A Mathematical Theory of Deep Convolutional Neural Networks for Feature Extraction

Thomas Wiatowski and Helmut Bölcskei, Fellow, IEEE

Abstract—Deep convolutional neural networks have led to breakthrough results in numerous practical machine learning tasks such as classification of images in the ImageNet dataset, 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 trainable classifier. The mathematical analysis of deep convolutional neural networks 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 trainable classifier (such as, e.g., a support vector machine (SVM) [6]) improves classification performance. Specifically, non-linear feature extractors (obtained, e.g., through the use of a so-called kernel in the context of SVMs) 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 pooling 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], [29], [21]. 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.

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).
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 [35, 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 subsampling and averaging. Finally, different network layers may be equipped with different convolutional transforms, different (Lipschitz-continuous) non-linearities, and different pooling operators.

Regarding translation invariance, it was argued, e.g., in [15–17, 20, 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 [41]. 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 structures 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 [41]. 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 $\overline{z}$. 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 $\langle x, y \rangle := \sum_{i=1}^{d} x_i y_i$, 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 $T \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$, $T_{i,j,k}$ 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 $T \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$ 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 $\int_{B_r(x)} f(x) \, dx$ for the integral of $f$ w.r.t. Lebesgue measure $\mu_L$. 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 $\|f\|_p := \left(\int_{B_r(x)} |f(x)|^p \, dx\right)^{1/p} < \infty$. $L^\infty(\mathbb{R}^d)$ denotes the space of Lebesgue-measurable functions $f : \mathbb{R}^d \rightarrow \mathbb{C}$ such that $\|f\|_{\infty} := \inf\{\alpha > 0 \mid |f(x)| \leq \alpha \text{ for a.e. } x \in \mathbb{R}^d\} < \infty$.

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 a function $f \in L^1(\mathbb{R}^d)$ by $\hat{f}(\omega) := \int_{\mathbb{R}^d} f(x)e^{-2\pi i(x,\omega)} \, dx$ and extend it in the usual way to $L^2(\mathbb{R}^d)$ [43, Theorem 7.9]. The convolution of $f \in L^1(\mathbb{R}^d)$ and $g \in L^\infty(\mathbb{R}^d)$ is $(f * g)(y) := \int_{\mathbb{R}^d} f(x)g(y-x) \, dx$. We write $(T_1 \cdot f)(x) := f(x-t)$, $t \in \mathbb{R}^d$, for the translation operator, and $(M_{\alpha} f)(x) := e^{2\pi i(x,\alpha)} f(x)$, $\alpha \in \mathbb{R}^d$, for the modulation operator. Involution is defined by $(f^*)(x) := \overline{f(-x)}$.

A multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ is an ordered $d$-tuple of non-negative integers $\alpha_i \in \mathbb{N}_0$. For a multi-index $\alpha \in \mathbb{N}_0^d$, $D^\alpha$ denotes the differential operator $D^\alpha := (\partial/\partial x_1)^{\alpha_1} \cdots (\partial/\partial x_d)^{\alpha_d}$, with order $|\alpha| := \sum_{i=1}^{d} \alpha_i$. If $|\alpha| = 0$, $D^0 f := f$, $f : \mathbb{R}^d \rightarrow \mathbb{C}$. 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 $C^\infty(\mathbb{R}^d, \mathbb{C})$. $C^\infty(\mathbb{R}^d, \mathbb{C})$ stands for the Schwartz space, i.e., the space of functions $f \in C^\infty(\mathbb{R}^d, \mathbb{C})$ whose derivatives $D^\alpha f$ along with the function itself are rapidly decaying [43, Section 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 $\nabla f$.

The space of continuous mappings $v : \mathbb{R}^p \rightarrow \mathbb{R}^q$ is written as $C^0(\mathbb{R}^p, \mathbb{R}^q)$, 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 $D^2 v$ its Jacobian tensor, with associated norms $\text{tr}(D^2 v)$ and $\text{det}(D^2 v)$.
\[\|v\|_\infty := \sup_{x \in \mathbb{R}^d} |v(x)|, \quad \|Dv\|_\infty := \sup_{x \in \mathbb{R}^d} |(Dv)(x)|, \quad \text{and} \quad \|D^2v\|_\infty := \sup_{x \in \mathbb{R}^d} |(D^2v)(x)|.\]

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, Definition 2.4] defines the feature vector \(\Phi_W(f)\) of the signal \(f \in L^2(\mathbb{R}^d)\) as the set

\[\Phi_W(f) := \bigcup_{n=0}^{\infty} \Phi^n_W(f),\]

where \(\Phi^n_W(f) := \{f \ast \psi_{(j,0)}\}\), and

\[\Phi^n_W(f) := \left\{\left(U[\lambda^{(j)}, \ldots, \lambda^{(p)}]f \ast \psi_{(j,0)}\right)_{\lambda^{(j)}, \ldots, \lambda^{(p)} \in \Lambda_W} : (j, k) \in \mathbb{Z} \cap \{j > -J, k \in \{0, \ldots, K-1\}\}\right\},\]

for all \(n \in \mathbb{N}\), with

\[U[\lambda^{(j)}, \ldots, \lambda^{(p)}]f := \int_{\mathbb{R}^d} f(x) \psi_{j,k}(x) \mathbb{1}_{\{x \in \mathbb{R}^d\}} dx,\]

Here, the index set \(\Lambda_W := \{-J, 0\} \cup \{j, k\} : j \in \mathbb{Z} \cap \{j > -J, k \in \{0, \ldots, K-1\}\}\) 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

\[\psi_{j,k}(x) := 2^d \psi(2^d r_k^{-1} x),\]

where \(\lambda = (j, k) \in \Lambda_W\) 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 \(r_k, k \in \{0, \ldots, K-1\}\), 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 [22].

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

\[\Pi_W := \{T_bI \psi_{\lambda}\}_{b \in \mathbb{R}^d, \lambda \in \Lambda_W},\]

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

\[\sum_{\lambda \in \Lambda_W} \int_{\mathbb{R}^d} |\langle f, T_bI \psi_{\lambda}\rangle|^2 db = \sum_{\lambda \in \Lambda_W} \|f \ast \psi_{\lambda}\|^2 = \|f\|^2,\]

for all \(f \in L^2(\mathbb{R}^d)\), where \(f, T_bI \psi_{\lambda} = (f \ast \psi_{\lambda})(b), (\lambda, b) \in \Lambda_W \times \mathbb{R}^d\), 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.

\[\text{We emphasize that the feature vector } \Phi_W(f) \text{ is a union of the sets of feature vectors } \Phi^n_W(f).\]

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

Remark 1. The function \(|f \ast \psi_{\lambda}|, \lambda \in \Lambda_W\), can be thought of as indicating the locations of singularities of \(f \in L^2(\mathbb{R}^d)\). Specifically, with the relation of \(|f \ast \psi_{\lambda}| \) to the Canny edge detector [46] as described in [37], in dimension \(d = 2\), we can think of \(|f \ast \psi_{\lambda}| = |f \ast \psi_{(j,k)}|\), \(\lambda = (j, k) \in \Lambda_W\), 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], [47] that the feature vector \(\Phi_W(f)\) 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] Theorem 2.10) that the feature extractor \(\Phi_W\) is translation-invariant in the sense of

\[\lim_{j \to \infty} \|\Phi_W(T_j f) - \Phi_W(f)\| = 0,\]

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, Theorem 2.12 establishes that \(\Phi_W\) 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. 4.6), it is shown in [22] Theorem 2.12) that there exists a constant \(C > 0\) such that for all \(f \in H_W\), and \(\tau \in C^1(\mathbb{R}^d, \mathbb{R}^d)\) with \(\|D\tau\|_\infty \leq \frac{1}{2d}\), the deformation error satisfies the following deformation stability bound

\[\|\Phi_W(F_\tau f) - \Phi_W(f)\| \leq C(2^{-J} \|\tau\|_\infty + J \|D\tau\|_\infty + \|D^2\tau\|_\infty)\|f\|_{H_W}.\]

Note that this upper bound goes to infinity as translation

\[\text{It is actually the assumption } \|D\tau\|_\infty \leq \frac{1}{2d}, \text{ rather than } \|D\tau\|_\infty \leq \frac{1}{2}, \text{ as stated in [22] Theorem 2.12], that is needed in [22] p. 1390] to establish that } \|\det(E - (D\tau)(x))\| \geq 1 - d\|D\tau\|_\infty \geq 1/2.\]
invariance through $J \to \infty$ 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 $\Phi_W(f)$ in (1) are indicated at the tips of the arrows.

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) without 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 $\Psi_n := \{T_\lambda g_{\lambda_n}\}_{b \in \mathbb{R}^d, \lambda_n \in \Lambda_n}$, for $L^2(\mathbb{R}^d)$ with countable index set $\Lambda_n$ (see Appendix A for a brief review of the theory of semi-discrete frames), followed by a non-linearity $M_n : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ 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

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

where $S_n \geq 1$ is the pooling factor and $P_n : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ 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 $g_{\lambda_n}$ are 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 $g_{\lambda_n}$ are obtained from a mother wavelet through scaling and rotation operations, see Section II). The corresponding semi-discrete signal transform is briefly reviewed in Appendices B and C.

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 $\Lambda$. Then, the mapping $f \mapsto \{f * g_b(\lambda)\}_{\lambda \in \Lambda}$ 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 (55), 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.
have been employed successfully in the literature in various feature extraction tasks \cite{[23],[54]–[61]}, but their use—apart from wavelets—in DCNNs appears to be new. We refer the reader to Appendix 1 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 \cite{[5]} emulates discrete-time pooling by sub-sampling \cite{[12],[13]} or by averaging \cite{[14],[15]}. Consider a one-dimensional discrete-time signal \( f \in L^2(\mathbb{Z}) := \{ f : \mathbb{Z} \to \mathbb{C} | \sum_{k \in \mathbb{Z}} |f(k)|^2 < \infty \} \). Sub-sampling by a factor of \( S \in \mathbb{N} \) in discrete time is defined by \cite{[62], Sec. 4}

\[
  f_d := f_d[S] := f_d[S']
\]

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

\[
  \widehat{h_d}(\theta) := \sum_{k \in \mathbb{Z}} h_d[k] e^{-2\pi ik\theta} = \frac{1}{S} \sum_{k=0}^{S-1} f_d(\theta - k/S).
\]

The translated copies of \( \widehat{f_d} \) in \cite{[6]} are a consequence of the 1-periodicity of the discrete-time sub-sampling operation. We therefore emulate the discrete-time sub-sampling operation in continuous time through the dilation operation

\[
  f \mapsto h := S^{d/2} f(S'), \quad f \in L^2(\mathbb{R}^d),
\]

which in the frequency domain amounts to dilation according to \( h = S^{-d/2} f(S^{-1}) \). The scaling by \( S^{d/2} \) in \cite{[7]} ensures unitarity of the continuous-time sub-sampling operation. The overall operation in \cite{[7]} fits into our general definition of pooling as it can be recovered from \cite{[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

\[
  f_a := (f_a * \phi_a)[S] := \frac{1}{S} \sum_{k=0}^{S-1} f_a(k/S).
\]

for the (typically compactly supported) “averaging kernel” \( \phi_a \in L^2(\mathbb{Z}) \) and the averaging factor \( S \in \mathbb{N} \). Taking \( \phi_a \) 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 kernel \( \phi_a \). The operation \cite{[8]} can be emulated in continuous time according to

\[
  f \mapsto S^{d/2}(f * \phi)(S'), \quad f \in L^2(\mathbb{R}^d),
\]

with the averaging window \( \phi \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) \). We note that \cite{[9]} can be recovered from \cite{[10]} by taking \( P(f) = f * \phi \), \( f \in L^2(\mathbb{R}^d) \), and noting that convolution with \( \phi \) is Lipschitz-continuous with Lipschitz constant \( R = \| \phi \|_1 \) (thanks to Young’s inequality \cite{[19]} Theorem 1.2.12) and trivially satisfies \( P f = 0 \) for \( f = 0 \). In the remainder of the paper, we refer to the operation in \cite{[5]} as "Lipschitz pooling through dilation" to indicate that \cite{[5]} essentially amounts to the application of a Lipschitz-continuous mapping followed by a continuous-time dilation. Note, however, that the operation in \cite{[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 \((\Psi_n, M_n, P_n)\) associated with individual network layers \( n \) and referred to as modules.

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

\[
  \Omega := \left( (\Psi_n, M_n, P_n) \right)_{n \in \mathbb{N}}
\]

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 \cite{[23]}.

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

\[
  U_n(\lambda, f) := U_n[\lambda_n] f := S_n^{d/2} P_n(M_n(f * g_{\lambda_n}))(S^{-1}_n).
\]

For \( n \in \mathbb{N} \), define the set \( \Lambda_1^n := \Lambda_1 \times \Lambda_2 \times \cdots \times \Lambda_n \). An ordered sequence \( q = (\lambda_1, \lambda_2, \ldots, \lambda_n) \in \Lambda^n \) 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

\[
  \| U_n[\lambda_n] f \|_2^2 = S_n^d \int_{\mathbb{R}^d} \left| P_n(M_n(f * g_{\lambda_n}))(S^{-1}_n x) \right|^2 \nu(x) \, dx
  = \int_{\mathbb{R}^d} \left| P_n(M_n(f * g_{\lambda_n}))(y) \right|^2 \nu(y) \, dy
  = \| P_n(M_n(f * g_{\lambda_n}))(y) \|^2 \leq R_n^2 \| M_n(f * g_{\lambda_n}) \|^2_2
  \leq L_n^2 R_n^2 \| f * g_{\lambda_n} \|^2_2 \leq B_n L_n^2 R_n^2 \| f \|^2_2.
\]

For the inequality in \cite{[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 \( \| P_n f \|_2^2 \leq R_n^2 \| f \|_2^2 \). Similar arguments lead to the first inequality in \cite{[12]}. The last step in \cite{[12]} is thanks to

\[
  \| f * g_{\lambda_n} \|^2_2 \leq \sum_{\lambda \in \Lambda_n} \| f * g_{\lambda_n} \|^2_2 \leq B_n \| f \|^2_2,
\]

which follows from the frame condition \cite{[30]} on \( \Psi_n \). We will also need the extension of the operator \( U_n \) to paths \( q \in \Lambda^n_1 \) according to

\[
  U[q] f := U[(\lambda_1, \lambda_2, \ldots, \lambda_n)] f := U_n[\lambda_n] \cdots U_2[\lambda_2] U_1[\lambda_1] f
\]

with \( U[e] f := f \). Note that the multi-stage operation \cite{[13]} is
again well-defined thanks to
\[ \|U[q]f\|_2^2 \leq \left( \prod_{k=1}^{n} B_k L_k^2 R_k^2 \right) \|f\|_2^2, \]
for \( q \in \Lambda_n^1 \) and \( f \in L^2(\mathbb{R}^d) \), which follows by repeated application of (12).

In scattering networks one atom \( \psi_\lambda, \lambda \in \Lambda_\omega \), in the wavelet frame \( \Psi_{\Lambda_\omega} \), namely the low-pass filter \( \psi_{(-L,0)} \), 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-sequenec \( \Omega = \{ (\psi_n, M_n, P_n) \}_{n \in \mathbb{N}} \) as the output-generating atom \( \chi_{n-1} := g_{\lambda_n^*} \), \( \lambda_n^* \in \Lambda_n \), of the \( (n-1) \)-th layer. The atoms \( \{ g_{\lambda_n} \}_{\lambda_n \in \Lambda_n \setminus \{ \lambda_n^* \}} \) in \( \Psi_n \) 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. From now on, with slight abuse of notation, we shall write \( \Lambda_n \) 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 \( \Phi_\Omega \) based on the module-sequenec \( \Omega \).

**Definition 3.** Let \( \Omega = \{ (\psi_n, M_n, P_n) \}_{n \in \mathbb{N}} \) be a module-sequenec. The feature extractor \( \Phi_\Omega \) based on \( \Omega \) maps \( L^2(\mathbb{R}^d) \) to its feature vector
\[ \Phi_\Omega(f) := \bigcup_{n=0}^{\infty} \Phi_n^\Omega(f), \]
where \( \Phi_n^\Omega(f) := \{(U[q]f) \ast \chi_n \}_{q \in \Lambda_n} \), for all \( n \in \mathbb{N} \).

The set \( \Phi_n^\Omega(f) \) 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 \( \Phi_\Omega : L^2(\mathbb{R}^d) \to (L^2(\mathbb{R}^d))^\Omega \), with \( \Omega := \bigcup_{n=0}^{\infty} \Lambda_n^1 \), is well-defined, i.e., \( \Phi_\Omega(f) \in (L^2(\mathbb{R}^d))^\Omega \), for all \( f \in L^2(\mathbb{R}^d) \), under a technical condition on the module-sequenec \( \Omega \) formalized as follows.

**Proposition 1.** Let \( \Omega = \{ (\psi_n, M_n, P_n) \}_{n \in \mathbb{N}} \) be a module-sequenec. Denote the frame upper bounds of \( \psi_n \) 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
\[ \max \{ B_n, B_n L_n^2 R_n^2 \} \leq 1, \quad \forall n \in \mathbb{N}, \]
then the feature extractor \( \Phi_\Omega : L^2(\mathbb{R}^d) \to (L^2(\mathbb{R}^d))^\Omega \) is well-defined, i.e., \( \Phi_\Omega(f) \in (L^2(\mathbb{R}^d))^\Omega \), 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.

**Definition 4.** Let \( \Omega = \{ (\psi_n, M_n, P_n) \}_{n \in \mathbb{N}} \) be a module-sequenec 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
\[ \max \{ B_n, B_n L_n^2 R_n^2 \} \leq 1, \quad \forall n \in \mathbb{N}, \]
is referred to as admissibility condition. Module-sequeneces 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 \( \Psi_n \) (e.g., the directional wavelet frame introduced in Section II), and \( R_n \) 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 (13) (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 \( \Psi_n \) 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 \( \Phi_\Omega \)

A. Vertical translation invariance

The following theorem states that under very mild decay conditions on the Fourier transforms \( \hat{\chi}_n \) of the output-generating atoms \( \chi_n \), the feature extractor \( \Phi_\Omega \) 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], [21], where it is informally argued that the network outputs generated at deeper layers tend to be more translation-invariant.

Theorem 1. Let \( \Omega = \left( (\Psi_n, M_n, P_n) \right)_{n \in \mathbb{N}} \) 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
\[
\Phi_\Omega^n(T_t f) = T_{f/\Omega} \Phi_\Omega^n(f), \quad (19)
\]
for all \( f \in L^2(\mathbb{R}^d), t \in \mathbb{R}^d, \) and \( n \in \mathbb{N} \), where \( T_{f/\Omega} \) refers to element-wise application of \( T_t \), i.e., \( T_{f/\Omega} = \{ T_t h | \forall h \in \Phi_\Omega^n(f) \} \).

ii) If, in addition, there exists a constant \( K > 0 \) (that does not depend on \( n \)) such that the Fourier transforms \( \hat{\chi}_n \) of the output-generating atoms \( \chi_n \) satisfy the decay condition
\[
|\hat{\chi}_n(\omega)| |\omega| \leq K, \quad \text{a.e. } \omega \in \mathbb{R}^d, \quad \forall n \in \mathbb{N}, \quad (20)
\]
then
\[
||| \Phi_\Omega^n(T_t f) - \Phi_\Omega^n(f) ||| \leq \frac{2\pi |t| K}{S_1 \cdot S_2 \cdots S_n} ||f||_2, \quad (21)
\]
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 \( P f = f \ast \phi \), 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 \( \{ \chi_n \}_{n \in \mathbb{N}} \) either to satisfy
\[
\sup_{n \in \mathbb{N}} \{ ||\chi_n||_1 + ||\nabla \chi_n||_1 \} < \infty,
\]
see, e.g., [43, Ch. 7], or to be uniformly band-limited in the sense of \( \text{supp}(\hat{\chi}_n) \subseteq B_t(0) \), for all \( n \in \mathbb{N}, \) with \( t \) 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], [21], where it is informally argued that pooling is crucial to get translation invariance of the extracted features. Furthermore, the condition \( \lim_{n \to \infty} S_1 \cdot S_2 \cdot \ldots \cdot S_n = \infty \) (easily met by taking \( S_n \geq 1, \) for all \( n \in \mathbb{N} \)) guarantees, thanks to (21), asymptotically full translation invariance according to
\[
\lim_{n \to \infty} ||| \Phi_\Omega^n(T_t f) - \Phi_\Omega^n(f) ||| = 0, \quad (22)
\]
for all \( f \in L^2(\mathbb{R}^d) \) and \( t \in \mathbb{R}^d \). This means that the features
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 \( M_n : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d) \) and \( P_n : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d) \) commute with the translation operator \( T_t \) in the sense of \([10]\). 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 \( \chi_n \) satisfy the decay condition \([20]\), then
\[
||| \Phi^n_\Omega(T_t f) - T_t \Phi^n_\Omega(f) ||| \leq 2 \pi |t| K |1/(S_1 \ldots S_n) - 1| ||f||_2,
\]
for all \( f \in L^2(\mathbb{R}^d) \) and \( t \in \mathbb{R}^d \).

Proof. The proof is given in Appendix [4].

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
\[
\| S_n^{d/2} P_n(f)(S_n \cdot) \|_2 = S_n^d \int_{\mathbb{R}^d} |P_n(f)(S_n x)|^2 \, dx
= \int_{\mathbb{R}^d} |P_n(f)(x)|^2 \, dx = \| P_n(f) \|_2^2 = \| f \|_2^2,
\]
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 \( \Phi^n_\Omega(T_t f) = T_t \Phi^n_\Omega(f) \), 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.

**Definition 5.** A signal class \( C \subseteq L^2(\mathbb{R}^d) \) is called deformation-insensitive if there exist \( \alpha, \beta, C > 0 \) such that for all \( f \in C, \omega \in C(\mathbb{R}^d, \mathbb{R}), \) and (possibly non-linear) \( \tau \in C^1(\mathbb{R}^d, \mathbb{R}) \) with \( ||D\tau||_\infty \leq \frac{1}{2d} \) it holds that
\[
||f - F_{\tau,\omega} f||_2 \leq C (||\tau||_\alpha^\alpha + ||\omega||_\beta^\beta).
\]
The deformation sensitivity bound for the feature extractor is coupling into Lipschitz continuity of $\Phi$ (see Appendix H for the corresponding technical details). This "deformation-insensitive" signal classes are the class of $\Omega = \{\Psi_n, M_n, P_n\}_{n \in \mathbb{N}}$ be an admissible module-sequence. There exists a constant $C > 0$ that does not depend on $\Omega$ such that for all $f \in L^2(R^d)$, $\omega \in C(R^d, R)$, and $\tau \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ with $||D\tau||_\infty \leq \frac{1}{2\tau}$, the feature extractor $\Phi_{\Omega}$ satisfies

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

(26)

**Proof.** The proof is given in Appendix H. 

First, we note that the bound in (26) holds for $\tau$ with sufficiently "small" Jacobian matrix, i.e., as long as $||D\tau||_\infty \leq \frac{1}{2\tau}$. We can think of this condition on the Jacobian matrix as follows: Let $f$ be an image of the handwritten digit "5" (see Fig. 5(a)), then, \(\{F_{\tau,\omega} f \mid \|D\tau\|_\infty < \frac{1}{2\tau}\} \) 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. 3 (b) and (c)). The condition $||D\tau||_\infty \leq \frac{1}{2\tau}$ 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 $\{F_{\tau,\omega} f \mid \|D\tau\|_\infty < \frac{1}{2\tau}\}$ can differ. The strength of Theorem 2 derives itself from the fact that the only condition on the underlying module-sequence $\Omega$ needed is admissibility according to (17), which as outlined in Section III can be easily be obtained by normalizing the frame elements of $\Psi_n$, 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 (13) to be completely independent of $\Omega$. All this is thanks to the decoupling technique used to prove Theorem 2 being completely independent of the structures of the frames $\Psi_n$ 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 $\Phi$, not only those generated by DCNNs.

The bound (4) for scattering networks reported in (22, Theorem 2.12) depends upon first-order ($D\tau$) and second-order ($D^2\tau$) derivatives of $\tau$. In contrast, our bound (26) depends on $\|D\tau\|$ implicitly only as we need to impose the condition $||D\tau||_\infty \leq \frac{1}{2\tau}$ for the bound to hold.

8 The ensuing argument is taken from (40).

9 We note that the condition $||D\tau||_\infty \leq \frac{1}{2\tau}$ is needed for the bound (4) to hold as well.
difference by referring to \(\Psi\) as deformation stability bound and to our bound \(\| F_r f \|_2 \leq C \| \tau \|_\infty^\alpha\), \(\forall f \in C \subseteq L^2(\mathbb{R}^d)\), as deformation sensitivity bound.

The dependence of the upper bound in \(\| F_r f \|_2 \leq C \| \tau \|_\infty^\alpha\) 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 \(\| F_r f \|_2 \leq C \| \tau \|_\infty^\alpha\) 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 \((F_r f)(x) = f(x - \tau(x))\) according to

\[
\| f - F_r f \|_2 \leq C \| \tau \|_\infty^\alpha, \quad \forall f \in C \subseteq L^2(\mathbb{R}^d),
\]

for the signal classes \(C \subseteq L^2(\mathbb{R}^d)\) 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 \(C\) and are specified in \([40]\).

For the vertical translation invariance result in Theorem 1 to apply 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 \(\| F_r f \|_2 \leq C \| \tau \|_\infty^\alpha\) applies to scattering networks reported in \([22]\) for the space

\[
H_W := \{ f \in L^2(\mathbb{R}^d) : \| f \|_{H_W} < \infty \},
\]

where

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

and \((\Lambda_W)^n\) denotes the set of paths \(q = (\lambda_1, \ldots, \lambda_n)\) of length \(n\) with \(\lambda_1, \ldots, \lambda_n \in \Lambda_W\). While \([22\text{, p. 1350}]\) cites numerical evidence on the series \(\sum_{q \in (\Lambda_W)^n} \| U[q] f \|_2^2\) being finite for a large class of signals \(f \in L^2(\mathbb{R}^d)\), it seems difficult to establish this analytically, let alone show that

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

In contrast, the deformation sensitivity bound \(\| F_r f \|_2 \leq C \| \tau \|_\infty^\alpha\) applies provably to the space of \(R\)-band-limited functions \(L^2_R(\mathbb{R}^d)\). 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 \(L^2_R(\mathbb{R}^d)\) is, trivially, independent of the module-sequence \(\Omega\).

V. Final remarks and outlook

It is interesting to note that the frame lower bounds \(A_n > 0\) of the semi-discrete frames \(\Psi_n\) 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 \(\Psi_n = \{ T_b I g_\lambda \}_{b \in \mathbb{R}^d, \lambda \in \Lambda_n}\), for all \(n \in \mathbb{N}\), satisfy the Bessel property

\[
\sum_{\lambda \in \Lambda_n} \int_{\mathbb{R}^d} |(f, T_b I g_\lambda)|^2 db = \sum_{\lambda \in \Lambda_n} \| f * g_\lambda \|_2^2 \leq B_n \| f \|_2^2,
\]

for all \(f \in L^2(\mathbb{R}^d)\) for some \(B_n > 0\), \(\forall f \in L^2(\mathbb{R}^d)\).

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 \(\Psi_n = \{ T_b I g_\lambda \}_{b \in \mathbb{R}^d, \lambda \in \Lambda_n}\) 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 \(\Psi_n\), which may result in the frame coefficients \((f, T_b I g_\lambda) = (f * g_\lambda)(b), (\lambda_n, b) \in \Lambda_n \times \mathbb{R}^d\), 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 \((29)\) is prudent. Interestingly, satisfying the frame property \((29)\) for all \(\Psi_n, n \in \mathbb{Z}\), does, however, not guarantee that the feature extractor \(\Psi_{11}\) has a trivial null-space, i.e., \(\Phi_{11}(f) = 0\) if and only if \(f = 0\). We refer the reader to \([65]\) for an example of a feature extractor with non-trivial null-space.

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 \(\Lambda\). 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

\[
A \| f \|_2^2 \leq \sum_{\lambda \in \Lambda} \int_{\mathbb{R}^d} |(f, T_b I g_\lambda)|^2 db = \sum_{\lambda \in \Lambda} \| f * g_\lambda \|_2^2 \leq B \| f \|_2^2, \quad \forall f \in L^2(\mathbb{R}^d).
\]

The functions \(\{ g_\lambda \}_{\lambda \in \Lambda}\) are called the atoms of the frame \(\Psi_\Lambda\). 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 $\Psi_\Lambda$ is defined in the weak sense as $S_\Lambda : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$,
$$S_\Lambda f := \sum_{\lambda \in \Lambda} \int_{\mathbb{R}^d} \langle f, T_b I g_\lambda \rangle (T_b I g_\lambda) \, db = \left( \sum_{\lambda \in \Lambda} g_\lambda \ast I g_\lambda \right) \ast f,$$
where $\langle f, T_b I g_\lambda \rangle = \langle f \ast g_\lambda \rangle (b)$, $(\lambda, b) \in \Lambda \times \mathbb{R}^d$, are called the frame coefficients. $S_\Lambda$ 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 $\Lambda$ 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 $\Lambda_W := \{(-J,0)\} \cup \{(j,k) \mid j \in \mathbb{Z} \text{ with } j > -J, \ k \in \{0,\ldots,K-1\}\}$ 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 $\Lambda$ 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 $L^2(\mathbb{R}^d)$ if and only if
$$A \leq \sum_{\lambda \in \Lambda} |\hat{g}_\lambda(\omega)|^2 \leq B,$$ 
for all $\omega \in \mathbb{R}^d$.

**Proof.** The proof is standard and can be found, e.g., in [30].

**Remark 2.** What behind Proposition 2 is a result on the unitary equivalence between operators $[70]$ Definition 5.19.3). Specifically, Proposition 2 follows from the fact that the multiplier $\sum_{\lambda \in \Lambda} |\hat{g}_\lambda|^2$ is unitarily equivalent to the frame operator $S_\Lambda$ in (31) according to
$$FS_\Lambda F^{-1} = \sum_{\lambda \in \Lambda} |\hat{g}_\lambda|^2,$$
where $F : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ denotes the Fourier transform. We refer the interested reader to [77] where the framework of unitary equivalence was formalized in the context of shift-invariant frames for $l^2(\mathbb{Z})$.

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

ii) The family of functions $\Psi_\Lambda^2 := \{T_b I g_\lambda^2\}_{(\lambda, b) \in \Lambda \times \mathbb{R}^d}$, $g_\lambda^2 := \mathcal{F}^{-1}\left(\hat{g_\lambda^2}\right)$, $\forall \lambda \in \Lambda$, is a semi-discrete Parseval frame for $L^2(\mathbb{R}^d)$, i.e., the frame bounds satisfy $\tilde{A}^2 = \tilde{B}^2 = 1$.

**Proof.** We start by proving statement i). As $\Psi_\Lambda$ is a frame for $L^2(\mathbb{R}^d)$, we have
$$A\|f\|^2_2 \leq \sum_{\lambda \in \Lambda} \|f \ast g_\lambda\|^2_2 \leq B\|f\|^2_2,$$ 
for all $f \in L^2(\mathbb{R}^d)$. With $g_\lambda = \sqrt{C} \tilde{g}_\lambda$, for all $\lambda \in \Lambda$, in (33) we get $A\|f\|^2_2 \leq \sum_{\lambda \in \Lambda} \|f \ast \sqrt{C} \tilde{g}_\lambda\|^2_2 \leq B\|f\|^2_2$, for all $f \in L^2(\mathbb{R}^d)$, which is equivalent to $\frac{1}{C}\|f\|^2_2 \leq \sum_{\lambda \in \Lambda} \|f \ast \tilde{g}_\lambda\|^2_2 \leq \frac{B}{C}\|f\|^2_2$, for all $f \in L^2(\mathbb{R}^d)$, and hence establishes i). To prove statement ii), we first note that $\tilde{g}_\lambda^2 = \hat{g}_\lambda^2\left(\sum_{\lambda \in \Lambda} \hat{g}_\lambda(\omega)^2\right)^{-1/2}$, for all $\lambda \in \Lambda$, and thus $\sum_{\lambda \in \Lambda} (\tilde{g}_\lambda^2(\omega)^2) = \sum_{\lambda \in \Lambda} (\hat{g}_\lambda(\omega)^2)\left(\sum_{\lambda \in \Lambda} \hat{g}_\lambda(\omega)^2\right)^{-1/2} = 1$, a.e. $\omega \in \mathbb{R}^d$.

Application of Proposition 3 then establishes that $\Psi_\Lambda^2$ is a semi-discrete Parseval frame for $L^2(\mathbb{R}^d)$, i.e., the frame bounds satisfy $\tilde{A}^2 = \tilde{B}^2 = 1$.

**APPENDIX B**

**EXAMPLES OF SEMI-DISCRETE FRAMES IN 1-D**

General 1-D semi-discrete frames are given by collections $\Psi = \{T_b I g_k\}_{(k, b) \in \mathbb{Z} \times \mathbb{R}}$ (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
$$A \leq \sum_{k \in \mathbb{Z}} \hat{g_k}(\omega)^2 \leq B,$$ 
for all $\omega \in \mathbb{R}$.

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 (33), where $g_k(x) := e^{2\pi ikx} g(x)$, $k \in \mathbb{Z}$, 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
$$A \leq \sum_{k \in \mathbb{Z}} \hat{g}(\omega - k)^2 \leq B,$$ 
for all $\omega \in \mathbb{R}$.

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

**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 tasks...
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

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

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

**Semi-discrete curvelet frames:** Curvelets, introduced in [34], [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].

**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], [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 $\Psi_{\text{TW}} := \{T_b I g_{(e,j)}(x)\}_{(e,j) \in \mathbb{A}_b \times \mathbb{B}}$, where $A_{\text{TW}} := \{(0,0), 0\} \cup \{(e,j) \mid e \in E \cap \{(0,0)\}, j \geq 0\}$, and $E := \{0, 1\}$. Here, the functions $\psi^e \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}^2)$ and a fine-scale function $\psi \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}^2)$ 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

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

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

$$\Psi_{\text{Dw}} := \{T_b I g_{(j,k)}\}_{(j,k) \in \mathbb{A}_{\text{Dw}} \times \mathbb{B}},$$

with $g_{(-l,0)}(x) := 2^{-2J} \phi(2^{-2J} x)$, $g_{(j,k)}(x) := 2^j \psi(2^j R_\theta x)$, where $A_{\text{Dw}} := \{(-J, 0)\} \cup \{(j,k) \mid j \in \mathbb{Z} \text{ with } j > -J, k \in \{0, \ldots, K-1\}\}$, $R_\theta$ is a $2 \times 2$ 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 $k = 0, \ldots, K-1$, 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 $\{g_{(j,k)}\}_{(j,k) \in \mathbb{A}_{\text{Dw}}}$ satisfy the Littlewood-Paley condition (32) according to

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

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

A semi-discrete curvelet frame for $L^2(\mathbb{R}^2)$ is a collection of functions according to $\Psi_{\text{Ac}} := \{T_b I g_{(j,l)}\}_{(j,l) \in \mathbb{A}_{\text{Ac}} \times \mathbb{B}}$, with $g_{(-l,0)}(x) := \phi(x)$, $g_{(j,l)}(x) := \psi_j(R_{\theta_j} x)$, where $A_{\text{Ac}} := \{(-1, 0)\} \cup \{(j,l) \mid j \geq 0, l = 0, \ldots, L_j-1\}$, $R_{\theta_j}$ is the rotation matrix defined in (39), and $\theta_{j,l} := \pi(2^{-\lfloor j/2 \rfloor} - 1)$, for $j \geq 0$, and $0 \leq l < L_j := 2^{J/2+2}$, 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

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

A.e. $\omega \in \mathbb{R}^2$. The $\psi_j, j \geq 0$, are designed to have their Fourier transforms $\hat{\psi}_j$ supported on a pair of opposite wedges of size...
specified by the parameter $\Lambda$ we need to verify the following:

\[ L^2 \times \omega \quad (32) \] according to atoms $\{ \} \in$ the dyadic corona

Semi-discrete ridgelet frames: Ridgelets, introduced in [79, 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 $\mathcal{A}_R := \{ T_{b}g_{j,l}(x) \}_{(j,l) \in \mathcal{A}_R} \in \mathbb{C}^2$, with $g_{0,0}(x) := \phi(x)$, $g_{j,l}(x) := \psi_{j,l}(x)$, where $\mathcal{A}_R := \{ (0,0) \} \cup \{ (j,l) \mid j \geq 1, l = 1, \ldots, 2^j - 1 \}$, and the atoms $\{ g_{j,l}(x) \}_{(j,l) \in \mathcal{A}_R}$ satisfy the Littlewood-Paley condition [33] according to

\[ A \leq \int_0^\infty \left| \hat{\phi}(\omega) \right|^2 + \sum_{j=1}^{2^j-1} \sum_{l=1}^{2^j-1} \left| \hat{\psi}_{j,l}(\omega) \right|^2 \leq B, \] (42)

where $\omega \in \mathbb{R}^2$. The $\psi_{j,l} \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$, $(j,l) \in \mathcal{A}_R \setminus \{(0,0)\}$, are designed to be constant in the direction specified by the parameter $2^{-j} \times 2^j$ 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, Proposition 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 [37], which deals with semi-discrete $\alpha$-curvelet frames.

APPENDIX D

NON-LINEARITIES

This appendix gives a brief overview of non-linearities $M : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ 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 $M : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ 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 $f = 0$.

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

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

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

\[ (MT_t f)(x) = \rho((T_t f)(x)) = \rho(f(x-t)) = T_t \rho(f(x)) = (T_t Mf)(x), \quad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}. \]

Modulus function: The modulus function

\[ | \cdot | : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d), \quad |f(x)| := |f(x)|, \]

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

\[ \|f| - |h\|_2^2 = \int_{\mathbb{R}^d} \|f(x) - |h(x)||^2 dx \leq \int_{\mathbb{R}^d} \|f(x) - h(x)||^2 dx = \|f - h\|_2^2, \]

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 $R : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$,

\[ (Rf)(x) := \max\{0, \text{Re}(f(x))\} + i \max\{0, \text{Im}(f(x))\}. \]

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

\[ |(Rf)(x) - (Rh)(x)| = \max\{0, \text{Re}(f(x))\} + i \max\{0, \text{Im}(f(x))\} \]

\[ \leq \max\{0, \text{Re}(f(x))\} + i \max\{0, \text{Im}(f(x))\} \]

\[ = \max\{0, \text{Re}(h(x))\} + i \max\{0, \text{Im}(h(x))\}| \]

\[ \leq |\text{Re}(f(x)) - \text{Re}(h(x))| + |\text{Im}(f(x)) - \text{Im}(h(x))| \]

\[ \leq |f(x) - h(x)| + |f(x) - h(x)| = 2|f(x) - h(x)|, \]

where we used the triangle inequality in (44).

\[ | \max\{0, a\} - \max\{0, b\}| = |a - b|, \quad \forall a, b \in \mathbb{R}, \]

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, Theorem 4.1] for constructions of functions $\phi, \psi_j$ satisfying (41) with $A = B = 1$.

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

$$\|Rf - Rh\|_2 = \left( \int_{\mathbb{R}^d} |(Rf)(x) - (Rh)(x)|^2 \, dx \right)^{1/2}$$

$$\leq 2 \left( \int_{\mathbb{R}^d} |f(x) - h(x)|^2 \, dx \right)^{1/2}$$

$$= 2 \|f - h\|_2,$$

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 $H : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$,

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

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

$$|\langle Hf \rangle(x) - \langle Hh \rangle(x) |$$

$$= \left| \tanh(\text{Re}(f(x))) + i \tanh(\text{Im}(f(x))) \right|$$

$$- \left| \tanh(\text{Re}(h(x))) + i \tanh(\text{Im}(h(x))) \right|$$

$$\leq \left| \tanh(\text{Re}(f(x))) - \tanh(\text{Re}(h(x))) \right|$$

$$+ \left| \tanh(\text{Im}(f(x))) - \tanh(\text{Im}(h(x))) \right|,$$

(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 $h : \mathbb{R} \to \mathbb{R}$ 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] Theorem 9.5.1. \hfill \square

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

$$|\langle Hf \rangle(x) - \langle Hh \rangle(x) |,$$

$$\leq \left| \text{Re}(f(x)) - \text{Re}(h(x)) \right|$$

$$+ \left| \text{Im}(f(x)) - \text{Im}(h(x)) \right|$$

$$\leq \left| f(x) - h(x) \right| + \left| f(x) - h(x) \right|$$

$$= 2 \left| f(x) - h(x) \right|.$$

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

$$\|Hf - Hh\|_2 = \left( \int_{\mathbb{R}^d} |\langle Hf \rangle(x) - \langle Hh \rangle(x) |^2 \, dx \right)^{1/2}$$

$$\leq 2 \left( \int_{\mathbb{R}^d} |f(x) - h(x)|^2 \, dx \right)^{1/2}$$

$$= 2 \|f - h\|_2,$$

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-linearity (see, e.g., [28], [29]) is defined as $P : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$,

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

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

$$\|\langle Pf \rangle(x) - \langle Ph \rangle(x) |$$

$$\leq \frac{1}{4} \left| \text{Re}(f(x)) - \text{Re}(h(x)) \right|$$

$$+ \frac{1}{4} \left| \text{Im}(f(x)) - \text{Im}(h(x)) \right|$$

$$\leq \frac{1}{2} \left| f(x) - h(x) \right|,$$

(49)

It now follows from (49) that

$$\|Pf - Ph\|_2 = \left( \int_{\mathbb{R}^d} |\langle Pf \rangle(x) - \langle Ph \rangle(x) |^2 \, dx \right)^{1/2}$$

$$\leq \frac{1}{2} \left( \int_{\mathbb{R}^d} |f(x) - h(x)|^2 \, dx \right)^{1/2}$$

$$= \frac{1}{2} \|f - h\|_2,$$

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

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

(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 $\|f_q\|_2 \leq \|f\|_2 < \infty$, and thus $f_q \in L^2(\mathbb{R}^d)$. The key idea of the proof is now—similarly

\[10\]Strictly 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], [29]. We incorporated the offset $\frac{1}{2}$ in order to satisfy the requirement $Pf = 0$ for $f = 0$.\]
to the proof of [22] Proposition 2.5)—to judiciously employ a telescoping series argument. We start by writing
\[
\left\| \mathbf{\Phi}(f) \right\|^2 = \sum_{n=0}^{\infty} \sum_{q \in \Lambda_n^*} \left\| f_q \ast \chi_n \right\|^2 \leq \lim_{N \to \infty} \sum_{n=0}^{N} \sum_{q \in \Lambda_n^*} \left\| f_q \ast \chi_n \right\|^2.
\]

(51)

The key step is then to establish that \( a_n \) can be upper-bounded according to
\[
a_n \leq b_n - b_{n+1}, \quad \forall n \in \mathbb{N}_0,
\]

(52)

with \( b_n := \sum_{q \in \Lambda_n^*} \left\| f_q \right\|^2 \), \( n \in \mathbb{N}_0 \), and to use this result in a telescoping series argument according to
\[
\sum_{n=0}^{N} a_n \leq \left( b_0 - b_{n+1} \right) + \left( b_1 - b_{n+1} \right) + \cdots + \left( b_{n-1} - b_{n+1} \right) = b_0 - b_{n+1} \geq 0 \tag{53}
\]

\[
\leq b_0 = \sum_{q \in \Lambda_0^*} \left\| f_q \right\|^2 = \| U[\varepsilon]f \|^2 = \| f \|^2. \tag{54}
\]

By (51) this then implies (50). We start by noting that (52) reads
\[
\sum_{q \in \Lambda_n^*} \left\| f_q \right\|^2 \leq \sum_{q \in \Lambda_n^*} \left\| f_q \right\|^2 - \sum_{q \in \Lambda_n^*} \| f_q \|^2, \tag{55}
\]

for all \( n \in \mathbb{N}_0 \), and proceed by examining the second term on the right hand side (RHS) of (55). Every path
\[
\hat{q} \in \Lambda_{n+1}^n = \Lambda_1 \times \cdots \times \Lambda_n \times \Lambda_{n+1}^n
\]
of length \( n+1 \) can be decomposed into a path \( q \in \Lambda_1^n \) of length \( n \) and an index \( \lambda_{n+1} \in \Lambda_{n+1}^n \) according to \( \hat{q} = (q, \lambda_{n+1}) \).

Thanks to (13) we have
\[
U[q] = U[(q, \lambda_{n+1})] = U_{n+1}[\lambda_{n+1}] U[q],
\]

which yields
\[
\sum_{q \in \Lambda_{n+1}^n} \left\| f_q \right\|^2 = \sum_{q \in \Lambda_1^n \lambda_{n+1} \in \Lambda_{n+1}^n} \| U_{n+1}[\lambda_{n+1}] f_q \|^2. \tag{56}
\]

Substituting the second term on the RHS of (55) by (56) now yields
\[
\sum_{q \in \Lambda_1^n \lambda_{n+1} \in \Lambda_{n+1}^n} \left( \| f_q \|^2 - \sum_{\lambda_{n+1} \in \Lambda_{n+1}^n} \| U_{n+1}[\lambda_{n+1}] f_q \|^2 \right) = \sum_{q \in \Lambda_1^n \lambda_{n+1} \in \Lambda_{n+1}^n} \left( \| f_q \|^2 - \sum_{\lambda_{n+1} \in \Lambda_{n+1}^n} \| U_{n+1}[\lambda_{n+1}] f_q \|^2 \right) \geq 0, \quad \forall n \in \mathbb{N}_0,
\]

which can be rewritten as
\[
\sum_{q \in \Lambda_1^n \lambda_{n+1} \in \Lambda_{n+1}^n} \left( \| f_q \|^2 + \sum_{\lambda_{n+1} \in \Lambda_{n+1}^n} \| U_{n+1}[\lambda_{n+1}] f_q \|^2 \right) \tag{57}
\]

\[
\leq \sum_{q \in \Lambda_1^n \lambda_{n+1} \in \Lambda_{n+1}^n} \| f_q \|^2, \quad \forall n \in \mathbb{N}_0.
\]

Next, note that the second term inside the sum on the left hand side (LHS) of (57) can be written as
\[
\sum_{\lambda_{n+1} \in \Lambda_{n+1}^n} \| U_{n+1}[\lambda_{n+1}] f_q \|^2
\]

\[
= \sum_{\lambda_{n+1} \in \Lambda_{n+1}^n} \int_{\mathbb{R}^d} |(U_{n+1}[\lambda_{n+1}] f_q)(x)|^2 \, dx
\]

\[
= \sum_{\lambda_{n+1} \in \Lambda_{n+1}^n} \int_{\mathbb{R}^d} |P_{n+1}(M_{n+1}(f_q \ast g_{\lambda_{n+1}}))(S_n x)|^2 \, dx
\]

\[
= \sum_{\lambda_{n+1} \in \Lambda_{n+1}^n} \int_{\mathbb{R}^d} |P_{n+1}(M_{n+1}(f_q \ast g_{\lambda_{n+1}}))(y)|^2 \, dy
\]

\[
= \sum_{\lambda_{n+1} \in \Lambda_{n+1}^n} \| P_{n+1}(M_{n+1}(f_q \ast g_{\lambda_{n+1}})) \|^2,
\]

(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 \( (f_q \ast g_{\lambda_{n+1}}) \in L^2(\mathbb{R}^d) \) thanks to Young’s inequality [63 Theorem 1.2.12]. We use the Lipschitz property of \( M_{n+1} \) and \( P_{n+1} \), i.e., \( \| M_{n+1} f_q \ast g_{\lambda_{n+1}} - M_{n+1} h \| \leq L_{n+1} \| f_q \ast g_{\lambda_{n+1}} - h \| \), and \( \| P_{n+1} f_q \ast g_{\lambda_{n+1}} - P_{n+1} h \| \leq R_{n+1} \| f_q \ast g_{\lambda_{n+1}} - h \| \), 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
\[
\| P_{n+1}(M_{n+1}(f_q \ast g_{\lambda_{n+1}})) \| \leq R_{n+1} \| f_q \ast g_{\lambda_{n+1}} \|, \quad \forall n \in \mathbb{N}_0.
\]

Substituting the second term inside the sum on the LHS of (57) by the upper bound resulting from insertion of (59) into (58) yields
\[
\sum_{q \in \Lambda_1^n \lambda_{n+1} \in \Lambda_{n+1}^n} \left( \| f_q \|^2 + L_{n+1}^2 R_{n+1}^2 \sum_{\lambda_{n+1} \in \Lambda_{n+1}^n} \| f_q \ast g_{\lambda_{n+1}} \|^2 \right)
\]

\[
\leq \sum_{q \in \Lambda_1^n \lambda_{n+1} \in \Lambda_{n+1}^n} \max\{1, L_n^2 R_{n+1}^2\} \left( \| f_q \|^2 + \sum_{\lambda_{n+1} \in \Lambda_{n+1}^n} \| f_q \ast g_{\lambda_{n+1}} \|^2 \right), \quad \forall n \in \mathbb{N}_0.
\]

(60)

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

which, when used in (60) yields
\[
\sum_{q \in \Lambda_1^n \lambda_{n+1} \in \Lambda_{n+1}^n} \left( \| f_q \|^2 + \sum_{\lambda_{n+1} \in \Lambda_{n+1}^n} \| U_{n+1}[\lambda_{n+1}] f_q \|^2 \right)
\]

\[
\leq \sum_{q \in \Lambda_1^n \lambda_{n+1} \in \Lambda_{n+1}^n} \max\{1, L_n^2 R_{n+1}^2\} B_{n+1} \| f_q \|^2
\]

\[
= \sum_{q \in \Lambda_1^n \lambda_{n+1} \in \Lambda_{n+1}^n} \max\{B_{n+1}, B_{n+1}L_n^2 R_{n+1}^2\} \| f_q \|^2,
\]

(61)

for all \( n \in \mathbb{N}_0 \). Finally, invoking the assumption
\[
\max\{B_{n+1}, B_{n+1}L_n^2 R_{n+1}^2\} \leq 1, \quad \forall n \in \mathbb{N},
\]

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
\[
U_n[\Lambda_n]T_t f = T_{t/S_n} U_n[\Lambda_n] f,
\]
for all \( f \in L^2(\mathbb{R}^d), t \in \mathbb{R}^d \), and \( \Lambda_n \in \Lambda_n \). With the definition of \( U[\cdot] \) in [13] this then yields
\[
U[\cdot]T_t f = T_{t/(S_1 \cdots S_n)} U[\cdot] f,
\]
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:
\[
\Phi_\Omega^n(T_t f) = \left\{ \left( U[\cdot]T_t f \right) * \chi_n \right\}_{q \in \Lambda_1^n}
= \left\{ \left( T_{t/(S_1 \cdots S_n)} U[\cdot] f \right) * \chi_n \right\}_{q \in \Lambda_1^n}
= \left\{ T_{t/(S_1 \cdots S_n)} \left( \left( U[\cdot] f \right) * \chi_n \right) \right\}_{q \in \Lambda_1^n}
= T_{t/(S_1 \cdots S_n)} \Phi_\Omega^n(f), \quad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}^d.
\]
To establish [62], we first define the unitary operator \( D_n : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d), D_n f := S_n^{d/2} f(S_n \cdot) \), and note that
\[
U_n[\Lambda_n]T_t f = S_n^{d/2} P_n \left( M_n \left( \left( T_t f \right) * g_{\Lambda_n} \right) \right) (S_n \cdot)
= D_n P_n \left( M_n \left( \left( T_t f \right) * g_{\Lambda_n} \right) \right)
= D_n P_n \left( M_n \left( T_t (f * g_{\Lambda_n}) \right) \right)
= D_n P_n \left( T_t \left( M_n (f * g_{\Lambda_n}) \right) \right)
= D_n T_t \left( P_n \left( \left( M_n (f * g_{\Lambda_n}) \right) \right) \right),
\]
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
\[
D_n T_t f = S_n^{d/2} f(S_n \cdot - t) = S_n^{d/2} f(S_n \cdot / t / S_n)
= T_{t/S_n} D_n f, \quad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}^d,
\]
in [65] yields
\[
U_n[\Lambda_n]T_t f = D_n T_t \left( P_n \left( \left( M_n (f * g_{\Lambda_n}) \right) \right) \right)
= T_{t/S_n} \left( D_n P_n \left( \left( M_n (f * g_{\Lambda_n}) \right) \right) \right)
= T_{t/S_n} U_n[\Lambda_n] f,
\]
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 \( f_q := U[\cdot] f \), for \( f \in L^2(\mathbb{R}^d) \). Thanks to [44], 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
\[
\left\| \Phi_\Omega^n(T_t f) - \Phi_\Omega^n(f) \right\|^2
= \left\| T_{t/(S_1 \cdots S_n)} \Phi_\Omega^n(f) - \Phi_\Omega^n(f) \right\|^2
= \sum_{q \in \Lambda_1^n} \left\| T_{t/(S_1 \cdots S_n)} (f_q * \chi_n) - f_q * \chi_n \right\|^2
= \sum_{q \in \Lambda_1^n} \left\| M_{-t/(S_1 \cdots S_n)} (f_q * \chi_n) - f_q * \chi_n \right\|^2,
\]
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, Theorem 1.2.12]) together with the relation
\[
\overline{T_t f} = M_{-t} \overline{f}, \quad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}^d.
\]
The key step is then to establish the upper bound
\[
\left\| M_{-t/(S_1 \cdots S_n)} (f_q * \chi_n) - f_q * \chi_n \right\|^2
\leq \frac{4\pi^2 |t|^2 K^2}{(S_1 \cdots S_n)^2} \| f_q \|^2, \quad \forall n \in \mathbb{N},
\]
where \( K > 0 \) corresponds to the constant in the decay condition [20], and to note that
\[
\sum_{q \in \Lambda_1^n} \| f_q \|^2 \leq \sum_{q \in \Lambda_1^{n-1}} \| f_q \|^2, \quad \forall n \in \mathbb{N},
\]
which follows from [62] thanks to
\[
0 \leq \sum_{q \in \Lambda_1^{n-1}} \| f_q \|^2 - \sum_{q \in \Lambda_1^n} \| f_q \|^2 = \sum_{q \in \Lambda_1^n} \| f_q \|^2, \quad \forall n \in \mathbb{N}.
\]
Iterating on [69] yields
\[
\sum_{q \in \Lambda_1^n} \| f_q \|^2 \leq \sum_{q \in \Lambda_1^{n-1}} \| f_q \|^2 \leq \cdots \leq \sum_{q \in \Lambda_1^1} \| f_q \|^2
= \| U[\cdot] f \|^2 \leq \| f \|^2, \quad \forall n \in \mathbb{N}.
\]
The identity [67] together with the inequalities [68] and [72] then directly imply
\[
\left\| \Phi_\Omega^n(T_t f) - \Phi_\Omega^n(f) \right\|^2 \leq \frac{4\pi^2 |t|^2 K^2}{(S_1 \cdots S_n)^2} \| f \|^2,
\]
for all \( n \in \mathbb{N} \). It remains to prove [68]. To this end, we first note that
\[
\left\| M_{-t/(S_1 \cdots S_n)} (f_q * \chi_n) - f_q * \chi_n \right\|^2
= \int_{\mathbb{R}^d} e^{-2\pi i (t, \omega)/(S_1 \cdots S_n)} - 1 \left( |\hat{\chi}_n(\omega)|^2 \right) |\hat{f}(\omega)|^2 d\omega.
\]
Since \( |e^{-2\pi i x} - 1| \leq 2|\pi | x | \), for all \( x \in \mathbb{R} \), it follows that
\[
|e^{-2\pi i (t, \omega)/(S_1 \cdots S_n)} - 1| \leq \frac{4\pi^2 |t, \omega|^2}{(S_1 \cdots S_n)^2}
\leq \frac{4\pi^2 |t|^2 |\omega|^2}{(S_1 \cdots S_n)^2},
\]
where in the last step we employed the Cauchy-Schwartz
inequality. Substituting (75) into (74) yields
\[
\| M_{-t/(S_1 \ldots S_n)} f_q \ast \chi_n \|_2^2 \\
\leq \frac{4\pi^2 |t|^2}{(S_1 \ldots S_n)^2} \int_{\mathbb{R}^d} |\omega|^2 |\chi_n(\omega)|^2 |\hat{f}_q(\omega)|^2 d\omega \\
\leq \frac{4\pi^2 |t|^2 K^2}{(S_1 \ldots S_n)^2} \int_{\mathbb{R}^d} |\hat{f}_q(\omega)|^2 d\omega \\
= \psi \| f_q \|_2^2 = \psi \| f_q \|_2^2, \\
\] (76)
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, let \( f_q := U[g] f \), 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
\[
\| | \| \Phi(f) - \Phi(f) \|_2 \| = \| \| T(f) - T(f) \|_2 \|_2 \|
= \sum_{q \in \mathbb{N}^+} \| (T(f) - T(f)) \|_2 \|
= \sum_{q \in \mathbb{N}^+} \| (M_{-t/(S_1 \ldots S_n)} - M_{-t}) f_q \ast \chi_n \|_2 \|
\] (78)
for all \( q \in \mathbb{N} \), where in (78) we used (19), and in (79) we employed Parseval’s formula [43, p. 189] (noting that \( f_q \ast \chi_n \in L^2(\mathbb{R}^d) \) thanks to Young’s inequality [63, Theorem 1.2.12]) together with the relation
\[
\hat{T}f = M_{-t} f, \quad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}^d. 
\]
The key step is then to establish the upper bound
\[
\| (M_{-t/(S_1 \ldots S_n)} - M_{-t}) f_q \ast \chi_n \|_2 \leq \frac{4\pi^2 |t|^2 K^2}{(S_1 \ldots S_n)^2} \frac{1}{|S_1 \ldots S_n|} - \frac{1^2}{\| f_q \|_2^2}, \\
\] (80)
where \( K \geq 0 \) 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):
\[
\| (M_{-t/(S_1 \ldots S_n)} - M_{-t}) f_q \ast \chi_n \|_2 \leq \int_{\mathbb{R}^d} \left| e^{2\pi i t(x)}/(S_1 \ldots S_n) - e^{-2\pi i t(x)}/(S_1 \ldots S_n) \right|^2 |\chi_n(\omega)|^2 \| f_q(\omega) \|_2^2 d\omega. \\
\] (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
\[
\left| e^{-2\pi i (t(x))/|S_1 \ldots S_n|} - e^{2\pi i (t(x))/|S_1 \ldots S_n|} \right|^2 \leq 4\pi^2 |t|^2 |\omega|^2 \frac{1}{|S_1 \ldots S_n|} - 1^2, \\
\] (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 14.B the proof of the deformation sensitivity bound (26) is based on two key ingredients. The first one, stated in Proposition 1 in Appendix G, establishes that the feature extractor \( \Phi(f) \) is Lipschitz-continuous with Lipschitz constant \( L_\Phi = 1 \), i.e.,
\[
\| | \| \Phi(f) - \Phi(f) \|_2 \| \leq \| f - f \|_2, \quad \forall f, h \in L^2(\mathbb{R}^d), \\
\] (83)
and needs the admissibility condition (17) only. The second ingredient, stated in Proposition 2 in Appendix G, is an upper bound on the deformation error \( \| f - F_{\tau,w}f \|_2 \) given by
\[
\| f - F_{\tau,w}f \|_2 \leq C(R)|\tau|_\infty + |\omega|_\infty \| f \|_2, \\
\] (84)
for all \( f \in L^2(\mathbb{R}^d) \), \( \omega \in C(\mathbb{R}, \mathbb{R}) \) and \( \tau \in C^1(\mathbb{R}^d, \mathbb{R}^d) \) \( \| D\tau \|_\infty \leq \frac{1}{2\pi} \). 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 \( h := F_{\tau,w}f \) for \( f \in L^2(\mathbb{R}^d) \), and we used \( F_{\tau,w}f \in L^2(\mathbb{R}^d) \), which is thanks to
\[
\| F_{\tau,w}f \|_2 \leq \int_{\mathbb{R}^d} |f(x - \tau(x))| d\omega \leq \| f \|_2, \\
\] obtained through the change of variables \( u = x - \tau(x) \), together with
\[
\frac{du}{dx} = |\det(E - (D\tau)(x))| \geq 1 - d\| D\tau \|_\infty \geq 1/2, \\
\] (86)
for \( x \in \mathbb{R}^d \). The first inequality in (86) follows from:

**Lemma 2.** [82, Corollary 1] Let \( M \in \mathbb{R}^{d \times d} \) be such that \( |M_{i,j}| \leq \sigma \), for all \( i, j \) with \( 1 \leq i, j \leq d \). If \( d\sigma \leq 1 \), then
\[
|\det(E - M)| \geq 1 - d\sigma. \\
\]
The second inequality in (86) is a consequence of the assumption \( \| D\tau \|_\infty \leq \frac{1}{2\pi} \). The proof is finalized by replacing the RHS of (85) by the RHS of (84).

**APPENDIX I**

**Proposition 3**

**Proposition 4.** Let \( \Omega = (\Psi_n, M_n, P_n)_{n \in \mathbb{N}} \) be an admissible module-sequence. The corresponding feature extractor \( \Phi(f) : L^2(\mathbb{R}^d) \to (L^2(\mathbb{R}^d))^\mathbb{N} \) is Lipschitz-continuous with Lipschitz constant \( L\Phi = 1 \), i.e.,
\[
\| | \| \Phi(f) - \Phi(f) \|_2 \| \leq \| f - f \|_2, \quad \forall f, h \in L^2(\mathbb{R}^d). \\
\] (87)

**Remark 4.** Proposition 3 generalizes Proposition 2.5, which shows that the wavelet-modulus feature extractor \( \Phi(f) \) generated by scattering networks is Lipschitz-continuous with Lipschitz constant \( L\Phi = 1 \). 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 again—similarly to the proof of Proposition 1 in Appendix E—to judiciously employ a telescoping series argument. For ease of notation, 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

$$
\|\phi_\Omega(f) - \phi_\Omega(h)\|_2^2 = \lim_{N \to \infty} \sum_{n=0}^{N} \sum_{q \in \Lambda_2^n} \|f_q \star \chi_n - h_q \star \chi_n\|_2^2.
$$

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

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

where here

$$
b_n := \sum_{q \in \Lambda_2^n} \|f_q - h_q\|_2^2, \quad \forall n \in \mathbb{N}_0,
$$

and to note that, similarly to (54),

$$
\sum_{n=0}^{N} a_n \leq \sum_{n=0}^{N} (b_n - b_{n+1}) = (b_0 - b_1) + (b_1 - b_2) + \cdots + (b_N - b_{N+1}) = b_0 - b_{N+1} \leq b_0 - b_{N+1} \geq 0
$$

$$
= \|f - h\|_2^2,
$$

which then yields (87) according to

$$
\|\phi_\Omega(f) - \phi_\Omega(h)\|_2^2 = \lim_{N \to \infty} \sum_{n=0}^{N} a_n \leq \lim_{N \to \infty} \|f - h\|_2^2 = \|f - h\|_2^2.
$$

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

$$
\sum_{q \in \Lambda_2^n} \|f_q \star \chi_n - h_q \star \chi_n\|_2^2 \\
\leq \sum_{q \in \Lambda_2^n} \|f_q - h_q\|_2^2 - \sum_{q \in \Lambda_1^{n+1}} \|f_q - h_q\|_2^2.
$$

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

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

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

$$
\sum_{\hat{q} \in \Lambda_1^{n+1}} \|f_{\hat{q}} - h_{\hat{q}}\|_2^2 = \sum_{\hat{q} \in \Lambda_1^{n+1}} \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|U_{n+1}[\lambda_{n+1}]f_{\hat{q}} - U_{n+1}[\lambda_{n+1}]h_{\hat{q}}\|_2^2.
$$

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

$$
\sum_{\hat{q} \in \Lambda_1^{n+1}} \|f_{\hat{q}} \star \chi_n - h_{\hat{q}} \star \chi_n\|_2^2 \\
+ \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|U_{n+1}[\lambda_{n+1}]f_{\hat{q}} - U_{n+1}[\lambda_{n+1}]h_{\hat{q}}\|_2^2 \\
\leq \sum_{\hat{q} \in \Lambda_1^{n+1}} \|f_{\hat{q}} - h_{\hat{q}}\|_2^2, \quad \forall n \in \mathbb{N}_0.
$$

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

$$
\sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|U_{n+1}[\lambda_{n+1}]f_{\hat{q}} - U_{n+1}[\lambda_{n+1}]h_{\hat{q}}\|_2^2 \\
\leq \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|P_{n+1}(M_{n+1}(f_{\hat{q}} \ast g_{\lambda_{n+1}})) \\
- P_{n+1}(M_{n+1}(h_{\hat{q}} \ast g_{\lambda_{n+1}})))\|_2^2,
$$

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

$$
\sum_{\hat{q} \in \Lambda_1^{n+1}} \|f_{\hat{q}} \star \chi_n - h_{\hat{q}} \star \chi_n\|_2^2 \\
+ \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|U_{n+1}[\lambda_{n+1}]f_{\hat{q}} - U_{n+1}[\lambda_{n+1}]h_{\hat{q}}\|_2^2 \\
\leq \sum_{\hat{q} \in \Lambda_1^{n+1}} \|(f_{\hat{q}} - h_{\hat{q}}) \ast \chi_n\|_2^2 \\
+ \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|(f_{\hat{q}} - h_{\hat{q}}) \ast g_{\lambda_{n+1}}\|_2^2,
$$

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 $\Psi_{n+1}$ for $L^2(\mathbb{R}^d)$ and $f_q, h_q \in L^2(\mathbb{R}^d)$, as established above, we have

$$
\|(f_{\hat{q}} - h_{\hat{q}}) \ast \chi_n\|_2^2 + \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|(f_{\hat{q}} - h_{\hat{q}}) \ast g_{\lambda_{n+1}}\|_2^2 \\
\leq B_{n+1}\|f_{\hat{q}} - h_{\hat{q}}\|_2^2,
$$

which, when used in (93) yields

$$
\sum_{\hat{q} \in \Lambda_1^{n+1}} \|f_{\hat{q}} \star \chi_n - h_{\hat{q}} \star \chi_n\|_2^2 \\
+ \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|U_{n+1}[\lambda_{n+1}]f_{\hat{q}} - U_{n+1}[\lambda_{n+1}]h_{\hat{q}}\|_2^2 \\
\leq \sum_{\hat{q} \in \Lambda_1^{n+1}} \max\{B_{n+1}, B_{n+1}L_{n+1}^2 R_{n+1}^2\}\|f_{\hat{q}} - h_{\hat{q}}\|_2^2,
$$

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

$$
\max\{B_{n}, B_{n}L_{n}^2 R_{n}^2\} \leq 1, \quad \forall n \in \mathbb{N},
$$

and $\lambda_{n+1} \in \Lambda_{n+1}$.
in (94) we get (91) and hence (88). This completes the proof.

APPENDIX J
PROPOSITION 5

Proposition 5. There exists a constant $C > 0$ such that for all $f \in L^2(R^d)$, $\omega \in C(R^d, R)$, and $\tau \in C^1(R^d, R^d)$ with $\|D\tau\|_\infty < \frac{1}{\pi}$, it holds that
\[ \|f - F_{\tau, \omega} f\|_2 \leq C(R\|\tau\|_\infty + \|\omega\|_\infty)\|f\|_2. \] (95)

Remark 5. A similar bound was derived in [22, App. B] for scattering networks, namely
\[ \|f \ast \omega(-t, 0) - F_{\tau}(f \ast \omega(-t, 0))\|_2 \leq C^2 - J + \|\tau\|_\infty\|f\|_2, \] (96)
for all $f \in L^2(R^d)$, where $\omega(-t, 0)$ is the low-pass filter of a semi-discrete directional wavelet frame for $L^2(R^d)$, 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 [63, 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
\[ (K f)(x) = \int_{R^d} k(x, u) f(u) du \] (97)
satisfying the signal-class specific identity
\[ K f = F_{\tau, \omega} f - f, \quad \forall f \in L^2(R^d), \]
and then upper-bound the deformation error $\|f - F_{\tau, \omega} f\|_2$ according to
\[ \|f - F_{\tau, \omega} f\|_2 = \|K f\|_2 \leq \|K\|_{2,2} \|f\|_2, \]
for all $f \in L^2(R^d)$. 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 $k : R^d \times R^d \rightarrow C$ be a locally integrable function satisfying
\[ (i) \sup_{x \in R^d} \int_{R^d} |k(x, u)| du \leq \alpha, \]
\[ (ii) \sup_{u \in R^d} \int_{R^d} |k(x, u)| dx \leq \alpha, \] (98)
where $\alpha > 0$. Then, $(K f)(x) = \int_{R^d} k(x, u) f(u) du$ is a bounded operator from $L^2(R^d)$ to $L^2(R^d)$ with operator norm $\|K\|_{2,2} \leq \alpha$.

We start by determining the integral operator $K$ in (97). To this end, consider $\eta \in S(R^d, C)$ such that $\hat{\eta}(\omega) = 1$, for all $\omega \in B_1(0)$. Setting $\gamma(x) := R^d \eta(Rx)$ yields $\gamma \in S(R^d, C)$ and $\hat{\gamma}(\omega) = \hat{\eta}(\omega/R)$. Thus, $\gamma(\omega) = 1$, for all $\omega \in B_H(0)$, and hence $f = f \gamma$, so that $f = f \gamma$, for all $f \in L^2(R^d)$. Next, we define the operator $A_{\gamma} : L^2(R^d) \rightarrow L^2(R^d)$, $A_{\gamma} f := f \gamma$, and note that $A_{\gamma}$ is well-defined, i.e., $A_{\gamma} f \in L^2(R^d)$, for all $f \in L^2(R^d)$, thanks to Young’s inequality [63, Theorem 1.2.12].

(91) now follows from Young’s inequality [63, Theorem 1.2.12] (since $f \in L^2(R^d)$ and $\gamma \in S(R^d, C) \subseteq L^1(R^d)$). Moreover, $A_{\gamma} f = f$, for all $f \in L^2(R^d)$. Setting $K := F_{\tau, \omega} A_{\gamma} - A_{\gamma}$, we get $K f = F_{\tau, \omega} A_{\gamma} f - A_{\gamma} f = F_{\tau, \omega} f - f$, for all $f \in L^2(R^d)$, as desired. Furthermore, it follows from
\[ (K f)(x) = \int_{R^d} k(x, u) f(u) du, \]
that the integral operator $K = F_{\tau, \omega} A_{\gamma} - A_{\gamma}$, i.e.,
\[ (K f)(x) = \int_{R^d} k(x, u) f(u) du, \]
has the kernel
\[ K(x, u) := e^{2\pi i \omega(x)} \gamma(x - \tau(x) - u) - \gamma(x - u). \] (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 $S \subseteq R^d \times R^d$, we have
\[ \int_S |k(x, u)| dx < \infty. \]
To this end, let $S \subseteq R^d \times R^d$ be a compact set. Next, choose compact sets $S_1, S_2 \subseteq R^d$ such that $S \subseteq S_1 \times S_2$. Thanks to $\gamma \in S(R^d, C)$, $\tau \in C^1(R^d, R^d)$, and $\omega \in C(R^d, R)$, all by assumption, the function $|k| : S_1 \times S_2 \rightarrow C$ is continuous as a composition of continuous functions, and therefore also Lebesgue-measurable. We further have
\[ \int_{S_1} \int_{S_2} |k(x, u)| dx du \leq \int_{S_1} \int_{R^d} |k(x, u)| dx du \]
\[ \leq \int_{S_1} \int_{R^d} \gamma(x - \tau(x) - u)| dx du + \int_{S_1} \int_{R^d} \gamma(x - u)| dx du \]
\[ \leq 2 \int_{S_1} \int_{R^d} |\gamma(y)| dy du + \int_{S_1} \int_{R^d} |\gamma(y)| dy du \]
\[ = 3\mu_L(S_1) \|\gamma\|_1 < \infty, \]
where the first term in (100) follows by the change of variables $y = x - \tau(x) - u$, together with
\[ \frac{dy}{dx} = |\det(E - (D\tau)(x))| \geq 1 - d\|D\tau\|_\infty \geq 1/2, \] (101)
for all $x \in R^d$. The arguments underlying (101) were already detailed at the end of Appendix [I]. It follows that $k$ is locally integrable owing to
\[ \int_{S_1} \int_{S_2} |k(x, u)| dx du \leq \int_{S_1 \times S_2} |k(x, u)| dx du \]
\[ = \int_{S_1} \int_{S_2} |k(x, u)| dx du < \infty, \] (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, Theorem 14.2] noting that $|k| : S_1 \times S_2 \rightarrow C$ 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 $\alpha$ of the form
\[ \alpha = C(R\|\tau\|_\infty + \|\omega\|_\infty), \quad \text{with } C > 0. \] (103)
This will be accomplished as follows: For \( x, u \in \mathbb{R}^d \), we parametrize the integral kernel according to \( h_{x,u}(t) := e^{2\pi i \omega(x)}(x-\tau(x)-u)-\gamma(x-u) \). A Taylor series expansion of \( h_{x,u}(t) \) w.r.t. the variable \( t \) now yields

\[
h_{x,u}(t) = h_{x,u}(0) + \int_0^t h'_{x,u}(\lambda) d\lambda = \int_0^t h'_{x,u}(\lambda) d\lambda, \tag{104}
\]

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

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

where

\[
h'_{x,u}(\lambda) = -e^{2\pi i \omega(x)}(\nabla_\gamma(x-\lambda \tau(x) - u), \tau(x)) + 2\pi i \omega(x)e^{2\pi i \lambda \omega(x)}(x-\lambda \tau(x) - u), \tag{106}
\]

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

\[
|h'_{x,u}(\lambda)| \leq |(\nabla_\gamma(x-\lambda \tau(x) - u), \tau(x))| + |2\pi i \omega(x)e^{2\pi i \lambda \omega(x)}(x-\lambda \tau(x) - u)| \leq |\tau(x)||\nabla_\gamma(x-\lambda \tau(x) - u)| + 2\pi |\omega(x)||\gamma(x-\lambda \tau(x) - u)|, \tag{107}
\]

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

\[
|k(x,u)| \leq \|\tau\|_\infty \int_0^1 |\nabla_\gamma(x-\lambda \tau(x) - u)| d\lambda + 2\pi |\omega\|_\infty \int_0^1 |\gamma(x-\lambda \tau(x) - u)| d\lambda. \tag{108}
\]

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

\[
\int_{\mathbb{R}^d} |k(x,u)| du \leq \|\tau\|_\infty \int_{\mathbb{R}^d} \int_0^1 |\nabla_\gamma(x-\lambda \tau(x) - u)| d\lambda du + 2\pi |\omega\|_\infty \int_{\mathbb{R}^d} \int_0^1 |\gamma(x-\lambda \tau(x) - u)| d\lambda du
\]

\[
= \|\tau\|_\infty \int_{\mathbb{R}^d} \int_0^1 |\nabla_\gamma(x-\lambda \tau(x) - u)| d\lambda du + 2\pi |\omega\|_\infty \int_{\mathbb{R}^d} \int_0^1 |\gamma(x-\lambda \tau(x) - u)| d\lambda du \tag{109}
\]

\[
= \|\tau\|_\infty \int_{\mathbb{R}^d} |\nabla_\gamma(y)| dy d\lambda + 2\pi |\omega\|_\infty \int_{\mathbb{R}^d} |\gamma(y)| dy d\lambda
\]

\[
= \|\tau\|_\infty \|\nabla_\gamma\|_1 + 2\pi |\omega\|_\infty \|\gamma\|_1. \tag{110}
\]

where (109) follows by application of the Fubini-Tonelli Theorem (84) Theorem 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

\[
\nabla_\gamma = R^{d+1} \nabla \eta(R \cdot), \|\gamma\|_1 = \|\eta\|_1, \text{ and } \|\nabla_\gamma\|_1 = R\|\nabla \eta\|_1 \text{ in (110) yields}
\]

\[
\sup_{x \in \mathbb{R}^d} \int_{\mathbb{R}^d} |k(x,u)| du \leq \|\tau\|_\infty \|\nabla_\gamma\|_1 + 2\pi |\omega\|_\infty \|\gamma\|_1 \leq \max \{\|\nabla \eta\|_1, 2\pi |\eta\|_1\} \langle R \|\tau\|_\infty + \|\omega\|_\infty \rangle, \tag{111}
\]

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

\[
\int_{\mathbb{R}^d} |k(x,u)| dx \leq \|\tau\|_\infty \int_{\mathbb{R}^d} \int_0^1 |\nabla_\gamma(x-\lambda \tau(x) - u)| d\lambda dx + 2\pi |\omega\|_\infty \int_{\mathbb{R}^d} \int_0^1 |\gamma(x-\lambda \tau(x) - u)| d\lambda dx
\]

\[
= \|\tau\|_\infty \int_{\mathbb{R}^d} \int_0^1 |\nabla_\gamma(x-\lambda \tau(x) - u)| dx d\lambda + 2\pi |\omega\|_\infty \int_{\mathbb{R}^d} \int_0^1 |\gamma(x-\lambda \tau(x) - u)| dx d\lambda \tag{112}
\]

\[
= \|\tau\|_\infty \int_{\mathbb{R}^d} |\nabla_\gamma(y)| dy d\lambda + 4\pi |\omega\|_\infty \int_{\mathbb{R}^d} |\gamma(y)| dy d\lambda \tag{113}
\]

\[
= \|\tau\|_\infty \|\nabla_\gamma\|_1 + 4\pi |\omega\|_\infty \|\gamma\|_1 \leq \max \{2\|\nabla \eta\|_1, 4\pi |\eta\|_1\} \langle R \|\tau\|_\infty + \|\omega\|_\infty \rangle. \tag{114}
\]

Here, again, (112) follows by application of the Fubini-Tonelli Theorem (84) Theorem 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

\[
C := \max \{2\|\nabla \eta\|_1, 4\pi |\eta\|_1\} \langle R \|\tau\|_\infty + \|\omega\|_\infty \rangle. \tag{115}
\]

This completes the proof. \( \square \)

ACKNOWLEDGMENTS

The authors would like to thank P. Grohs, S. Mallat, R. Alaisfari, 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. of IEEE International Symposium on Information Theory (ISIT), pp. 1212–1216, 2015.
[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, 2013.
[3] C. M. Bishop, Pattern recognition and machine learning. Springer, 2009.
[4] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern classification. John Wiley, 2nd ed., 2001.
[5] Y. LeCun and C. Cortes, “The MNIST database of handwritten digits.” http://yann.lecun.com/exdb/mnist/, 1998.
[6] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
