \{T_b I g_{(j,l)}\}_{(j,l) \in \Lambda_R, b \in \mathbb{R}^2}$, with $g_{(0,0)}(x) := \phi(x)$, $g_{(j,l)}(x) := \psi_{(j,l)}(x)$, where $\Lambda_R := \{(0,0)\} \cup \{(j,l) \mid j \geq 1, l = 1, \dots, 2^j - 1\}$, and the atoms $\{g_{(j,l)}\}_{(j,l) \in \Lambda_R}$ satisfy the Littlewood-Paley condition (32) according to

$$A \leq |\widehat{\phi}(\omega)|^2 + \sum_{j=1}^{\infty} \sum_{l=1}^{2^j-1} |\widehat{\psi}_{(j,l)}(\omega)|^2 \leq B, \quad (42)$$

a.e. $\omega \in \mathbb{R}^2$. The $\psi_{(j,l)} \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$, $(j,l) \in \Lambda_R \setminus \{(0,0)\}$, are designed to be constant in the direction specified by the parameter l , and to have Fourier transforms $\widehat{\psi}_{(j,l)}$ supported on a pair of opposite wedges of size $2^{-j} \times 2^j$ in the dyadic corona $\{\omega \in \mathbb{R}^2 \mid 2^j \leq |\omega| \leq 2^{j+1}\}$, see Fig. 8 (right). We refer the reader to [37, Prop. 6] for constructions of functions $\phi, \psi_{(j,l)}$ satisfying (42) with $A = B = 1$.

Remark 3: For further examples of interesting structured semi-discrete frames, we refer to [36], which discusses semi-discrete shearlet frames, and [35], which deals with semi-discrete α -curvelet frames.

APPENDIX D NON-LINEARITIES

This appendix gives a brief overview of non-linearities $M : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$ that are widely used in the deep learning literature and that fit into our theory. For each example, we establish how it satisfies the conditions on $M : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$ in Theorems 1 and 2 and in Corollary 1. Specifically, we need to verify the following:

- (i) Lipschitz continuity: There exists a constant $L \geq 0$ such that $\|Mf - Mh\|_2 \leq L\|f - h\|_2$, for all $f, h \in L^2(\mathbb{R}^d)$.

- (ii) $Mf = 0$ for $f = 0$.

All non-linearities considered here are pointwise (memoryless) operators in the sense of

$$M : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d), \quad (Mf)(x) = \rho(f(x)), \quad (43)$$

where $\rho : \mathbb{C} \rightarrow \mathbb{C}$. An immediate consequence of this property is that the operator M commutes with the translation operator T_t (see Theorem 2 and Corollary 1):

$$\begin{aligned} (MT_t f)(x) &= \rho((T_t f)(x)) = \rho(f(x-t)) = T_t \rho(f(x)) \\ &= (T_t Mf)(x), \quad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}^d. \end{aligned}$$

Modulus function: The modulus function

$$|\cdot| : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d), \quad |f|(x) := |f(x)|,$$

has been applied successfully in the deep learning literature, e.g., in [16] and [21], and most prominently in scattering networks [22]. Lipschitz continuity with $L = 1$ follows from

$$\begin{aligned} \| |f| - |h| \|_2^2 &= \int_{\mathbb{R}^d} ||f(x)| - |h(x)||^2 dx \\ &\leq \int_{\mathbb{R}^d} |f(x) - h(x)|^2 dx = \|f - h\|_2^2, \end{aligned}$$

for $f, h \in L^2(\mathbb{R}^d)$, by the reverse triangle inequality. Furthermore, obviously $|f| = 0$ for $f = 0$, and finally $|\cdot|$ is pointwise as (43) is satisfied with $\rho(x) := |x|$.

Rectified linear unit: The rectified linear unit non-linearity (see, e.g., [26], [27]) is defined as $R : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$,

$$(Rf)(x) := \max\{0, \operatorname{Re}(f(x))\} + i \max\{0, \operatorname{Im}(f(x))\}.$$

We start by establishing that R is Lipschitz-continuous with $L = 2$. To this end, fix $f, h \in L^2(\mathbb{R}^d)$. We have

$$\begin{aligned} |(Rf)(x) - (Rh)(x)| &= |\max\{0, \operatorname{Re}(f(x))\} + i \max\{0, \operatorname{Im}(f(x))\} \\ &\quad - (\max\{0, \operatorname{Re}(h(x))\} + i \max\{0, \operatorname{Im}(h(x))\})| \\ &\leq |\max\{0, \operatorname{Re}(f(x))\} - \max\{0, \operatorname{Re}(h(x))\}| \\ &\quad + |\max\{0, \operatorname{Im}(f(x))\} - \max\{0, \operatorname{Im}(h(x))\}| \end{aligned} \quad (44)$$

$$\begin{aligned} &\leq |\operatorname{Re}(f(x)) - \operatorname{Re}(h(x))| + |\operatorname{Im}(f(x)) - \operatorname{Im}(h(x))| \end{aligned} \quad (45)$$

$$\leq |f(x) - h(x)| + |f(x) - h(x)| = 2|f(x) - h(x)|, \quad (46)$$

where we used the triangle inequality in (44),

$$|\max\{0, a\} - \max\{0, b\}| \leq |a - b|, \quad \forall a, b \in \mathbb{R},$$

in (45), and the Lipschitz continuity (with $L = 1$) of the mappings $\text{Re} : \mathbb{C} \rightarrow \mathbb{R}$ and $\text{Im} : \mathbb{C} \rightarrow \mathbb{R}$ in (46). We therefore get

$$\begin{aligned} \|Rf - Rh\|_2 &= \left(\int_{\mathbb{R}^d} |(Rf)(x) - (Rh)(x)|^2 dx \right)^{1/2} \\ &\leq 2 \left(\int_{\mathbb{R}^d} |f(x) - h(x)|^2 dx \right)^{1/2} \\ &= 2 \|f - h\|_2, \end{aligned}$$

which establishes Lipschitz continuity of R with Lipschitz constant $L = 2$. Furthermore, obviously $Rf = 0$ for $f = 0$, and finally (43) is satisfied with $\rho(x) := \max\{0, \text{Re}(x)\} + i \max\{0, \text{Im}(x)\}$.

Hyperbolic tangent: The hyperbolic tangent non-linearity (see, e.g., [15]–[17]) is defined as $H : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$,

$$(Hf)(x) := \tanh(\text{Re}(f(x))) + i \tanh(\text{Im}(f(x))),$$

where $\tanh(x) := \frac{e^x - e^{-x}}{e^x + e^{-x}}$. We start by proving that H is Lipschitz-continuous with $L = 2$. To this end, fix $f, h \in L^2(\mathbb{R}^d)$. We have

$$\begin{aligned} |(Hf)(x) - (Hh)(x)| &= |\tanh(\text{Re}(f(x))) + i \tanh(\text{Im}(f(x))) \\ &\quad - (\tanh(\text{Re}(h(x))) + i \tanh(\text{Im}(h(x))))| \\ &\leq |\tanh(\text{Re}(f(x))) - \tanh(\text{Re}(h(x)))| \\ &\quad + |\tanh(\text{Im}(f(x))) - \tanh(\text{Im}(h(x)))|, \end{aligned} \quad (47)$$

where, again, we used the triangle inequality. In order to further upper-bound (47), we show that \tanh is Lipschitz-continuous. To this end, we make use of the following result.

Lemma 1: Let $h : \mathbb{R} \rightarrow \mathbb{R}$ be a continuously differentiable function satisfying $\sup_{x \in \mathbb{R}} |h'(x)| \leq L$. Then, h is Lipschitz-continuous with Lipschitz constant L .

Proof: See [81, Th. 9.5.1]. \square

Since $\tanh'(x) = 1 - \tanh^2(x)$, $x \in \mathbb{R}$, we have $\sup_{x \in \mathbb{R}} |\tanh'(x)| \leq 1$. By Lemma 1 we can therefore conclude that \tanh is Lipschitz-continuous with $L = 1$, which when used in (47), yields

$$\begin{aligned} |(Hf)(x) - (Hh)(x)| &\leq |\text{Re}(f(x)) - \text{Re}(h(x))| \\ &\quad + |\text{Im}(f(x)) - \text{Im}(h(x))| \\ &\leq |f(x) - h(x)| + |f(x) - h(x)| \\ &= 2|f(x) - h(x)|. \end{aligned}$$

Here, again, we used the Lipschitz continuity (with $L = 1$) of $\text{Re} : \mathbb{C} \rightarrow \mathbb{R}$ and $\text{Im} : \mathbb{C} \rightarrow \mathbb{R}$. Putting things together, we obtain

$$\begin{aligned} \|Hf - Hh\|_2 &= \left(\int_{\mathbb{R}^d} |(Hf)(x) - (Hh)(x)|^2 dx \right)^{1/2} \\ &\leq 2 \left(\int_{\mathbb{R}^d} |f(x) - h(x)|^2 dx \right)^{1/2} \\ &= 2 \|f - h\|_2, \end{aligned}$$

which proves that H is Lipschitz-continuous with $L = 2$. Since $\tanh(0) = 0$, we trivially have $Hf = 0$ for $f = 0$. Finally, (43) is satisfied with $\rho(x) := \tanh(\text{Re}(x)) + i \tanh(\text{Im}(x))$.

Shifted logistic sigmoid: The shifted logistic sigmoid non-linearity¹⁰ (see, e.g., [28], [29]) is defined as $P : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$,

$$(Pf)(x) := \text{sig}(\text{Re}(f(x))) + i \text{sig}(\text{Im}(f(x))),$$

where $\text{sig}(x) := \frac{1}{1+e^{-x}} - \frac{1}{2}$. We first establish that P is Lipschitz-continuous with $L = \frac{1}{2}$. To this end, fix $f, h \in L^2(\mathbb{R}^d)$. We have

$$\begin{aligned} |(Pf)(x) - (Ph)(x)| &= |\text{sig}(\text{Re}(f(x))) + i \text{sig}(\text{Im}(f(x))) \\ &\quad - (\text{sig}(\text{Re}(h(x))) + i \text{sig}(\text{Im}(h(x))))| \\ &\leq |\text{sig}(\text{Re}(f(x))) - \text{sig}(\text{Re}(h(x)))| \\ &\quad + |\text{sig}(\text{Im}(f(x))) - \text{sig}(\text{Im}(h(x)))|, \end{aligned} \quad (48)$$

where, again, we employed the triangle inequality. As before, to further upper-bound (48), we show that sig is Lipschitz-continuous. Specifically, we apply Lemma 1 with $\text{sig}'(x) = \frac{e^{-x}}{(1+e^{-x})^2}$, $x \in \mathbb{R}$, and hence $\sup_{x \in \mathbb{R}} |\text{sig}'(x)| \leq \frac{1}{4}$, to conclude that sig is Lipschitz-continuous with $L = \frac{1}{4}$. When used in (48) this yields (together with the Lipschitz continuity, with $L = 1$, of $\text{Re} : \mathbb{C} \rightarrow \mathbb{R}$ and $\text{Im} : \mathbb{C} \rightarrow \mathbb{R}$)

$$\begin{aligned} |(Pf)(x) - (Ph)(x)| &\leq \frac{1}{4} |\text{Re}(f(x)) - \text{Re}(h(x))| \\ &\quad + \frac{1}{4} |\text{Im}(f(x)) - \text{Im}(h(x))| \leq \frac{1}{4} |f(x) - h(x)| \\ &\quad + \frac{1}{4} |f(x) - h(x)| = \frac{1}{2} |f(x) - h(x)|. \end{aligned} \quad (49)$$

It now follows from (49) that

$$\begin{aligned} \|Pf - Ph\|_2 &= \left(\int_{\mathbb{R}^d} |(Pf)(x) - (Ph)(x)|^2 dx \right)^{1/2} \\ &\leq \frac{1}{2} \left(\int_{\mathbb{R}^d} |f(x) - h(x)|^2 dx \right)^{1/2} \\ &= \frac{1}{2} \|f - h\|_2, \end{aligned}$$

which establishes Lipschitz continuity of P with $L = \frac{1}{2}$. Since $\text{sig}(0) = 0$, we trivially have $Pf = 0$ for $f = 0$. Finally, (43) is satisfied with $\rho(x) := \text{sig}(\text{Re}(x)) + i \text{sig}(\text{Im}(x))$.

APPENDIX E PROOF OF PROPOSITION 1

We need to show that $\Phi_\Omega(f) \in (L^2(\mathbb{R}^d))^{\mathcal{Q}}$, for all $f \in L^2(\mathbb{R}^d)$. This will be accomplished by proving an even stronger result, namely

$$\|\Phi_\Omega(f)\| \leq \|f\|_2, \quad \forall f \in L^2(\mathbb{R}^d), \quad (50)$$

which, by $\|f\|_2 < \infty$, establishes the claim. For ease of notation, we let $f_q := U[q]f$, for $f \in L^2(\mathbb{R}^d)$, in the following. Thanks to (14) and (17), we have $\|f_q\|_2 \leq \|f\|_2 < \infty$,

¹⁰Strictly speaking, it is actually the sigmoid function $x \mapsto \frac{1}{1+e^{-x}}$ rather than the shifted sigmoid function $x \mapsto \frac{1}{1+e^{-x}} - \frac{1}{2}$ that is used in [28] and [29]. We incorporated the offset $\frac{1}{2}$ in order to satisfy the requirement $Pf = 0$ for $f = 0$.

and thus $f_q \in L^2(\mathbb{R}^d)$. The key idea of the proof is now—similarly to the proof of [22, Prop. 2.5]—to judiciously employ a telescoping series argument. We start by writing

$$\begin{aligned} |||\Phi_\Omega(f)|||^2 &= \sum_{n=0}^{\infty} \sum_{q \in \Lambda_1^n} \|f_q * \chi_n\|_2^2 \\ &= \lim_{N \rightarrow \infty} \sum_{n=0}^N \underbrace{\sum_{q \in \Lambda_1^n} \|f_q * \chi_n\|_2^2}_{:=a_n}. \end{aligned} \quad (51)$$

The key step is then to establish that a_n can be upper-bounded according to

$$a_n \leq b_n - b_{n+1}, \quad \forall n \in \mathbb{N}_0, \quad (52)$$

with $b_n := \sum_{q \in \Lambda_1^n} \|f_q\|_2^2$, $n \in \mathbb{N}_0$, and to use this result in a telescoping series argument according to

$$\begin{aligned} \sum_{n=0}^N a_n &\leq \sum_{n=0}^N (b_n - b_{n+1}) = (b_0 - b_1) + (b_1 - b_2) \\ &\quad + \cdots + (b_N - b_{N+1}) = b_0 - \underbrace{b_{N+1}}_{\geq 0} \end{aligned} \quad (53)$$

$$\leq b_0 = \sum_{q \in \Lambda_1^0} \|f_q\|_2^2 = \|U[e]f\|_2^2 = \|f\|_2^2. \quad (54)$$

By (51) this then implies (50). We start by noting that (52) reads

$$\sum_{q \in \Lambda_1^n} \|f_q * \chi_n\|_2^2 \leq \sum_{q \in \Lambda_1^n} \|f_q\|_2^2 - \sum_{q \in \Lambda_1^{n+1}} \|f_q\|_2^2, \quad (55)$$

for all $n \in \mathbb{N}_0$, and proceed by examining the second term on the right hand side (RHS) of (55). Every path

$$\tilde{q} \in \Lambda_1^{n+1} = \underbrace{\Lambda_1 \times \cdots \times \Lambda_n}_{=\Lambda_1^n} \times \Lambda_{n+1}$$

of length $n+1$ can be decomposed into a path $q \in \Lambda_1^n$ of length n and an index $\lambda_{n+1} \in \Lambda_{n+1}$ according to $\tilde{q} = (q, \lambda_{n+1})$. Thanks to (13) we have

$$U[\tilde{q}] = U[(q, \lambda_{n+1})] = U_{n+1}[\lambda_{n+1}]U[q],$$

which yields

$$\sum_{\tilde{q} \in \Lambda_1^{n+1}} \|f_{\tilde{q}}\|_2^2 = \sum_{q \in \Lambda_1^n} \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|U_{n+1}[\lambda_{n+1}]f_q\|_2^2. \quad (56)$$

Substituting the second term on the RHS of (55) by (56) now yields

$$\begin{aligned} &\sum_{q \in \Lambda_1^n} \|f_q * \chi_n\|_2^2 \\ &\leq \sum_{q \in \Lambda_1^n} \left(\|f_q\|_2^2 - \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|U_{n+1}[\lambda_{n+1}]f_q\|_2^2 \right), \quad \forall n \in \mathbb{N}_0, \end{aligned}$$

which can be rewritten as

$$\begin{aligned} &\sum_{q \in \Lambda_1^n} \left(\|f_q * \chi_n\|_2^2 + \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|U_{n+1}[\lambda_{n+1}]f_q\|_2^2 \right) \\ &\leq \sum_{q \in \Lambda_1^n} \|f_q\|_2^2, \quad \forall n \in \mathbb{N}_0. \end{aligned} \quad (57)$$

Next, note that the second term inside the sum on the left hand side (LHS) of (57) can be written as

$$\begin{aligned} &\sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|U_{n+1}[\lambda_{n+1}]f_q\|_2^2 \\ &= \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \int_{\mathbb{R}^d} |(U_{n+1}[\lambda_{n+1}]f_q)(x)|^2 dx \\ &= \sum_{\lambda_{n+1} \in \Lambda_{n+1}} S_{n+1}^d \int_{\mathbb{R}^d} \left| P_{n+1}(M_{n+1}(f_q * g_{\lambda_{n+1}}))(S_{n+1}x) \right|^2 dx \\ &= \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \int_{\mathbb{R}^d} \left| P_{n+1}(M_{n+1}(f_q * g_{\lambda_{n+1}}))(y) \right|^2 dy \\ &= \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|P_{n+1}(M_{n+1}(f_q * g_{\lambda_{n+1}}))\|_2^2, \end{aligned} \quad (58)$$

for all $n \in \mathbb{N}_0$. Noting that $f_q \in L^2(\mathbb{R}^d)$, as established above, and $g_{\lambda_{n+1}} \in L^1(\mathbb{R}^d)$, by assumption, it follows that $(f_q * g_{\lambda_{n+1}}) \in L^2(\mathbb{R}^d)$ thanks to Young's inequality [63, Th. 1.2.12]. We use the Lipschitz property of M_{n+1} and P_{n+1} , i.e., $\|M_{n+1}(f_q * g_{\lambda_{n+1}}) - M_{n+1}h\|_2 \leq L_{n+1}\|f_q * g_{\lambda_{n+1}} - h\|$, and $\|P_{n+1}(f_q * g_{\lambda_{n+1}}) - P_{n+1}h\|_2 \leq R_{n+1}\|f_q * g_{\lambda_{n+1}} - h\|$, together with $M_{n+1}h = 0$ and $P_{n+1}h = 0$ for $h = 0$, to upper-bound the term inside the sum in (58) according to

$$\begin{aligned} &\|P_{n+1}(M_{n+1}(f_q * g_{\lambda_{n+1}}))\|_2^2 \leq R_{n+1}^2 \|M_{n+1}(f_q * g_{\lambda_{n+1}})\|_2^2 \\ &\leq L_{n+1}^2 R_{n+1}^2 \|f_q * g_{\lambda_{n+1}}\|_2^2, \quad \forall n \in \mathbb{N}_0. \end{aligned} \quad (59)$$

Substituting the second term inside the sum on the LHS of (57) by the upper bound resulting from insertion of (59) into (58) yields

$$\begin{aligned} &\sum_{q \in \Lambda_1^n} \left(\|f_q * \chi_n\|_2^2 + L_{n+1}^2 R_{n+1}^2 \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|f_q * g_{\lambda_{n+1}}\|_2^2 \right) \\ &\leq \sum_{q \in \Lambda_1^n} \max\{1, L_{n+1}^2 R_{n+1}^2\} \left(\|f_q * \chi_n\|_2^2 \right. \\ &\quad \left. + \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|f_q * g_{\lambda_{n+1}}\|_2^2 \right), \quad \forall n \in \mathbb{N}_0. \end{aligned} \quad (60)$$

As the functions $\{g_{\lambda_{n+1}}\}_{\lambda_{n+1} \in \Lambda_{n+1}} \cup \{\chi_n\}$ are the atoms of the semi-discrete frame Ψ_{n+1} for $L^2(\mathbb{R}^d)$ and $f_q \in L^2(\mathbb{R}^d)$, as established above, we have

$$\|f_q * \chi_n\|_2^2 + \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|f_q * g_{\lambda_{n+1}}\|_2^2 \leq B_{n+1} \|f_q\|_2^2,$$

which, when used in (60) yields

$$\begin{aligned} &\sum_{q \in \Lambda_1^n} \left(\|f_q * \chi_n\|_2^2 + \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|U_{n+1}[\lambda_{n+1}]f_q\|_2^2 \right) \\ &\leq \sum_{q \in \Lambda_1^n} \max\{1, L_{n+1}^2 R_{n+1}^2\} B_{n+1} \|f_q\|_2^2 \\ &= \sum_{q \in \Lambda_1^n} \max\{B_{n+1}, B_{n+1} L_{n+1}^2 R_{n+1}^2\} \|f_q\|_2^2, \end{aligned} \quad (61)$$

for all $n \in \mathbb{N}_0$. Finally, invoking the assumption

$$\max\{B_n, B_n L_n^2 R_n^2\} \leq 1, \quad \forall n \in \mathbb{N},$$

in (61) yields (57) and thereby completes the proof.

APPENDIX F
PROOF OF THEOREM 1

We start by proving i). The key step in establishing (19) is to show that the operator U_n , $n \in \mathbb{N}$, defined in (10) satisfies the relation

$$U_n[\lambda_n]T_t f = T_{t/S_n} U_n[\lambda_n]f, \quad (62)$$

for all $f \in L^2(\mathbb{R}^d)$, $t \in \mathbb{R}^d$, and $\lambda_n \in \Lambda_n$. With the definition of $U[q]$ in (13) this then yields

$$U[q]T_t f = T_{t/(S_1 \dots S_n)} U[q]f, \quad (63)$$

for all $f \in L^2(\mathbb{R}^d)$, $t \in \mathbb{R}^d$, and $q \in \Lambda_1^n$. The identity (19) is then a direct consequence of (63) and the translation-covariance of the convolution operator:

$$\begin{aligned} \Phi_\Omega^n(T_t f) &= \{(U[q]T_t f) * \chi_n\}_{q \in \Lambda_1^n} \\ &= \{(T_{t/(S_1 \dots S_n)} U[q]f) * \chi_n\}_{q \in \Lambda_1^n} \\ &= \{T_{t/(S_1 \dots S_n)}((U[q]f) * \chi_n)\}_{q \in \Lambda_1^n} \\ &= T_{t/(S_1 \dots S_n)}\{(U[q]f) * \chi_n\}_{q \in \Lambda_1^n} \\ &= T_{t/(S_1 \dots S_n)} \Phi_\Omega^n(f), \quad \forall f \in L^2(\mathbb{R}^d), \quad \forall t \in \mathbb{R}^d. \end{aligned}$$

To establish (62), we first define the unitary operator

$$D_n : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d), \quad D_n f := S_n^{d/2} f(S_n \cdot),$$

and note that

$$\begin{aligned} U_n[\lambda_n]T_t f &= S_n^{d/2} P_n(M_n((T_t f) * g_{\lambda_n}))(S_n \cdot) \\ &= D_n P_n(M_n((T_t f) * g_{\lambda_n})) \\ &= D_n P_n(M_n(T_t(f * g_{\lambda_n}))) \\ &= D_n P_n(T_t(M_n(f * g_{\lambda_n}))) \quad (64) \\ &= D_n T_t(P_n((M_n(f * g_{\lambda_n})))) \quad (65) \end{aligned}$$

for all $f \in L^2(\mathbb{R}^d)$ and $t \in \mathbb{R}^d$, where in (64) and (65) we employed

$$M_n T_t = T_t M_n, \quad \text{and} \quad P_n T_t = T_t P_n,$$

for all $n \in \mathbb{N}$ and $t \in \mathbb{R}^d$, respectively, both of which are by assumption. Next, using

$$\begin{aligned} D_n T_t f &= S_n^{d/2} f(S_n \cdot - t) = S_n^{d/2} f(S_n(\cdot - t/S_n)) \\ &= T_{t/S_n} D_n f, \quad \forall f \in L^2(\mathbb{R}^d), \quad \forall t \in \mathbb{R}^d, \end{aligned}$$

in (65) yields

$$\begin{aligned} U_n[\lambda_n]T_t f &= D_n T_t(P_n((M_n(f * g_{\lambda_n})))) \\ &= T_{t/S_n}(D_n P_n((M_n(f * g_{\lambda_n})))) \\ &= T_{t/S_n} U_n[\lambda_n]f, \end{aligned}$$

for all $f \in L^2(\mathbb{R}^d)$ and $t \in \mathbb{R}^d$. This completes the proof of i).

Next, we prove ii). For ease of notation, again, we let $f_q := U[q]f$, for $f \in L^2(\mathbb{R}^d)$. Thanks to (14) and the admissibility

condition (17), we have $\|f_q\|_2 \leq \|f\|_2 < \infty$, and thus $f_q \in L^2(\mathbb{R}^d)$. We first write

$$\begin{aligned} &|||\Phi_\Omega^n(T_t f) - \Phi_\Omega^n(f)|||^2 \\ &= |||T_{t/(S_1 \dots S_n)} \Phi_\Omega^n(f) - \Phi_\Omega^n(f)|||^2 \quad (66) \end{aligned}$$

$$\begin{aligned} &= \sum_{q \in \Lambda_1^n} \|T_{t/(S_1 \dots S_n)}(f_q * \chi_n) - f_q * \chi_n\|_2^2 \\ &= \sum_{q \in \Lambda_1^n} \|M_{-t/(S_1 \dots S_n)}(\widehat{f_q * \chi_n}) - \widehat{f_q * \chi_n}\|_2^2, \quad (67) \end{aligned}$$

for all $n \in \mathbb{N}$, where in (66) we used (19), and in (67) we employed Parseval's formula [43, p. 189] (noting that $(f_q * \chi_n) \in L^2(\mathbb{R}^d)$ thanks to Young's inequality [63, Th. 1.2.12]) together with the relation

$$\widehat{T_t f} = M_{-t} \widehat{f}, \quad \forall f \in L^2(\mathbb{R}^d), \quad \forall t \in \mathbb{R}^d.$$

The key step is then to establish the upper bound

$$\begin{aligned} &\|M_{-t/(S_1 \dots S_n)}(\widehat{f_q * \chi_n}) - \widehat{f_q * \chi_n}\|_2^2 \\ &\leq \frac{4\pi^2 |t|^2 K^2}{(S_1 \dots S_n)^2} \|f_q\|_2^2, \quad \forall n \in \mathbb{N}, \quad (68) \end{aligned}$$

where $K > 0$ corresponds to the constant in the decay condition (20), and to note that

$$\sum_{q \in \Lambda_1^n} \|f_q\|_2^2 \leq \sum_{q \in \Lambda_1^{n-1}} \|f_q\|_2^2, \quad \forall n \in \mathbb{N}, \quad (69)$$

which follows from (52) thanks to

$$0 \leq \sum_{q \in \Lambda_1^{n-1}} \|f_q * \chi_{n-1}\|_2^2 = a_{n-1} \leq b_{n-1} - b_n \quad (70)$$

$$= \sum_{q \in \Lambda_1^{n-1}} \|f_q\|_2^2 - \sum_{q \in \Lambda_1^n} \|f_q\|_2^2, \quad \forall n \in \mathbb{N}. \quad (71)$$

Iterating on (69) yields

$$\begin{aligned} \sum_{q \in \Lambda_1^n} \|f_q\|_2^2 &\leq \sum_{q \in \Lambda_1^{n-1}} \|f_q\|_2^2 \leq \dots \leq \sum_{q \in \Lambda_1^0} \|f_q\|_2^2 \\ &= \|U[e]f\|_2^2 = \|f\|_2^2, \quad \forall n \in \mathbb{N}. \quad (72) \end{aligned}$$

The identity (67) together with the inequalities (68) and (72) then directly imply

$$|||\Phi_\Omega^n(T_t f) - \Phi_\Omega^n(f)|||^2 \leq \frac{4\pi^2 |t|^2 K^2}{(S_1 \dots S_n)^2} \|f\|_2^2, \quad (73)$$

for all $n \in \mathbb{N}$. It remains to prove (68). To this end, we first note that

$$\begin{aligned} &\|M_{-t/(S_1 \dots S_n)}(\widehat{f_q * \chi_n}) - \widehat{f_q * \chi_n}\|_2^2 \\ &= \int_{\mathbb{R}^d} |e^{-2\pi i \langle t, \omega \rangle / (S_1 \dots S_n)} - 1|^2 |\widehat{\chi_n}(\omega)|^2 |\widehat{f_q}(\omega)|^2 d\omega. \quad (74) \end{aligned}$$

Since $|e^{-2\pi i x} - 1| \leq 2\pi |x|$, for all $x \in \mathbb{R}$, it follows that

$$\begin{aligned} |e^{-2\pi i \langle t, \omega \rangle / (S_1 \dots S_n)} - 1|^2 &\leq \frac{4\pi^2 |\langle t, \omega \rangle|^2}{(S_1 \dots S_n)^2} \\ &\leq \frac{4\pi^2 |t|^2 |\omega|^2}{(S_1 \dots S_n)^2}, \quad (75) \end{aligned}$$

where in the last step we employed the Cauchy-Schwartz inequality. Substituting (75) into (74) yields

$$\begin{aligned} & \|M_{-t/(S_1 \dots S_n)}(\widehat{f_q * \chi_n}) - \widehat{f_q * \chi_n}\|_2^2 \\ & \leq \frac{4\pi^2 |t|^2}{(S_1 \dots S_n)^2} \int_{\mathbb{R}^d} |\omega|^2 |\widehat{\chi_n}(\omega)|^2 |\widehat{f_q}(\omega)|^2 d\omega \\ & \leq \frac{4\pi^2 |t|^2 K^2}{(S_1 \dots S_n)^2} \int_{\mathbb{R}^d} |\widehat{f_q}(\omega)|^2 d\omega \end{aligned} \quad (76)$$

$$= \frac{4\pi^2 |t|^2 K^2}{(S_1 \dots S_n)^2} \|\widehat{f_q}\|_2^2 = \frac{4\pi^2 |t|^2 K^2}{(S_1 \dots S_n)^2} \|f_q\|_2^2, \quad (77)$$

for all $n \in \mathbb{N}$, where in (76) we employed the decay condition (20), and in the last step, again, we used Parseval's formula [43, p. 189]. This establishes (68) and thereby completes the proof of ii).

APPENDIX G PROOF OF COROLLARY 1

The key idea of the proof is—similarly to the proof of ii) in Theorem 1—to upper-bound the deviation from perfect covariance in the frequency domain. For ease of notation, again, we let $f_q := U[q]f$, for $f \in L^2(\mathbb{R}^d)$. Thanks to (14) and the admissibility condition (17), we have $\|f_q\|_2 \leq \|f\|_2 < \infty$, and thus $f_q \in L^2(\mathbb{R}^d)$. We first write

$$\begin{aligned} & |||\Phi_\Omega^n(T_t f) - T_t \Phi_\Omega^n(f)|||^2 \\ & = |||T_{t/(S_1 \dots S_n)} \Phi_\Omega^n(f) - T_t \Phi_\Omega^n(f)|||^2 \end{aligned} \quad (78)$$

$$\begin{aligned} & = \sum_{q \in \Lambda_1^n} \|(T_{t/(S_1 \dots S_n)} - T_t)(f_q * \chi_n)\|_2^2 \\ & = \sum_{q \in \Lambda_1^n} \|(M_{-t/(S_1 \dots S_n)} - M_{-t})(\widehat{f_q * \chi_n})\|_2^2, \end{aligned} \quad (79)$$

for all $n \in \mathbb{N}$, where in (78) we used (19), and in (79) we employed Parseval's formula [43, p. 189] (noting that $(f_q * \chi_n) \in L^2(\mathbb{R}^d)$ thanks to Young's inequality [63, Th. 1.2.12]) together with the relation

$$\widehat{T_t f} = M_{-t} \widehat{f}, \quad \forall f \in L^2(\mathbb{R}^d), \quad \forall t \in \mathbb{R}^d.$$

The key step is then to establish the upper bound

$$\begin{aligned} & \|(M_{-t/(S_1 \dots S_n)} - M_{-t})(\widehat{f_q * \chi_n})\|_2^2 \\ & \leq 4\pi^2 |t|^2 K^2 |1/(S_1 \dots S_n) - 1|^2 \|f_q\|_2^2, \end{aligned} \quad (80)$$

where $K > 0$ corresponds to the constant in the decay condition (20). Arguments similar to those leading to (73) then complete the proof. It remains to prove (80):

$$\begin{aligned} & \|(M_{-t/(S_1 \dots S_n)} - M_{-t})(\widehat{f_q * \chi_n})\|_2^2 \\ & = \int_{\mathbb{R}^d} |e^{-2\pi i \langle t, \omega \rangle / (S_1 \dots S_n)} - e^{-2\pi i \langle t, \omega \rangle}|^2 |\widehat{\chi_n}(\omega)|^2 |\widehat{f_q}(\omega)|^2 d\omega. \end{aligned} \quad (81)$$

Since $|e^{-2\pi i x} - e^{-2\pi i y}| \leq 2\pi |x - y|$, for all $x, y \in \mathbb{R}$, it follows that

$$\begin{aligned} & |e^{-2\pi i \langle t, \omega \rangle / (S_1 \dots S_n)} - e^{-2\pi i \langle t, \omega \rangle}|^2 \\ & \leq 4\pi^2 |t|^2 |\omega|^2 |1/(S_1 \dots S_n) - 1|^2, \end{aligned} \quad (82)$$

where, again, we employed the Cauchy-Schwartz inequality. Substituting (82) into (81), and employing arguments similar to those leading to (77), establishes (80) and thereby completes the proof.

APPENDIX H PROOF OF THEOREM 2

As already mentioned at the beginning of Section IV-B, the proof of the deformation sensitivity bound (26) is based on two key ingredients. The first one, stated in Proposition 4 in Appendix I, establishes that the feature extractor Φ_Ω is Lipschitz-continuous with Lipschitz constant $L_\Omega = 1$, i.e.,

$$|||\Phi_\Omega(f) - \Phi_\Omega(h)||| \leq \|f - h\|_2, \quad \forall f, h \in L^2(\mathbb{R}^d), \quad (83)$$

and needs the admissibility condition (17) only. The second ingredient, stated in Proposition 5 in Appendix J, is an upper bound on the deformation error $\|f - F_{\tau, \omega} f\|_2$ given by

$$\|f - F_{\tau, \omega} f\|_2 \leq C(R\|\tau\|_\infty + \|\omega\|_\infty) \|f\|_2, \quad (84)$$

for all $f \in L_R^2(\mathbb{R}^d)$, and is valid under the assumptions $\omega \in C(\mathbb{R}^d, \mathbb{R})$ and $\tau \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ with $\|D\tau\|_\infty \leq \frac{1}{2d}$. We now show how (83) and (84) can be combined to establish the deformation sensitivity bound (26). To this end, we first apply (83) with $h := F_{\tau, \omega} f = e^{2\pi i \omega(\cdot)} f(\cdot - \tau(\cdot))$ to get

$$|||\Phi_\Omega(f) - \Phi_\Omega(F_{\tau, \omega} f)||| \leq \|f - F_{\tau, \omega} f\|_2, \quad (85)$$

for all $f \in L^2(\mathbb{R}^d)$. Here, we used $F_{\tau, \omega} f \in L^2(\mathbb{R}^d)$, which is thanks to

$$\|F_{\tau, \omega} f\|_2^2 = \int_{\mathbb{R}^d} |f(x - \tau(x))|^2 dx \leq 2\|f\|_2^2,$$

obtained through the change of variables $u = x - \tau(x)$, together with

$$\frac{du}{dx} = |\det(E - (D\tau)(x))| \geq 1 - d\|D\tau\|_\infty \geq 1/2, \quad (86)$$

for $x \in \mathbb{R}^d$. The first inequality in (86) follows from:

Lemma 2: [82, Cor. 1]: Let $M \in \mathbb{R}^{d \times d}$ be such that $|M_{i,j}| \leq a$, for all i, j with $1 \leq i, j \leq d$. If $da \leq 1$, then

$$|\det(E - M)| \geq 1 - da.$$

The second inequality in (86) is a consequence of the assumption $\|D\tau\|_\infty \leq \frac{1}{2d}$. The proof is finalized by replacing the RHS of (85) by the RHS of (84).

APPENDIX I PROPOSITION 4

Proposition 4: Let $\Omega = ((\Psi_n, M_n, P_n))_{n \in \mathbb{N}}$ be an admissible module-sequence. The corresponding feature extractor $\Phi_\Omega : L^2(\mathbb{R}^d) \rightarrow (L^2(\mathbb{R}^d))^\mathbb{Q}$ is Lipschitz-continuous with Lipschitz constant $L_\Omega = 1$, i.e.,

$$|||\Phi_\Omega(f) - \Phi_\Omega(h)||| \leq \|f - h\|_2, \quad \forall f, h \in L^2(\mathbb{R}^d). \quad (87)$$

Remark 4: Proposition 4 generalizes [22, Prop. 2.5], which shows that the wavelet-modulus feature extractor Φ_W generated by scattering networks is Lipschitz-continuous with Lipschitz constant $L_W = 1$. Specifically, our generalization

allows for general semi-discrete frames (i.e., general convolution kernels), general Lipschitz-continuous non-linearities M_n , and general Lipschitz-continuous operators P_n , all of which can be different in different layers. Moreover, thanks to the admissibility condition (17), the Lipschitz constant $L_\Omega = 1$ in (87) is completely independent of the frame upper bounds B_n and the Lipschitz-constants L_n and R_n of M_n and P_n , respectively.

Proof: The key idea of the proof is—similarly to the proof of Proposition 1 in Appendix E—to judiciously employ a telescoping series argument. For ease of notation, we let $f_q := U[q]f$ and $h_q := U[q]h$, for $f, h \in L^2(\mathbb{R}^d)$. Thanks to (14) and the admissibility condition (17), we have $\|f_q\|_2 \leq \|f\|_2 < \infty$ and $\|h_q\|_2 \leq \|h\|_2 < \infty$ and thus $f_q, h_q \in L^2(\mathbb{R}^d)$. We start by writing

$$\begin{aligned} |||\Phi_\Omega(f) - \Phi_\Omega(h)|||^2 &= \sum_{n=0}^{\infty} \sum_{q \in \Lambda_1^n} \|f_q * \chi_n - h_q * \chi_n\|_2^2 \\ &= \lim_{N \rightarrow \infty} \sum_{n=0}^N \underbrace{\sum_{q \in \Lambda_1^n} \|f_q * \chi_n - h_q * \chi_n\|_2^2}_{=: a_n}. \end{aligned}$$

As in the proof of Proposition 1 in Appendix E, the key step is to show that a_n can be upper-bounded according to

$$a_n \leq b_n - b_{n+1}, \quad \forall n \in \mathbb{N}_0, \quad (88)$$

where here

$$b_n := \sum_{q \in \Lambda_1^n} \|f_q - h_q\|_2^2, \quad \forall n \in \mathbb{N}_0,$$

and to note that, similarly to (54),

$$\begin{aligned} \sum_{n=0}^N a_n &\leq \sum_{n=0}^N (b_n - b_{n+1}) = (b_0 - b_1) + (b_1 - b_2) \\ &\quad + \cdots + (b_N - b_{N+1}) = b_0 - \underbrace{b_{N+1}}_{\geq 0} \\ &\leq b_0 = \sum_{q \in \Lambda_1^0} \|f_q - h_q\|_2^2 = \|U[e]f - U[e]h\|_2^2 \\ &= \|f - h\|_2^2, \end{aligned}$$

which then yields (87) according to

$$\begin{aligned} |||\Phi_\Omega(f) - \Phi_\Omega(h)|||^2 &= \lim_{N \rightarrow \infty} \sum_{n=0}^N a_n \leq \lim_{N \rightarrow \infty} \|f - h\|_2^2 \\ &= \|f - h\|_2^2. \end{aligned}$$

Writing out (88), it follows that we need to establish

$$\begin{aligned} &\sum_{q \in \Lambda_1^n} \|f_q * \chi_n - h_q * \chi_n\|_2^2 \\ &\leq \sum_{q \in \Lambda_1^n} \|f_q - h_q\|_2^2 - \sum_{q \in \Lambda_1^{n+1}} \|f_q - h_q\|_2^2, \end{aligned} \quad (89)$$

for all $n \in \mathbb{N}_0$. We start by examining the second term on the RHS of (89) and note that, thanks to the decomposition

$$\tilde{q} \in \Lambda_1^{n+1} = \underbrace{\Lambda_1 \times \cdots \times \Lambda_n}_{=\Lambda_1^n} \times \Lambda_{n+1}$$

and $U[\tilde{q}] = U[(q, \lambda_{n+1})] = U_{n+1}[\lambda_{n+1}]U[q]$, by (13), we have

$$\begin{aligned} \sum_{\tilde{q} \in \Lambda_1^{n+1}} \|f_{\tilde{q}} - h_{\tilde{q}}\|_2^2 &= \sum_{q \in \Lambda_1^n} \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|U_{n+1}[\lambda_{n+1}]f_q \\ &\quad - U_{n+1}[\lambda_{n+1}]h_q\|_2^2. \end{aligned} \quad (90)$$

Substituting (90) into (89) and rearranging terms, we obtain

$$\begin{aligned} &\sum_{q \in \Lambda_1^n} \left(\|f_q * \chi_n - h_q * \chi_n\|_2^2 \right. \\ &\quad \left. + \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|U_{n+1}[\lambda_{n+1}]f_q - U_{n+1}[\lambda_{n+1}]h_q\|_2^2 \right) \\ &\leq \sum_{q \in \Lambda_1^n} \|f_q - h_q\|_2^2, \quad \forall n \in \mathbb{N}_0. \end{aligned} \quad (91)$$

We next note that the second term inside the sum on the LHS of (91) satisfies

$$\begin{aligned} &\sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|U_{n+1}[\lambda_{n+1}]f_q - U_{n+1}[\lambda_{n+1}]h_q\|_2^2 \\ &\leq \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|P_{n+1}(M_{n+1}(f_q * g_{\lambda_{n+1}})) \\ &\quad - P_{n+1}(M_{n+1}(h_q * g_{\lambda_{n+1}}))\|_2^2, \end{aligned} \quad (92)$$

where we employed arguments similar to those leading to (58). Substituting the second term inside the sum on the LHS of (91) by the upper bound (92), and using the Lipschitz property of M_{n+1} and P_{n+1} yields

$$\begin{aligned} &\sum_{q \in \Lambda_1^n} \left(\|f_q * \chi_n - h_q * \chi_n\|_2^2 \right. \\ &\quad \left. + \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|U_{n+1}[\lambda_{n+1}]f_q - U_{n+1}[\lambda_{n+1}]h_q\|_2^2 \right) \\ &\leq \sum_{q \in \Lambda_1^n} \max\{1, L_{n+1}^2 R_{n+1}^2\} \left(\|f_q - h_q\|_2^2 \right. \\ &\quad \left. + \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|(f_q - h_q) * g_{\lambda_{n+1}}\|_2^2 \right), \end{aligned} \quad (93)$$

for all $n \in \mathbb{N}_0$. As the functions $\{g_{\lambda_{n+1}}\}_{\lambda_{n+1} \in \Lambda_{n+1}} \cup \{\chi_n\}$ are the atoms of the semi-discrete frame Ψ_{n+1} for $L^2(\mathbb{R}^d)$ and $f_q, h_q \in L^2(\mathbb{R}^d)$, as established above, we have

$$\begin{aligned} &\|(f_q - h_q) * \chi_n\|_2^2 + \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|(f_q - h_q) * g_{\lambda_{n+1}}\|_2^2 \\ &\leq B_{n+1} \|f_q - h_q\|_2^2, \end{aligned}$$

which, when used in (93) yields

$$\begin{aligned} & \sum_{q \in \Lambda_1^n} \left(\|f_q * \chi_n - h_q * \chi_n\|_2^2 \right. \\ & + \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|U_{n+1}[\lambda_{n+1}]f_q - U_{n+1}[\lambda_{n+1}]h_q\|_2^2 \Big) \\ & \leq \sum_{q \in \Lambda_1^n} \max\{B_{n+1}, B_{n+1}L_{n+1}^2R_{n+1}^2\} \|f_q - h_q\|_2^2, \end{aligned} \quad (94)$$

for all $n \in \mathbb{N}_0$. Finally, invoking the admissibility condition

$$\max\{B_n, B_nL_n^2R_n^2\} \leq 1, \quad \forall n \in \mathbb{N},$$

in (94) we get (91) and hence (88). This completes the proof. \square

APPENDIX J PROPOSITION 5

Proposition 5: There exists a constant $C > 0$ such that for all $f \in L_R^2(\mathbb{R}^d)$, $\omega \in C(\mathbb{R}^d, \mathbb{R})$, and $\tau \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ with $\|D\tau\|_\infty \leq \frac{1}{2d}$, it holds that

$$\|f - F_{\tau, \omega} f\|_2 \leq C(R\|\tau\|_\infty + \|\omega\|_\infty) \|f\|_2. \quad (95)$$

Remark 5: A similar bound was derived in [22, App. B] for scattering networks, namely

$$\|f * \psi_{(-J, 0)} - F_\tau(f * \psi_{(-J, 0)})\|_2 \leq C2^{-J+d} \|\tau\|_\infty \|f\|_2, \quad (96)$$

for all $f \in L^2(\mathbb{R}^d)$, where $\psi_{(-J, 0)}$ is the low-pass filter of a semi-discrete directional wavelet frame for $L^2(\mathbb{R}^d)$, and $(F_\tau f)(x) = f(x - \tau(x))$. The techniques for proving (95) and (96) are related in the sense of both employing Schur's Lemma [63, App. I.1] and a Taylor series expansion argument [83, p. 411]. The signal-class specificity of our bound (95) comes with new technical elements detailed at the beginning of the proof.

Proof: We first determine an integral operator

$$(Kf)(x) = \int_{\mathbb{R}^d} k(x, u) f(u) du \quad (97)$$

satisfying the signal-class specific identity

$$Kf = F_{\tau, \omega} f - f, \quad \forall f \in L^2(\mathbb{R}^d),$$

and then upper-bound the deformation error $\|f - F_{\tau, \omega} f\|_2$ according to

$$\|f - F_{\tau, \omega} f\|_2 = \|F_{\tau, \omega} f - f\|_2 = \|Kf\|_2 \leq \|K\|_{2,2} \|f\|_2,$$

for all $f \in L_R^2(\mathbb{R}^d)$. Application of Schur's Lemma, stated below, then yields

$$\|K\|_{2,2} \leq C(R\|\tau\|_\infty + \|\omega\|_\infty), \quad \text{with } C > 0,$$

which completes the proof.

Schur's Lemma [63, App. I.1]: Let $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{C}$ be a locally integrable function satisfying

$$\begin{aligned} (i) \quad & \sup_{x \in \mathbb{R}^d} \int_{\mathbb{R}^d} |k(x, u)| du \leq \alpha, \\ (ii) \quad & \sup_{u \in \mathbb{R}^d} \int_{\mathbb{R}^d} |k(x, u)| dx \leq \alpha, \end{aligned} \quad (98)$$

where $\alpha > 0$. Then, $(Kf)(x) = \int_{\mathbb{R}^d} k(x, u) f(u) du$ is a bounded operator from $L^2(\mathbb{R}^d)$ to $L^2(\mathbb{R}^d)$ with operator norm $\|K\|_{2,2} \leq \alpha$.

We start by determining the integral operator K in (97). To this end, consider $\eta \in S(\mathbb{R}^d, \mathbb{C})$ such that $\widehat{\eta}(\omega) = 1$, for all $\omega \in B_1(0)$. Setting $\gamma(x) := R^d \eta(Rx)$ yields $\gamma \in S(\mathbb{R}^d, \mathbb{C})$ and $\widehat{\gamma}(\omega) = \widehat{\eta}(\omega/R)$. Thus, $\widehat{\gamma}(\omega) = 1$, for all $\omega \in B_R(0)$, and hence $\widehat{f} = \widehat{f} \cdot \widehat{\gamma}$, so that $f = f * \gamma$, for all $f \in L_R^2(\mathbb{R}^d)$. Next, we define the operator $A_\gamma : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$, $A_\gamma f := f * \gamma$, and note that A_γ is well-defined, i.e., $A_\gamma f \in L^2(\mathbb{R}^d)$, for all $f \in L^2(\mathbb{R}^d)$, thanks to Young's inequality [63, Th. 1.2.12] (since $f \in L^2(\mathbb{R}^d)$ and $\gamma \in S(\mathbb{R}^d, \mathbb{C}) \subseteq L^1(\mathbb{R}^d)$). Moreover, $A_\gamma f = f$, for all $f \in L_R^2(\mathbb{R}^d)$. Setting $K := F_{\tau, \omega} A_\gamma - A_\gamma$, we get $Kf = F_{\tau, \omega} A_\gamma f - A_\gamma f = F_{\tau, \omega} f - f$, for all $f \in L_R^2(\mathbb{R}^d)$, as desired. Furthermore, it follows from

$$(F_{\tau, \omega} A_\gamma f)(x) = e^{2\pi i \omega(x)} \int_{\mathbb{R}^d} \gamma(x - \tau(x) - u) f(u) du,$$

that the integral operator $K = F_{\tau, \omega} A_\gamma - A_\gamma$, i.e.,

$$(Kf)(x) = \int_{\mathbb{R}^d} k(x, u) f(u) du,$$

has the kernel

$$k(x, u) := e^{2\pi i \omega(x)} \gamma(x - \tau(x) - u) - \gamma(x - u). \quad (99)$$

Before we can apply Schur's Lemma to establish an upper bound on $\|K\|_{2,2}$, we need to verify that k in (99) is locally integrable, i.e., we need to show that for every compact set $S \subseteq \mathbb{R}^d \times \mathbb{R}^d$ we have

$$\int_S |k(x, u)| dx du < \infty.$$

To this end, let $S \subseteq \mathbb{R}^d \times \mathbb{R}^d$ be a compact set. Next, choose compact sets $S_1, S_2 \subseteq \mathbb{R}^d$ such that $S \subseteq S_1 \times S_2$. Thanks to $\gamma \in S(\mathbb{R}^d, \mathbb{C})$, $\tau \in C^1(\mathbb{R}^d, \mathbb{R}^d)$, and $\omega \in C(\mathbb{R}^d, \mathbb{R})$, all by assumption, the function $|k| : S_1 \times S_2 \rightarrow \mathbb{C}$ is continuous as a composition of continuous functions, and therefore also Lebesgue-measurable. We further have

$$\begin{aligned} \int_{S_1} \int_{S_2} |k(x, u)| dx du & \leq \int_{S_1} \int_{\mathbb{R}^d} |k(x, u)| dx du \\ & \leq \int_{S_1} \int_{\mathbb{R}^d} |\gamma(x - \tau(x) - u)| dx du + \int_{S_1} \int_{\mathbb{R}^d} |\gamma(x - u)| dx du \\ & \leq 2 \int_{S_1} \int_{\mathbb{R}^d} |\gamma(y)| dy du + \int_{S_1} \int_{\mathbb{R}^d} |\gamma(y)| dy du \\ & = 3\mu_L(S_1) \|\gamma\|_1 < \infty, \end{aligned} \quad (100)$$

where the first term in (100) follows by the change of variables $y = x - \tau(x) - u$, together with

$$\frac{dy}{dx} = |\det(E - (D\tau)(x))| \geq 1 - d\|D\tau\|_\infty \geq 1/2, \quad (101)$$

for all $x \in \mathbb{R}^d$. The arguments underlying (101) were already detailed at the end of Appendix H. It follows that k is locally integrable owing to

$$\begin{aligned} \int_S |k(x, u)| dx du & \leq \int_{S_1 \times S_2} |k(x, u)| dx du \\ & = \int_{S_1} \int_{S_2} |k(x, u)| dx du < \infty, \end{aligned} \quad (102)$$

where the first step in (102) follows from $S \subseteq S_1 \times S_2$, the second step is thanks to the Fubini-Tonelli Theorem [84, Th. 14.2] noting that $|k| : S_1 \times S_2 \rightarrow \mathbb{C}$ is Lebesgue-measurable (as established above) and non-negative, and the last step is due to (100). Next, we need to verify conditions (i) and (ii) in (98) and determine the corresponding $\alpha > 0$. In fact, we seek a specific constant α of the form

$$\alpha = C(R\|\tau\|_\infty + \|\omega\|_\infty), \quad \text{with } C > 0. \quad (103)$$

This will be accomplished as follows: For $x, u \in \mathbb{R}^d$, we parametrize the integral kernel in (99) according to $h_{x,u}(t) := e^{2\pi i t \omega(x)} \gamma(x - t\tau(x) - u) - \gamma(x - u)$. A Taylor series expansion [83, p. 411] of $h_{x,u}(t)$ w.r.t. the variable t now yields

$$h_{x,u}(t) = \underbrace{h_{x,u}(0)}_{=0} + \int_0^t h'_{x,u}(\lambda) d\lambda = \int_0^t h'_{x,u}(\lambda) d\lambda, \quad (104)$$

for $t \in \mathbb{R}$, where $h'_{x,u}(t) = (\frac{d}{dt} h_{x,u})(t)$. Note that $h_{x,u} \in C^1(\mathbb{R}, \mathbb{C})$ thanks to $\gamma \in S(\mathbb{R}^d, \mathbb{C})$. Setting $t = 1$ in (104) we get

$$|k(x, u)| = |h_{x,u}(1)| \leq \int_0^1 |h'_{x,u}(\lambda)| d\lambda, \quad (105)$$

where

$$h'_{x,u}(\lambda) = -e^{2\pi i \lambda \omega(x)} \langle \nabla \gamma(x - \lambda \tau(x) - u), \tau(x) \rangle + 2\pi i \omega(x) e^{2\pi i \lambda \omega(x)} \gamma(x - \lambda \tau(x) - u), \quad (106)$$

for $\lambda \in [0, 1]$. We further have

$$\begin{aligned} |h'_{x,u}(\lambda)| &\leq |\langle \nabla \gamma(x - \lambda \tau(x) - u), \tau(x) \rangle| \\ &\quad + |2\pi i \omega(x) \gamma(x - \lambda \tau(x) - u)| \\ &\leq |\tau(x)| |\nabla \gamma(x - \lambda \tau(x) - u)| \\ &\quad + 2\pi |\omega(x)| |\gamma(x - \lambda \tau(x) - u)|. \end{aligned} \quad (107)$$

Now, using $|\tau(x)| \leq \sup_{y \in \mathbb{R}^d} |\tau(y)| = \|\tau\|_\infty$ and $|\omega(x)| \leq \sup_{y \in \mathbb{R}^d} |\omega(y)| = \|\omega\|_\infty$ in (107), together with (105), we get the upper bound

$$\begin{aligned} |k(x, u)| &\leq \|\tau\|_\infty \int_0^1 |\nabla \gamma(x - \lambda \tau(x) - u)| d\lambda \\ &\quad + 2\pi \|\omega\|_\infty \int_0^1 |\gamma(x - \lambda \tau(x) - u)| d\lambda. \end{aligned} \quad (108)$$

Next, we integrate (108) w.r.t. u to establish (i) in (98):

$$\begin{aligned} &\int_{\mathbb{R}^d} |k(x, u)| du \\ &\leq \|\tau\|_\infty \int_{\mathbb{R}^d} \int_0^1 |\nabla \gamma(x - \lambda \tau(x) - u)| d\lambda du \\ &\quad + 2\pi \|\omega\|_\infty \int_{\mathbb{R}^d} \int_0^1 |\gamma(x - \lambda \tau(x) - u)| d\lambda du \\ &= \|\tau\|_\infty \int_0^1 \int_{\mathbb{R}^d} |\nabla \gamma(x - \lambda \tau(x) - u)| du d\lambda \\ &\quad + 2\pi \|\omega\|_\infty \int_0^1 \int_{\mathbb{R}^d} |\gamma(x - \lambda \tau(x) - u)| du d\lambda \end{aligned} \quad (109)$$

$$\begin{aligned} &= \|\tau\|_\infty \int_0^1 \int_{\mathbb{R}^d} |\nabla \gamma(y)| dy d\lambda \\ &\quad + 2\pi \|\omega\|_\infty \int_0^1 \int_{\mathbb{R}^d} |\gamma(y)| dy d\lambda \\ &= \|\tau\|_\infty \|\nabla \gamma\|_1 + 2\pi \|\omega\|_\infty \|\gamma\|_1, \end{aligned} \quad (110)$$

where (109) follows by application of the Fubini-Tonelli Theorem [84, Th. 14.2] noting that the functions $(u, \lambda) \mapsto |\nabla \gamma(x - \lambda \tau(x) - u)|$, $(u, \lambda) \in \mathbb{R}^d \times [0, 1]$, and $(u, \lambda) \mapsto |\gamma(x - \lambda \tau(x) - u)|$, $(u, \lambda) \in \mathbb{R}^d \times [0, 1]$, are both non-negative and continuous (and thus Lebesgue-measurable) as compositions of continuous functions. Finally, using $\gamma = R^d \eta(R \cdot)$, and thus $\nabla \gamma = R^{d+1} \nabla \eta(R \cdot)$, $\|\gamma\|_1 = \|\eta\|_1$, and $\|\nabla \gamma\|_1 = R \|\nabla \eta\|_1$ in (110) yields

$$\begin{aligned} \sup_{x \in \mathbb{R}^d} \int_{\mathbb{R}^d} |k(x, u)| du &\leq R \|\tau\|_\infty \|\nabla \eta\|_1 + 2\pi \|\omega\|_\infty \|\eta\|_1 \\ &\leq \max \{ \|\nabla \eta\|_1, 2\pi \|\eta\|_1 \} (R \|\tau\|_\infty + \|\omega\|_\infty), \end{aligned} \quad (111)$$

which establishes an upper bound of the form (i) in (98) that exhibits the desired structure for α . Condition (ii) in (98) is established similarly by integrating (108) w.r.t. x according to

$$\begin{aligned} &\int_{\mathbb{R}^d} |k(x, u)| dx \\ &\leq \|\tau\|_\infty \int_{\mathbb{R}^d} \int_0^1 |\nabla \gamma(x - \lambda \tau(x) - u)| d\lambda dx \\ &\quad + 2\pi \|\omega\|_\infty \int_{\mathbb{R}^d} \int_0^1 |\gamma(x - \lambda \tau(x) - u)| d\lambda dx \\ &= \|\tau\|_\infty \int_0^1 \int_{\mathbb{R}^d} |\nabla \gamma(x - \lambda \tau(x) - u)| dx d\lambda \\ &\quad + 2\pi \|\omega\|_\infty \int_0^1 \int_{\mathbb{R}^d} |\gamma(x - \lambda \tau(x) - u)| dx d\lambda \end{aligned} \quad (112)$$

$$\begin{aligned} &\leq 2 \|\tau\|_\infty \int_0^1 \int_{\mathbb{R}^d} |\nabla \gamma(y)| dy d\lambda \\ &\quad + 4\pi \|\omega\|_\infty \int_0^1 \int_{\mathbb{R}^d} |\gamma(y)| dy d\lambda \\ &= 2 \|\tau\|_\infty \|\nabla \gamma\|_1 + 4\pi \|\omega\|_\infty \|\gamma\|_1 \\ &\leq \max \{ 2 \|\nabla \eta\|_1, 4\pi \|\eta\|_1 \} (R \|\tau\|_\infty + \|\omega\|_\infty). \end{aligned} \quad (113)$$

Here, again, (112) follows by application of the Fubini-Tonelli Theorem [84, Th. 14.2] noting that the functions $(x, \lambda) \mapsto |\nabla \gamma(x - \lambda \tau(x) - u)|$, $(x, \lambda) \in \mathbb{R}^d \times [0, 1]$, and $(x, \lambda) \mapsto |\gamma(x - \lambda \tau(x) - u)|$, $(x, \lambda) \in \mathbb{R}^d \times [0, 1]$, are both non-negative and continuous (and thus Lebesgue-measurable) as compositions of continuous functions. The inequality (113) follows from a change of variables argument similar to the one in (100) and (101). Combining (111) and (114), we finally get (103) with

$$C := \max \{ 2 \|\nabla \eta\|_1, 4\pi \|\eta\|_1 \}. \quad (115)$$

This completes the proof. \square

ACKNOWLEDGMENTS

The authors would like to thank P. Grohs, S. Mallat, R. Alaifari, M. Tschannen, and G. Kutyniok for helpful discussions and comments on the paper.

REFERENCES

- [1] T. Wiatowski and H. Bölcskei, "Deep convolutional neural networks based on semi-discrete frames," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 1212–1216.
- [2] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

- [3] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2009.
- [4] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Hoboken, NJ, USA: Wiley, 2001.
- [5] Y. LeCun and C. Cortes. (1998). *The MNIST Database of Handwritten Digits*. [Online]. Available: <http://yann.lecun.com/exdb/mnist/>
- [6] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
- [7] Y. LeCun *et al.*, "Handwritten digit recognition with a back-propagation network," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 1990, pp. 396–404.
- [8] D. E. Rumelhart, G. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, J. L. McClelland and D. E. Rumelhart, Eds. Cambridge, MA, USA: MIT Press, 1986, pp. 318–362.
- [9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
- [10] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May/Jun. 2010, pp. 253–256.
- [11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
- [12] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier, "Parseval networks: Improving robustness to adversarial examples," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 854–863.
- [13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 770–778.
- [14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: <http://www.deeplearningbook.org>
- [15] F. J. Huang and Y. LeCun, "Large-scale learning with SVM and convolutional nets for generic object categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 284–291.
- [16] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2009, pp. 2146–2153.
- [17] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.
- [18] M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun, "Efficient learning of sparse representations with an energy-based model," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2006, pp. 1137–1144.
- [19] N. Pinto, D. D. Cox, and J. J. DiCarlo, "Why is real-world visual object recognition hard," *PLoS Comput. Biol.*, vol. 4, no. 1, pp. 151–156, 2008.
- [20] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 994–1000.
- [21] J. Mutch and D. G. Lowe, "Multiclass object recognition with sparse, localized features," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 11–18.
- [22] S. Mallat, "Group invariant scattering," *Commun. Pure Appl. Math.*, vol. 65, no. 10, pp. 1331–1398, 2012.
- [23] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1872–1886, Aug. 2013.
- [24] L. Sifre, "Rigid-motion scattering for texture classification," Ph.D. dissertation, Centre Math. Appl., École Polytechn., Palaiseau, France, 2014.
- [25] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4114–4128, Aug. 2014.
- [26] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2011, pp. 315–323.
- [27] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.
- [28] A.-R. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 14–22, Jan. 2011.
- [29] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2010, pp. 249–256.
- [30] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed. San Diego, CA, USA: Academic Press, 2009.
- [31] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 7, pp. 710–732, Jul. 1992.
- [32] M. Unser, "Texture classification and segmentation using wavelet frames," *IEEE Trans. Image Process.*, vol. 4, no. 11, pp. 1549–1560, Nov. 1995.
- [33] P. Vandergheynst and J.-F. Gobbers, "Directional dyadic wavelet transforms: Design and algorithms," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 363–372, Apr. 2002.
- [34] E. J. Candès and D. L. Donoho, "Continuous curvelet transform: II. Discretization and frames," *Appl. Comput. Harmon. Anal.*, vol. 19, no. 2, pp. 198–222, 2005.
- [35] P. Grohs, S. Keiper, G. Kutyniok, and M. Schäfer, "Cartoon approximation with α -curvelets," *J. Fourier Anal. Appl.*, vol. 22, no. 6, pp. 1235–1293, 2015.
- [36] G. Kutyniok and D. Labate, Eds., *Shearlets: Multiscale Analysis for Multivariate Data*. Cambridge, MA, USA: Birkhäuser, 2012.
- [37] P. Grohs, "Ridgelet-type frame decompositions for Sobolev spaces related to linear transport," *J. Fourier Anal. Appl.*, vol. 18, no. 2, pp. 309–325, 2012.
- [38] E. J. Candès and D. L. Donoho, "New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities," *Commun. Pure Appl. Math.*, vol. 57, no. 2, pp. 219–266, Feb. 2004.
- [39] K. Guo, G. Kutyniok, and D. Labate, "Sparse multidimensional representations using anisotropic dilation and shear operators," in *Wavelets and Splines*, G. Chen and M. J. Lai, Eds. Brentwood, TN, USA: Nashboro Press, 2006, pp. 189–201.
- [40] P. Grohs, T. Wiatowski, and H. Bölcskei, "Deep convolutional neural networks on cartoon functions," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 1163–1167.
- [41] S. T. Ali, J. P. Antoine, and J. P. Gazeau, "Continuous frames in Hilbert space," *Ann. Phys.*, vol. 222, no. 1, pp. 1–37, 1993.
- [42] G. Kaiser, *A Friendly Guide to Wavelets*. Cambridge, MA, USA: Birkhäuser, 1994.
- [43] W. Rudin, *Functional Analysis*, 2nd ed., New York, NY, USA: McGraw-Hill, 1991.
- [44] J. P. Antoine, R. Murrenzi, P. Vandergheynst, and S. T. Ali, *Two-Dimensional Wavelets and Their Relatives*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
- [45] T. S. Lee, "Image representation using 2D Gabor wavelets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 10, pp. 959–971, Oct. 1996.
- [46] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.
- [47] E. Oyallon and S. Mallat, "Deep roto-translation scattering for object classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 2865–2873.
- [48] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.
- [49] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
- [50] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, May 2010.
- [51] S. Chen, C. F. N. Cowan, and P. M. Grant, "Orthogonal least squares learning algorithm for radial basis function networks," *IEEE Trans. Neural Netw.*, vol. 2, no. 2, pp. 302–309, Mar. 1991.
- [52] G. Kutyniok and D. Labate, "Introduction to shearlets," in *Shearlets: Multiscale Analysis for Multivariate Data*. Cambridge, MA, USA: Birkhäuser, 2012, pp. 1–38.
- [53] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1992.
- [54] D. P. W. Ellis, X. Zeng, and J. H. McDermott, "Classifying soundtracks with audio texture features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2011, pp. 5880–5883.
- [55] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.
- [56] J. Lin and L. Qu, "Feature extraction based on Morlet wavelet and its application for mechanical fault diagnosis," *J. Sound Vibrat.*, vol. 234, no. 1, pp. 135–148, 2000.

- [57] G. Y. Chen, T. D. Bui, and A. Krzyżak, "Rotation invariant pattern recognition using ridgelets, wavelet cycle-spinning and Fourier features," *Pattern Recognit.*, vol. 38, no. 12, pp. 2314–2322, 2005.
- [58] Y.-L. Qiao, C.-Y. Song, and C.-H. Zhao, "M-band ridgelet transform based texture classification," *Pattern Recognit. Lett.*, vol. 31, no. 3, pp. 244–249, 2010.
- [59] S. Arivazhagan, L. Ganesan, and T. G. S. Kumar, "Texture classification using ridgelet transform," *Pattern Recognit. Lett.*, vol. 27, no. 16, pp. 1875–1883, 2006.
- [60] J. Ma and G. Plonka, "The curvelet transform," *IEEE Signal Process. Mag.*, vol. 27, no. 2, pp. 118–133, Mar. 2010.
- [61] L. Dettori and L. Semler, "A comparison of wavelet, ridgelet, and curvelet-based texture classification algorithms in computed tomography," *Comput. Biol. Med.*, vol. 37, no. 4, pp. 486–498, 2007.
- [62] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.
- [63] L. Grafakos, *Classical Fourier Analysis*, 2nd ed., New York, NY, USA: Springer, 2008.
- [64] T. Wiatowski, M. Tschannen, A. Stanić, P. Grohs, and H. Bölcskei, "Discrete deep feature extraction: A theory and new architectures," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 2149–2158.
- [65] D. L. Donoho, "Sparse components of images and optimal atomic decompositions," *Construct. Approx.*, vol. 17, no. 3, pp. 353–382, 2001.
- [66] T. Wiatowski, P. Grohs, and H. Bölcskei, "Energy propagation in deep convolutional neural networks," *IEEE Trans. Inf. Theory*, to be published.
- [67] A. J. E. M. Janssen, "The duality condition for Weyl-Heisenberg frames," in *Gabor Analysis: Theory and Applications*, H. G. Feichtinger and T. Strohmer, Eds. Cambridge, MA, USA: Birkhäuser, 1998, pp. 33–84.
- [68] A. Ron and Z. Shen, "Frames and stable bases for shift-invariant subspaces of $L^2(\mathbb{R}^d)$," *Canad. J. Math.*, vol. 47, no. 5, pp. 1051–1094, 1995.
- [69] M. Frazier, B. Jawerth, and G. Weiss, *Littlewood-Paley Theory and the Study of Function Spaces*. Providence, RI, USA: American Mathematical Society, 1991.
- [70] A. W. Naylor and G. R. Sell, *Linear Operator Theory in Engineering and Science*. New York, NY, USA: Springer, 1982.
- [71] H. Bölcskei, F. Hlawatsch, and H. G. Feichtinger, "Frame-theoretic analysis of oversampled filter banks," *IEEE Trans. Signal Process.*, vol. 46, no. 12, pp. 3256–3268, Dec. 1998.
- [72] A. J. E. M. Janssen, "Duality and biorthogonality for Weyl-Heisenberg frames," *J. Fourier Anal. Appl.*, vol. 1, no. 4, pp. 403–436, 1995.
- [73] I. Daubechies, H. J. Landau, and Z. Landau, "Gabor time-frequency lattices and the Wexler-Raz identity," *J. Fourier Anal. Appl.*, vol. 1, no. 4, pp. 438–478, 1995.
- [74] K. Gröchenig, *Foundations of Time-Frequency Analysis*. Cambridge, MA, USA: Birkhäuser, 2001.
- [75] I. Daubechies, A. Grossmann, and Y. Meyer, "Painless nonorthogonal expansions," *J. Math. Phys.*, vol. 27, no. 5, pp. 1271–1283, 1986.
- [76] K. Gröchenig and S. Samarah, "Nonlinear approximation with local Fourier bases," *Construct. Approx.*, vol. 16, no. 3, pp. 317–331, 2000.
- [77] C.-H. Lee, J.-L. Shih, K.-M. Yu, and H.-S. Lin, "Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features," *IEEE Trans. Multimedia*, vol. 11, no. 4, pp. 670–682, Jun. 2009.
- [78] D. Donoho and G. Kutyniok, "Microlocal analysis of the geometric separation problem," *Commun. Pure Appl. Math.*, vol. 66, no. 1, pp. 1–47, 2013.
- [79] E. J. Candès, *Ridgelets: Theory and Applications*, Ph.D. dissertation, Stanford Univ., Stanford, CA, USA: 1998.
- [80] E. J. Candès and D. L. Donoho, "Ridgelets: A key to higher-dimensional intermittency?" *Philos. Trans. Roy. Soc. London A, Math. Phys. Sci.*, vol. 357, no. 1760, pp. 2495–2509, Sep. 1999.
- [81] M. Searcoid, *Metric Spaces*. London, U.K.: Springer, 2007.
- [82] R. P. Brent, J.-A. H. Osborn, and W. D. Smith, "Note on best possible bounds for determinants of matrices close to the identity matrix," *Linear Algebra Appl.*, vol. 466, pp. 21–26, Feb. 2015.
- [83] W. Rudin, *Real and Complex Analysis*, 2nd ed., New York, NY, USA: McGraw-Hill, 1983.
- [84] E. DiBenedetto, *Real Analysis*. Cambridge, MA, USA: Birkhäuser, 2002.

Thomas Wiatowski was born in Strzelce Opolskie, Poland, on December 20, 1987, and received the BSc and MSc degrees, both in Mathematics, from the Technical University of Munich, Germany, in 2010 and 2012, respectively. In 2012 he was a researcher with the Institute of Computational Biology at the Helmholtz Zentrum in Munich, Germany. He joined ETH Zurich in 2013, where he graduated with the Dr. sc. degree in 2017. His research interests are in deep machine learning, mathematical signal processing, and applied harmonic analysis.

Helmut Bölcskei (S'94–M'98–SM'02–F'09) was born in Mödling, Austria, on May 29, 1970, and received the Dipl.-Ing. and Dr. techn. degrees in electrical engineering from Vienna University of Technology, Vienna, Austria, in 1994 and 1997, respectively. In 1998 he was with Vienna University of Technology. From 1999 to 2001 he was a postdoctoral researcher in the Information Systems Laboratory, Department of Electrical Engineering, and in the Department of Statistics, Stanford University, Stanford, CA. He was in the founding team of Iospan Wireless Inc., a Silicon Valley-based startup company (acquired by Intel Corporation in 2002) specialized in multiple-input multiple-output (MIMO) wireless systems for high-speed Internet access, and was a co-founder of Celestrius AG, Zurich, Switzerland. From 2001 to 2002 he was an Assistant Professor of Electrical Engineering at the University of Illinois at Urbana-Champaign. He has been with ETH Zurich since 2002, where he is a Professor of Electrical Engineering. He was a visiting researcher at Philips Research Laboratories Eindhoven, The Netherlands, ENST Paris, France, and the Heinrich Hertz Institute Berlin, Germany. His research interests are in information theory, mathematical signal processing, machine learning, and statistics.

He received the 2001 IEEE Signal Processing Society Young Author Best Paper Award, the 2006 IEEE Communications Society Leonard G. Abraham Best Paper Award, the 2010 Vodafone Innovations Award, the ETH "Golden Owl" Teaching Award, is a Fellow of the IEEE, a 2011 EURASIP Fellow, was a Distinguished Lecturer (2013–2014) of the IEEE Information Theory Society, an Erwin Schrödinger Fellow (1999–2001) of the Austrian National Science Foundation (FWF), was included in the 2014 Thomson Reuters List of Highly Cited Researchers in Computer Science, and is the 2016 Padovani Lecturer of the IEEE Information Theory Society. He served as an associate editor of the IEEE TRANSACTIONS ON INFORMATION THEORY, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and the EURASIP Journal on Applied Signal Processing. He was editor-in-chief of the IEEE TRANSACTIONS ON INFORMATION THEORY during the period 2010–2013. He served on the editorial board of the *IEEE Signal Processing Magazine* and is currently on the editorial boards of *Foundations and Trends in Networking and Foundations and Trends in Communications and Information Theory*. He was TPC co-chair of the 2008 IEEE International Symposium on Information Theory and the 2016 IEEE Information Theory Workshop and currently serves as 2nd Vice President of the IEEE Information Theory Society. He has been a delegate of the president of ETH Zurich for faculty appointments since 2008.