A Mathematical Theory of Deep Convolutional Neural Networks for Feature Extraction

Thomas Wiatowski and Helmut Bölcskei
Dept. IT & EE, ETH Zurich, Switzerland

December 22, 2015

Abstract

Deep convolutional neural networks have led to breakthrough results in practical feature extraction applications. The mathematical analysis of such networks was initiated by Mallat, 2012. Specifically, Mallat considered so-called scattering networks based on semi-discrete shift-invariant wavelet frames and modulus non-linearities in each network layer, and proved translation invariance (asymptotically in the wavelet scale parameter) and deformation stability of the corresponding feature extractor. The purpose of this paper is to develop Mallat's theory further by allowing for general convolution kernels, or in more technical parlance, general semi-discrete shift-invariant frames (including Weyl-Heisenberg, curvelet, shearlet, ridgelet, and wavelet frames) and general Lipschitz-continuous non-linearities (e.g., rectified linear units, shifted logistic sigmoids, hyperbolic tangents, and modulus functions), as well as pooling through sub-sampling, all of which can be different in different network layers. The resulting generalized network enables extraction of significantly wider classes of features than those resolved by Mallat's wavelet-modulus scattering network. We prove deformation stability for a larger class of deformations than those considered by Mallat, and we establish a new translation invariance result which is of vertical nature in the sense of the network depth determining the amount of invariance. Moreover, our results establish that deformation stability and vertical translation invariance are guaranteed by the network structure per se rather than the specific convolution kernels and non-linearities. This offers an explanation for the tremendous success of deep convolutional neural networks in a wide variety of practical feature extraction applications. The mathematical techniques we employ are based on continuous frame theory, as developed by Ali et al., 1993, and Kaiser, 1994, and allow to completely detach our proofs from the algebraic structures of the underlying frames and the particular form of the Lipschitz non-linearities.

Keywords: Deep convolutional neural networks, scattering networks, frame theory, feature extraction, signal classification.

This paper was presented in part at the 2015 IEEE International Symposium on Information Theory (ISIT) [1].

I. Introduction

A central task in signal classification is feature extraction [2]–[4]. For example, if the classification task is to decide whether an image contains a certain handwritten digit [5], the features to be extracted correspond, e.g., to the edges of the digit. The idea behind feature extraction is that feeding characteristic features of signals to be classified—rather than the signals themselves—to a trainable classifier (such as, e.g., a support vector machine (SVM) [6]) improves classification performance. Sticking to the example of handwritten digits, we would, moreover, want the feature extractor to be invariant to the digits' spatial location within the image, which motivates the use of translation-invariant feature extractors. In addition, we would also like the feature extractor to be robust with respect to (w.r.t.) handwriting styles. This can be accomplished by demanding stability w.r.t. non-linear deformations.

Spectacular success in many practical classification tasks has been reported for feature extractors generated by so-called deep convolutional neural networks [2], [7]–[11]. These networks are composed of multiple layers, each of which computes convolutional transforms, followed by non-linearities and pooling[1] operations. While deep convolutional neural networks can be used to perform classification directly [2], [7], [9]–[11], typically based on the output of the last network layer, they can also act as stand-alone feature extractors [12]–[18] with the extracted features fed into a classifier such as, e.g., a SVM. The present paper follows the latter philosophy and studies deep convolutional neural networks as stand-alone feature extractors.

Deep convolutional neural network-based feature extractors are typically distinguished according to whether the filters (i.e., the convolution kernels) employed are learned (i.e., determined from a training data set through optimization) or pre-specified (i.e., chosen a priori, possibly taking into account structural properties of the data set). While learning the filters, e.g., based on labeled data in a supervised fashion [12], [13], leads to good classification performance for large data sets, in small data sets overfitting [14] may result in performance limitations. Learning filters based on unlabeled data in an unsupervised fashion [13]–[15] can sometimes be a remedy. Pre-specified filters [13], [14], [16]–[18] (including structured filters such as wavelets[2] [13], [16]–[18], and unstructured filters such as random filters [13], [14]), on the other hand, have been found to work well on data sets of varying sizes.

The mathematical analysis of feature extractors generated by deep convolutional neural networks was initiated by Mallat in [19]. Mallat's theory applies to so-called scattering networks, where signals are

[1]In 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).

[2]Here, the structure results from the filters being obtained from a mother wavelet through scaling (and rotation) operations.

propagated through layers that compute semi-discrete wavelet transforms (i.e., convolutional transforms with pre-specified filters that are obtained from a mother wavelet through scaling (and rotation) operations), followed by modulus non-linearities. 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 [20]–[22].

Contributions and relation to Mallat's theory. The past two decades have seen extensive research [23]–[32] devoted to developing structured transforms adapted to a variety of features, most prominently, curvelet [28]–[30] and shearlet [31], [32] transforms, both of which are known to be very effective in extracting features characterized by curved edges in images. It is thus natural to ask whether Mallat's theory of scattering networks can be extended to general semi-discrete transforms (i.e., convolutional transforms with general filters that depend on some discrete indices), including curvelet and shearlet transforms. Moreover, certain image [21], [33] and audio [22] classification problems suggest that scattering networks with different semi-discrete transforms in different layers would be desirable. Furthermore, deep neural network-based feature extractors that were found to work well in practice employ a wide range of non-linearities, beyond the modulus function [13], [18], [19], namely, hyperbolic tangents [12]–[14], rectified linear units [34], [35], and logistic sigmoids [36], [37]; in addition, these non-linearities can be different in different network layers. Regarding translation invariance it was argued, e.g., in [12]–[14], [17], [18], that in practice invariance of the extracted features is crucially governed by the network depth and by pooling operations (such as, e.g., max-pooling [13], [14], [17], [18], average-pooling [12], [13], or sub-sampling [16]). In contrast, Mallat's translation invariance result [19] (in this paper referred to as *horizontal* translation invariance) is asymptotic in wavelet scales. Another aspect that was found to be desirable in practice [20], [33], but is not contained in Mallat's theory [19], is sub-sampling to reduce redundancy in the extracted features.

The goal of this paper is to develop a mathematical theory of deep convolutional neural networks for feature extraction that addresses all the points raised above and contains Mallat's wavelet-modulus scattering networks as a special case. Specifically, we extend Mallat's theory to allow for general semi-discrete transforms (including Weyl-Heisenberg (Gabor), wavelet, curvelet, shearlet, and ridgelet transforms), general Lipschitz-continuous non-linearities (e.g., rectified linear units, shifted logistic sigmoids, hyperbolic tangents, and modulus operations), and pooling through sub-sampling. Moreover, in our theory different network layers may be equipped with different semi-discrete transforms, different Lipschitz-

continuous non-linearities, and different sub-sampling factors. We prove that the resulting generalized feature extractor is translation-invariant and deformation-stable. More specifically, we obtain (i) a new translation invariance result (referred to as *vertical* translation invariance) which shows that the depth of the network determines the extent to which the feature extractor is translation-invariant, (ii) a new deformation stability bound valid for a class of non-linear deformations that is larger than that in [19], and (iii) an explicit and easy-to-verify condition on the signal transforms, the non-linearities' Lipschitz constants, and the sub-sampling factors to guarantee *vertical* translation invariance and deformation stability. Particularizing our new translation invariance result to Mallat's scattering networks, we find that asymptotics in the wavelet scale parameter, as in [19], are not needed to ensure invariance. Perhaps surprisingly, our results establish that deformation stability and vertical translation invariance are guaranteed by the network structure per se rather than the specific convolution kernels and non-linearities. This offers an explanation for the tremendous success of deep convolutional neural networks in a wide variety of practical feature extraction applications.

In terms of mathematical techniques, we note that the proofs in Mallat's theory hinge critically on the wavelet transform's structural properties such as isotropic scaling[3] and a constant number of wavelets across scales, as well as on additional technical conditions such as the vanishing moment condition on the mother wavelet. The mathematical tools employed in our theory, on the other hand, are completely detached from the algebraic structures[4] of the semi-discrete transforms, the nature of the non-linearities—as long as they are Lipschitz—and the values of the sub-sampling factors. Moreover, we show that the scattering admissibility condition [19, Theorem 2.6] is not needed for Mallat's feature extractor to be *vertically* translation-invariant and deformation-stable, where the latter is even w.r.t. the larger class of deformations considered here. The mathematical engine behind our results is the theory of continuous frames [38], [39].

Notation and preparatory material. The complex conjugate of $z \in \mathbb{C}$ is denoted by $\overline{z}$. We write $\mathrm{Re}(z)$ for the real, and $\mathrm{Im}(z)$ for the imaginary part of $z \in \mathbb{C}$. The Euclidean inner product of $x, y \in \mathbb{C}^d$ is $\langle x, y\rangle := \sum_{i=1}^{d} x_i \overline{y_i}$, with associated norm $|x| := \sqrt{\langle x, x\rangle}$. We denote the identity matrix by $E \in \mathbb{R}^{d\times d}$. For the matrix $M \in \mathbb{R}^{d\times d}$, $M_{i,j}$ designates the entry in its i-th row and j-th column, and for a tensor $T \in \mathbb{R}^{d\times d\times d}$, $T_{i,j,k}$ refers to its (i,j,k)-th component. The supremum norm of a matrix $M \in \mathbb{R}^{d\times d}$ is defined as $|M|_\infty := \sup_{i,j} |M_{i,j}|$, and the supremum norm of a tensor $T \in \mathbb{R}^{d\times d\times d}$ is

[3]Isotropic scaling of multi-dimensional signals uses the same scaling factor in all directions.

[4]Algebraic structure here refers to the structural relationship between the convolution kernels in a given semi-discrete transformation, i.e., scaling (and rotation) operations in the case of the wavelet transform as considered by Mallat in [19].

$|T|_\infty := \sup_{i,j,k} |T_{i,j,k}|$. 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 $f : \mathbb{R}^d \to \mathbb{C}$, we write $\int_{\mathbb{R}^d} f(x)\mathrm{d}x$ for the integral of f w.r.t. Lebesgue measure μ_L. For $p \in [1, \infty)$, $L^p(\mathbb{R}^d)$ stands for the space of Lebesgue-measurable functions $f : \mathbb{R}^d \to \mathbb{C}$ satisfying $\|f\|_p := (\int_{\mathbb{R}^d} |f(x)|^p \mathrm{d}x)^{1/p} < \infty$. $L^\infty(\mathbb{R}^d)$ denotes the space of Lebesgue-measurable functions $f : \mathbb{R}^d \to \mathbb{C}$ such that $\|f\|_\infty := \inf\{\alpha > 0 \mid |f(x)| \leq \alpha \text{ for a.e.}^5\ x \in \mathbb{R}^d\} < \infty$. For $f, g \in L^2(\mathbb{R}^d)$ we set $\langle f, g\rangle := \int_{\mathbb{R}^d} f(x)\overline{g(x)}\mathrm{d}x$. The tensor product of functions $f, g : \mathbb{R}^d \to \mathbb{C}$ is $(f \otimes g)(x, y) := f(x)g(y)$, $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$. Id : $L^p(\mathbb{R}^d) \to L^p(\mathbb{R}^d)$ stands for the identity operator on $L^p(\mathbb{R}^d)$. The operator norm of the bounded linear operator $A : L^p(\mathbb{R}^d) \to L^q(\mathbb{R}^d)$ is $\|A\|_{p,q} := \sup_{\|f\|_p=1} \|Af\|_q$. We denote the Fourier transform of $f \in L^1(\mathbb{R}^d)$ by $\widehat{f}(\omega) := \int_{\mathbb{R}^d} f(x)e^{-2\pi i\langle x,\omega\rangle}\mathrm{d}x$ and extend it in the usual way to $L^2(\mathbb{R}^d)$ [40, Theorem 7.9]. The convolution of $f \in L^2(\mathbb{R}^d)$ and $g \in L^1(\mathbb{R}^d)$ is $(f * g)(y) := \int_{\mathbb{R}^d} f(x)g(y - x)\mathrm{d}x$. We write $(T_t f)(x) := f(x - t)$, $t \in \mathbb{R}^d$, for the translation operator, and $(M_\omega f)(x) := e^{2\pi i\langle x,\omega\rangle} f(x)$, $\omega \in \mathbb{R}^d$, for the modulation operator. Involution is defined by $(If)(x) := \overline{f(-x)}$. For $R > 0$, the space of R-band-limited functions is denoted as $L^2_R(\mathbb{R}^d) := \{f \in L^2(\mathbb{R}^d) \mid \mathrm{supp}(\widehat{f}) \subseteq B_R(0)\}$. For a countable set $\mathcal{Q}$, $(L^2(\mathbb{R}^d))^{\mathcal{Q}}$ denotes the space of sets $s := \{s_q\}_{q\in\mathcal{Q}}$, $s_q \in L^2(\mathbb{R}^d)$, for all $q \in \mathcal{Q}$, satisfying $|||s||| := (\sum_{q\in\mathcal{Q}} \|s_q\|_2^2)^{1/2} < \infty$. A multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ is an ordered d-tupel of non-negative integers $\alpha_i \in \mathbb{N}_0$. For a multi-index $\alpha \in \mathbb{N}_0^d$, D^α denotes the differential operator $D^\alpha := (\partial/\partial x_1)^{\alpha_1} \ldots (\partial/\partial x_d)^{\alpha_d}$, with order $|\alpha| := \sum_{i=1}^d \alpha_i$. If $|\alpha| = 0$, $D^\alpha f := f$, for $f : \mathbb{R}^d \to \mathbb{C}$. The space of functions $f : \mathbb{R}^d \to \mathbb{C}$ 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 by $C^\infty(\mathbb{R}^d, \mathbb{C})$. $S(\mathbb{R}^d, \mathbb{C})$ 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 [40, 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 $f : \mathbb{R}^d \to \mathbb{C}$ as ∇f. The space of continuous mappings $v : \mathbb{R}^p \to \mathbb{R}^q$ is $C(\mathbb{R}^p, \mathbb{R}^q)$, and for $k, p, q \in \mathbb{N}$, the space of k-times continuously differentiable mappings $v : \mathbb{R}^p \to \mathbb{R}^q$ is written as $C^k(\mathbb{R}^p, \mathbb{R}^q)$. For a mapping $v : \mathbb{R}^d \to \mathbb{R}^d$ we let Dv be its Jacobian matrix, and $D^2 v$ its Jacobian tensor, with associated norms $\|v\|_\infty := \sup_{x\in\mathbb{R}^d} |v(x)|$, $\|Dv\|_\infty := \sup_{x\in\mathbb{R}^d} |(Dv)(x)|_\infty$, and $\|D^2 v\|_\infty := \sup_{x\in\mathbb{R}^d} |(D^2 v)(x)|_\infty$.

[5] Throughout "a.e." is w.r.t. Lebesgue measure.

II. MALLAT'S WAVELET-MODULUS FEATURE EXTRACTOR

We set the stage by first reviewing Mallat's feature extractor [19], the basis of which is a multi-stage architecture that involves wavelet transforms followed by modulus non-linearities. Specifically, Mallat [19, Definition 2.4] defines the extracted features $\Phi_M(f)$ of a signal $f \in L^2(\mathbb{R}^d)$ as the set of low-pass filtered functions

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

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

$$\Phi_M^n(f) := \big\{|\cdots| \, |f * \psi_{\lambda^{(1)}}| * \psi_{\lambda^{(2)}}| \cdots * \psi_{\lambda^{(n)}}| * \psi_{(-J,0)}\big\}_{\lambda^{(1)},\ldots,\lambda^{(n)} \in \Lambda_{DW} \backslash \{(-J,0)\}}, \tag{2}$$

for all $n \in \mathbb{N}$. Here, the index set $\Lambda_{DW} := \big\{(-J,0)\big\} \cup \big\{(j,k) \mid j \in \mathbb{Z} \text{ with } j > -J,\ k \in \{0,\ldots,K-1\}\big\}$ contains pairs of scales j and directions k, and

$$\psi_\lambda(x) := 2^{dj}\psi(2^j r_k^{-1} x), \qquad \lambda = (j,k) \in \Lambda_{DW} \backslash \{(-J,0)\}, \tag{3}$$

are directional wavelets [23], [41], [42] 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 $(-J,0) \in \Lambda_{DW}$ 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 (3).

The functions $\{\psi_\lambda\}_{\lambda \in \Lambda_{DW}}$ are taken to form a semi-discrete shift-invariant Parseval frame $\Psi_{\Lambda_{DW}} := \{T_b I \psi_\lambda\}_{b \in \mathbb{R}^d, \lambda \in \Lambda_{DW}}$ for $L^2(\mathbb{R}^d)$ [38], [39], [41] and hence satisfy

$$\sum_{\lambda \in \Lambda_{DW}} \int_{\mathbb{R}^d} |\langle f, T_b I \psi_\lambda \rangle|^2 \mathrm{d}b = \sum_{\lambda \in \Lambda_{DW}} \|f * \psi_\lambda\|_2^2 = \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d),$$

where $\langle f, T_b I \psi_\lambda \rangle = (f * \psi_\lambda)(b)$, $(\lambda, b) \in \Lambda_{DW} \times \mathbb{R}^d$, are the underlying frame coefficients. Note that for given $\lambda \in \Lambda_{DW}$, we actually have a continuum of frame coefficients as the translation parameter $b \in \mathbb{R}^d$. In Appendix A, we give a brief review of the general theory of semi-discrete shift-invariant frames, and in Appendices B and C we collect structured example frames in 1-D and 2-D, respectively.

The architecture corresponding to the feature extractor Φ_M in (1), illustrated in Figure 1, is known as *scattering network* [20], and employs the frame $\Psi_{\Lambda_{DW}}$ and the modulus non-linearity $|\cdot|$ in every network layer. For given $n \in \mathbb{N}$, the set $\Phi_M^n(f)$ in (2) corresponds to the features of the function f generated in the n-th network layer, see Figure 1.

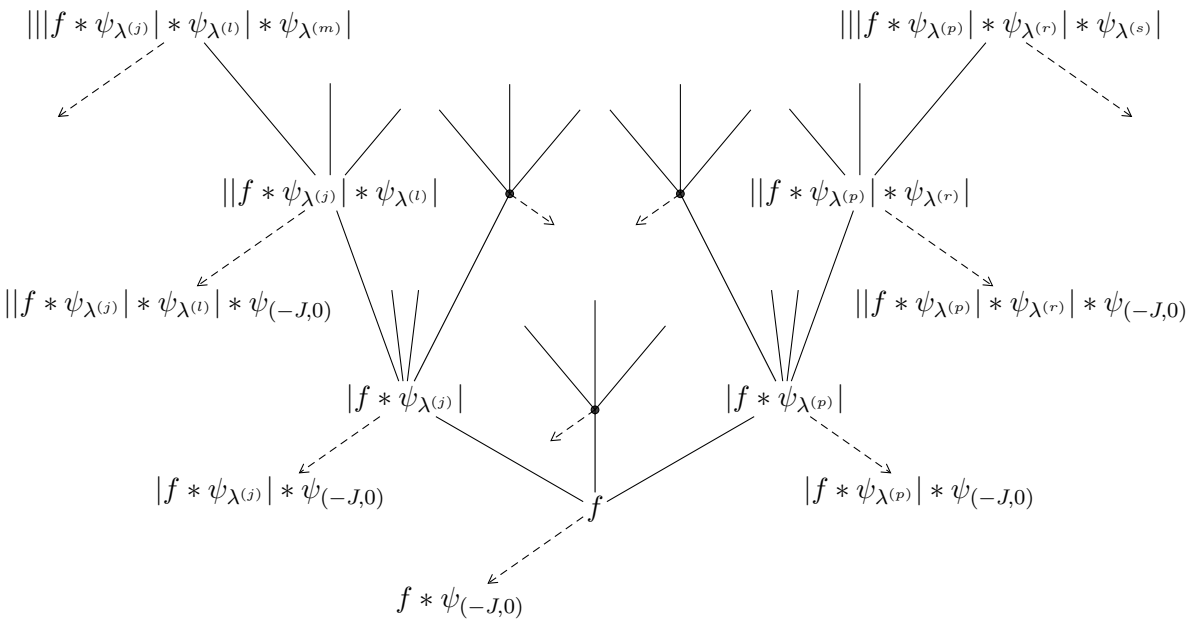

Fig. 1: Mallat's scattering network architecture based on wavelet filtering and modulus non-linearities. The features $\Phi_M(f)$ in (1), here indicated at the tips of the arrows, are generated from outputs in all layers of the network.

Remark 1. *The function $|f * \psi_\lambda|$, $\lambda \in \Lambda_{DW}\backslash\{(-J,0)\}$, can be thought of as indicating the locations of singularities of $f \in L^2(\mathbb{R}^d)$. Specifically, with the relation of $|f * \psi_\lambda|$ to the Canny edge detector [43] as described in [44], in dimension $d = 2$, we can think of $|f * \psi_\lambda| = |f * \psi_{(j,k)}|$, $\lambda = (j,k) \in \Lambda_{DW}\backslash\{(-J,0)\}$, as an image at scale j specifying the locations of edges of the image f that are oriented in direction k. Furthermore, it was argued in [20], [22], [33] that the features $\Phi_M^1(f)$ generated in the first layer of the scattering network are very similar, in dimension $d = 1$, to mel frequency cepstral coefficients [45], and in dimension $d = 2$ to SIFT-descriptors [46], [47].*

It is shown in [19, Theorem 10] that the feature extractor Φ_M is translation-invariant in the sense of

$$\lim_{J\to\infty} |||\Phi_M(T_t f) - \Phi_M(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d),\, \forall t \in \mathbb{R}^d. \tag{4}$$

Note that this invariance result is asymptotic in the scale parameter $J \in \mathbb{Z}$, an aspect discussed in more detail in Section V. Furthermore, Mallat proved in [19, Theorem 2.12] that Φ_M is stable w.r.t. deformations of the form

$$(F_\tau f)(x) := f(x - \tau(x)).$$

More formally, for the normed function space $(H_M, \|\cdot\|_{H_M})$ defined in (23) below, Mallat established

that there exists a constant $C > 0$ such that for all $f \in H_M$, and all $\tau \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ with[6] $\|D\tau\|_\infty \leq \frac{1}{2d}$, the deformation error satisfies

$$|||\Phi_M(F_\tau f) - \Phi_M(f)||| \leq C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_{H_M}. \tag{5}$$

The following technical condition on the mother wavelet ψ, referred to as the scattering admissibility condition in [19, Theorem 2.6], is of crucial importance in Mallat's proofs of translation invariance (4) and deformation stability (5): The mother wavelet ψ is said to be scattering-admissible if there exists a function $\rho : \mathbb{R}^d \to \mathbb{R}^+$ with $|\widehat{\rho}(2^J\omega)| \leq |\widehat{\psi_{(-J,0)}}(2\omega)|$, $\omega \in \mathbb{R}^d$, $\widehat{\rho}(0) = 1$, and a $\nu \in \mathbb{R}^d$, such that

$$\inf_{1\leq\omega\leq 2} \sum_{j=-\infty}^{\infty} \sum_{k=0}^{K-1} |\widehat{\psi}(2^{-j}r_k^{-1}\omega)|^2 \Delta(2^{-j}r_k^{-1}\omega) > 0, \tag{6}$$

where

$$\Delta(\omega) := |\widehat{\rho}(\omega - \nu)|^2 - \sum_{k=1}^{\infty} k\big(1 - |\widehat{\rho}(2^{-k}(\omega - \nu))|^2\big).$$

We refer the reader to Section V for an in-depth discussion of Mallat's scattering admissibility condition. Here, we conclude by noting that, to the best of our knowledge, no mother wavelet $\psi \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$, for $d \geq 2$, satisfying the scattering admissibility condition has been reported in the literature.

In practice signal classification based on Mallat's feature extractor is performed as follows. First, the function f and the wavelet frame atoms $\{\psi_\lambda\}_{\lambda\in\Lambda_{DW}}$ are discretized to finite-dimensional vectors. The resulting scattering network then computes the finite-dimensional feature vector $\Phi_M(f)$, whose dimension is typically reduced through an orthogonal least squares step [48], and feeds the result into a supervised classifier such as, e.g., a SVM. State-of-the-art results were reported for various classification tasks such as handwritten digit recognition [20], texture discrimination [20], [21], and musical genre classification [22].

III. Generalized feature extractor

As already mentioned, scattering networks follow the architecture of deep convolutional neural networks [2], [7]–[18] in the sense of cascading convolutions (with atoms $\{\psi_\lambda\}_{\lambda\in\Lambda_{DW}}$ of the wavelet frame $\Psi_{\Lambda_{DW}}$) and non-linearities, namely, the modulus function. On the other hand, general deep convolutional neural networks as studied in the literature exhibit a number of additional features:

[6]It is actually the assumption $\|D\tau\|_\infty \leq \frac{1}{2d}$, rather than $\|D\tau\|_\infty \leq \frac{1}{2}$ as stated in [19, Theorem 2.12], that is needed in [19, p. 1390] to establish that $|\det(E - (D\tau)(x))| \geq 1 - d\|D\tau\|_\infty \geq 1/2$.

- a wide variety of filters are employed, namely pre-specified unstructured filters such as random filters [13], [14], and filters that are learned in a supervised [12], [13] or an unsupervised [13]–[15] fashion.
- a wide variety of non-linearities are employed such as, e.g., hyperbolic tangents [12]–[14], rectified linear units [34], [35], and logistic sigmoids [36], [37].
- convolution and the application of a non-linearity is typically followed by a pooling operation such as, e.g., max-pooling [13], [14], [17], [18], average-pooling [12], [13], or sub-sampling [16].
- the filters, non-linearities, and pooling operations are allowed to be different in different network layers.

The purpose of this paper is to develop a mathematical theory of deep convolutional neural networks for feature extraction that encompasses all of the aspects above, apart from max-pooling and average-pooling. Formally, we generalize Mallat's feature extractor Φ_M as follows. In the n-th network layer, we replace the wavelet-modulus convolution 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 shift-invariant frame $\Psi_n := \{T_b I g_{\lambda_n}\}_{b \in \mathbb{R}^d, \lambda_n \in \Lambda_n}$ for $L^2(\mathbb{R}^d)$ with countable index set Λ_n (see Appendix A for a brief overview of the theory of semi-discrete shift-invariant 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 $f, h \in L^2(\mathbb{R}^d)$, with $M_n f = 0$ for $f = 0$. The output of this non-linearity, $M_n(f * g_{\lambda_n})$, is then sub-sampled by a factor of $R_n \geq 1$ according to

$$(M_n(f * g_{\lambda_n}))(R_n \cdot).$$

The operation $f \mapsto f(R_n \cdot)$, for $R_n \geq 1$, emulates sub-sampling or decimation as used in multi-rate signal processing [49] where sub-sampling by a factor of R amounts to retaining only every R-th sample.

As the atoms g_{λ_n} are arbitrary in our generalization, they can, of course, also be taken to be structured, e.g., Weyl-Heisenberg functions, curvelets, shearlets, ridgelets, or wavelets as considered by Mallat in [19] (where the atoms g_{λ_n} are obtained from a mother wavelet through scaling (and rotation) operations, see Section II). These signal transforms have been employed successfully in various feature extraction tasks [50]–[58], see Appendices B and C, but their use—apart from wavelets—in deep convolutional neural networks appears to be new. Furthermore, our generalization comprises Mallat-type feature extractors based on general (i.e., not necessarily tight) wavelet frames [20]–[22], [33], and allows for different mother wavelets in different layers [22].

We refer the reader to Appendix D for a detailed discussion of several relevant example non-linearities (e.g., rectified linear units, shifted logistic sigmoids, hyperbolic tangents, and, of course, the modulus) that fit into our framework. Another novel aspect of our theory is a translation invariance result that formalizes the idea of the features becoming more translation-invariant with increasing network depth (see, e.g., [12]–[14], [17], [18]). This notion of translation invariance is in stark contrast to that used by Mallat (4), which is asymptotic in the scale parameter J, and does not depend on the network depth. We honor this difference by referring to Mallat's result as *horizontal* translation invariance and to ours as *vertical* translation invariance.

Finally, on a methodological level, we systematically introduce frame theory and the theory of Lipschitz-continuous operators into the field of deep learning. Specifically, the conditions on the atoms g_{λ_n} for the network to be deformation-stable and vertically translation-invariant are so mild as to easily be satisfied by *learned* filters. In essence, this shows that deformation stability and vertical translation invariance are induced by the network structure per se rather than the filter characteristics and the specific nature of the non-linearities. We feel that this insight offers an explanation for the impressive performance of deep convolutional neural networks in a wide variety of practical classification tasks.

Although it may seem that our generalizations require more sophisticated mathematical techniques than those employed in [19], it actually turns out that our approach leads to significantly simpler and, in particular, shorter proofs. We hasten to add, however, that the notion of translation invariance we consider, namely vertical translation invariance, is fundamentally different from horizontal translation invariance as used by Mallat. Specifically, by letting $J \to \infty$ Mallat guarantees translation invariance in every network layer, whereas vertical translation invariance only builds up with increasing network depth.

We next state definitions and collect preliminary results needed for the mathematical analysis of our generalized feature extraction network. The basic building blocks of the network we consider are the triplets (Ψ_n, M_n, R_n) associated with individual network layers and referred to as *modules*.

Definition 1. *For $n \in \mathbb{N}$, let $\Psi_n = \{T_b I g_{\lambda_n}\}_{b \in \mathbb{R}^d, \lambda_n \in \Lambda_n}$ be a semi-discrete shift-invariant frame for $L^2(\mathbb{R}^d)$, let $M_n : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ be a Lipschitz-continuous operator with $M_n f = 0$ for $f = 0$, and let $R_n \geq 1$ be a sub-sampling factor. Then, the sequence of triplets*

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

is referred to as a module-sequence.

The following definition introduces the concept of paths on index sets, which will prove helpful in characterizing the generalized feature extraction network. The idea for this formalism is due to Mallat [19, Definition 2.2].

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

$$U_n(\lambda_n, f) := U_n[\lambda_n]f := (M_n(f * g_{\lambda_n}))(R_n\cdot). \tag{7}$$

For $1 \leq n < \infty$, define the set $\Lambda_1^n := \Lambda_1 \times \Lambda_2 \times \cdots \times \Lambda_n$. An ordered sequence $q = (\lambda_1, \lambda_2, \ldots, \lambda_n) \in \Lambda_1^n$ is called a path. For the empty path $e := \emptyset$ we set $\Lambda_1^0 := \{e\}$ and $U_0[e]f := f$, for all $f \in L^2(\mathbb{R}^d)$.

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

$$\begin{aligned}\|U_n[\lambda_n]f\|_2^2 &= \int_{\mathbb{R}^d} |(M_n(f * g_{\lambda_n}))(R_n x)|^2 \mathrm{d}x = R_n^{-d} \int_{\mathbb{R}^d} |(M_n(f * g_{\lambda_n}))(y)|^2 \mathrm{d}y \\ &= R_n^{-d} \|M_n(f * g_{\lambda_n})\|_2^2 \leq R_n^{-d} L_n^2 \|f * g_{\lambda_n}\|_2^2 \leq B_n R_n^{-d} L_n^2 \|f\|_2^2. \end{aligned} \tag{8}$$

Here, we used the Lipschitz continuity of M_n according to $\|M_n f - M_n h\|_2^2 \leq L_n^2 \|f - h\|_2^2$, together with $M_n h = 0$ for $h = 0$ to get $\|M_n f\|_2^2 \leq L_n^2 \|f\|_2^2$. The last step in (8) is thanks to

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

which follows from the frame condition (24) on Ψ_n. We will also need the extension of the operator U_n to paths $q \in \Lambda_1^n$ according to

$$U[q]f = U[(\lambda_1, \lambda_2, \ldots, \lambda_n)]f := U_n[\lambda_n] \cdots U_2[\lambda_2] U_1[\lambda_1] f, \tag{9}$$

with $U[e]f := U_0[e]f = f$. Note that the multi-stage operation (9) is again well-defined as

$$\|U[q]f\|_2^2 \leq \left(\prod_{k=1}^{n} B_k R_k^{-d} L_k^2 \right) \|f\|_2^2, \qquad \forall q \in \Lambda_1^n, \forall f \in L^2(\mathbb{R}^d), \tag{10}$$

which follows by repeated application of (8).

In Mallat's construction one atom ψ_λ, $\lambda \in \Lambda_{DW}$, in the frame $\Psi_{\Lambda_{DW}}$, namely the low-pass filter $\psi_{(-J,0)}$, is singled out to generate the extracted features according to (2), see also Figure 1. We follow Mallat's construction and designate one of the atoms in each frame in the module-sequence

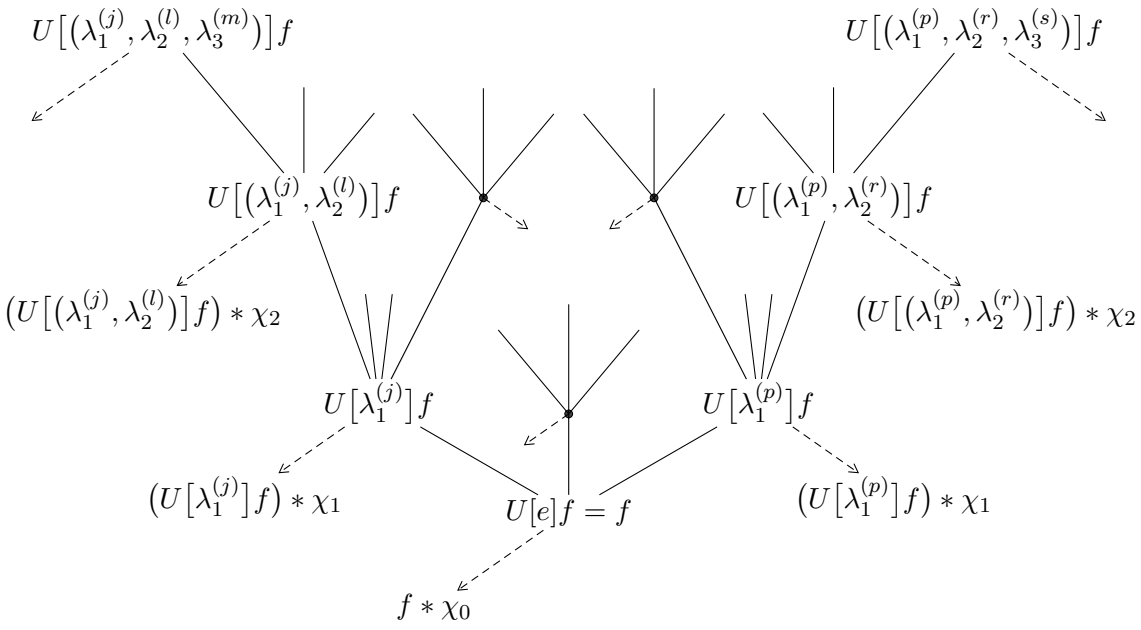

Fig. 2: Network architecture underlying the generalized feature extractor (11). The index $\lambda_n^{(k)}$ corresponds to the k-th atom $g_{\lambda_n^{(k)}}$ of the frame Ψ_n associated with the n-th network layer. The function χ_n is the output-generating atom of the n-th layer, where $n=0$ corresponds to the root of the network.

$\Omega=\big((\Psi_n,M_n,R_n)\big)_{n\in\mathbb{N}}$ as the output-generating atom $\chi_{n-1}:=g_{\lambda_n^*}$, $\lambda_n^*\in\Lambda_n$, of the $(n-1)$-th layer. The atoms $\{g_{\lambda_n}\}_{\lambda_n\in\Lambda_n\backslash\{\lambda_n^*\}}\cup\{\chi_{n-1}\}$ in Ψ_n are thus used across two consecutive layers in the sense of χ_{n-1} generating the output in the $(n-1)$-th layer, and the $\{g_{\lambda_n}\}_{\lambda_n\in\Lambda_n\backslash\{\lambda_n^*\}}$ propagating signals to the n-th layer according to (7), see Figure 2. Note, however, that our theory does not require the output-generating atoms to be low-pass filters[7] (as is the case for Mallat's feature extractor (1)), rather a very mild decay condition is needed only, see Theorem 2. From now on, with slight abuse of notation, we shall write Λ_n for $\Lambda_n\backslash\{\lambda_n^*\}$ as well.

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

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

$$\Phi_\Omega(f):=\bigcup_{n=0}^{\infty}\{(U[q]f)*\chi_n\}_{q\in\Lambda_1^n}. \tag{11}$$

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

For $q \in \Lambda_1^n$, the feature $(U[q]f) * \chi_n$ is generated in the n-th layer of the network. The collection of features generated in the n-th network layer is denoted by Φ_Ω^n, i.e.,

$$\Phi_\Omega^n(f) := \{(U[q]f) * \chi_n\}_{q \in \Lambda_1^n},$$

and the overall features are given by

$$\Phi_\Omega(f) = \bigcup_{n=0}^{\infty} \Phi_\Omega^n(f).$$

The feature extractor $\Phi_\Omega : L^2(\mathbb{R}^d) \to (L^2(\mathbb{R}^d))^{\mathcal{Q}}$, where $\mathcal{Q} := \bigcup_{n=0}^{\infty} \Lambda_1^n$, is well-defined, i.e., $\Phi_\Omega(f) \in (L^2(\mathbb{R}^d))^{\mathcal{Q}}$, for all $f \in L^2(\mathbb{R}^d)$, under a technical condition on the module-sequence Ω formalized as follows.

Proposition 1. *Let $\Omega = \big((\Psi_n, M_n, R_n)\big)_{n\in\mathbb{N}}$ be a module-sequence, denote the frame upper bounds of Ψ_n by $B_n > 0$, the Lipschitz constants of the operators M_n by $L_n > 0$, and the sub-sampling factors by $R_n \geq 1$. If*

$$\max\{B_n, B_n R_n^{-d} L_n^2\} \leq 1, \quad \forall n \in \mathbb{N}, \tag{12}$$

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

Proof. The proof is given in Appendix E. □

As condition (12) is of central importance, we formalize it as follows.

Definition 4. *Let $\Omega = \big((\Psi_n, M_n, R_n)\big)_{n\in\mathbb{N}}$ be a module-sequence with frame upper bounds $B_n > 0$, Lipschitz constants $L_n > 0$, and sub-sampling factors $R_n \geq 1$. The condition*

$$\max\{B_n, B_n R_n^{-d} L_n^2\} \leq 1, \quad \forall n \in \mathbb{N}, \tag{13}$$

is referred to as weak admissibility condition. Module-sequences that satisfy (13) *are called weakly admissible.*

We chose the qualifier *weak* in Definition 4 to indicate that the admissibility condition (13) is easily met in practice. To see this, first note that L_n is set through the non-linearity M_n (e.g., the modulus non-linearity $M_n = |\cdot|$ has $L_n = 1$, for all $n \in \mathbb{N}$, see Appendix D), and B_n is determined through the frame Ψ_n (e.g., the directional wavelet frame introduced in Section II has $B_n = 1$, for all $n \in \mathbb{N}$).

Depending on the desired amount of translation invariance of the features Φ_Ω^n generated in the n-th network layer (see Section IV-B for details), we fix the sub-sampling factor $R_n \geq 1$ (e.g., $R_n = 2$, for all $n \in \mathbb{N}$). Obviously, condition (13) is met if

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

which can be satisfied by simply normalizing the frame elements of Ψ_n accordingly. We refer to Proposition 3 in Appendix A for corresponding normalization techniques, which, as explained in Section IV, do not affect our deformation stability and translation invariance results.

IV. Properties of the generalized feature extractor

A. Deformation stability

The following theorem states that the generalized feature extractor Φ_Ω defined in (11) is stable w.r.t. time-frequency deformations of the form

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

This class of deformations is wider than that considered in Mallat's theory, which deals with translation-like deformations of the form $f(x - \tau(x))$ only. Modulation-like deformations $e^{2\pi i \omega(x)} f(x)$ occur, e.g., if the signal is subject to an unwanted modulation, and we therefore have access to a bandpass version of $f \in L^2(\mathbb{R}^d)$ only.

Theorem 1. *Let $\Omega = \big((\Psi_n, M_n, R_n)\big)_{n\in\mathbb{N}}$ be a weakly admissible module-sequence. The corresponding feature extractor Φ_Ω is stable on the space of R-band-limited functions $L_R^2(\mathbb{R}^d)$ w.r.t. deformations $(F_{\tau,\omega} f)(x) = e^{2\pi i \omega(x)} f(x - \tau(x))$, i.e., there exists a universal constant $C > 0$ (that does not depend on Ω) such that for all $f \in L_R^2(\mathbb{R}^d)$, all $\omega \in C(\mathbb{R}^d, \mathbb{R})$, and all $\tau \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ with $\|D\tau\|_\infty \leq \frac{1}{2d}$, it holds that*

$$|||\Phi_\Omega(F_{\tau,\omega} f) - \Phi_\Omega(f)||| \leq C\big(R\|\tau\|_\infty + \|\omega\|_\infty\big)\|f\|_2. \tag{14}$$

Proof. The proof is given in Appendix F. □

Theorem 1 shows that deformation stability in the sense of (5) is retained for the generalized feature extractor Φ_Ω. Similarly to Mallat's deformation stability bound (5), the bound in (14) holds for deformations τ with sufficiently "small" Jacobian matrix, i.e., as long as $\|D\tau\|_\infty \leq \frac{1}{2d}$. Note, however, that (5) depends on the scale parameter J. This is problematic as Mallat's horizontal translation invariance

result (4) requires $J \to \infty$, and the upper bound in (5) goes to infinity for $J \to \infty$ as a consequence of $J\|D\tau\|_\infty \to \infty$. The deformation stability bound (14), in contrast, is completely decoupled from the vertical translation invariance result stated in Theorem 2 in Section IV-B.

The strength of the deformation stability result in Theorem 1 derives itself from the fact that the only condition on the underlying module-sequence Ω for (14) to hold is the weak admissibility condition (13), which as outlined in Section III, can easily be met by normalizing the frame elements of Ψ_n, for all $n \in \mathbb{N}$, appropriately. This normalization does not have an impact on the constant C in (14). More specifically, C is shown in (86) to be completely independent of Ω. All this is thanks to the technique we use for proving Theorem 1 being completely independent of the algebraic structures of the frames Ψ_n, of the particular form of the operators M_n, and of the specific sub-sampling factors R_n. This is accomplished through a generalization[8] of [19, Proposition 2.5] stated in Proposition 4 in Appendix F, and the upper bound on $\|F_{\tau,\omega} f - f\|_2$ for R-band-limited functions detailed in Proposition 5 in Appendix F.

B. Vertical translation invariance

The next result states that under very mild decay conditions on the Fourier transforms $\widehat{\chi_n}$ of the output-generating atoms χ_n, the network 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 [12]–[14], [17], [18], where it is informally argued that the network's outputs generated at deeper layers tend to be more translation-invariant. Before presenting formal statements, we note that the vertical nature of our translation invariance result is in stark contrast to the horizontal nature of Mallat's result (4), where translation invariance is achieved asymptotically in the scale parameter J. We hasten to add, that $J \to \infty$ in Mallat's scattering network yields, however, translation invariance for the features in *each* network layer.

Theorem 2. *Let $\Omega = \big((\Psi_n, M_n, R_n)\big)_{n\in\mathbb{N}}$ be a weakly admissible module-sequence, and assume that the operators $M_n : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ commute with the translation operator T_t, i.e.,*

$$M_n T_t f = T_t M_n f, \quad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}^d, \forall n \in \mathbb{N}. \tag{15}$$

[8]This generalization is in the sense of allowing for general semi-discrete shift-invariant frames, general Lipschitz-continuous operators, and sub-sampling.

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

$$\Phi_\Omega^n(T_t f) = T_{\frac{t}{R_1 R_2 \dots R_n}} \Phi_\Omega^n(f), \quad \forall f \in L^2(\mathbb{R}^d),\, \forall t \in \mathbb{R}^d,\, \forall n \in \mathbb{N}, \tag{16}$$

where $T_t \Phi_\Omega^n(f)$ refers to element-wise application of T_t, i.e., $T_t \Phi_\Omega^n(f) := \{T_t h \,|\, \forall h \in \Phi_\Omega^n(f)\}$.

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

$$|\widehat{\chi_n}(\omega)||\omega| \leq K, \qquad \textit{a.e. } \omega \in \mathbb{R}^d,\ \forall n \in \mathbb{N}_0, \tag{17}$$

then

$$|||\Phi_\Omega^n(T_t f) - \Phi_\Omega^n(f)||| \leq \frac{2\pi|t|K}{R_1 \dots R_n} \|f\|_2, \quad \forall f \in L^2(\mathbb{R}^d),\, \forall t \in \mathbb{R}^d. \tag{18}$$

Proof. The proof is given in Appendix I. □

We first note that all pointwise (i.e., memoryless) non-linearities $M_n : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ satisfy the commutation condition (15). 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 as employed by Mallat in [19], are, indeed, pointwise and hence covered by Theorem 2. We refer the reader to Appendix D for a brief review of corresponding example non-linearities. Moreover, note that (17) can easily be met by taking the output-generating atoms $\{\chi_n\}_{n \in \mathbb{N}_0}$ either to satisfy

$$\sup_{n \in \mathbb{N}_0} \{\|\chi_n\|_1 + \|\nabla \chi_n\|_1\} < \infty, \tag{19}$$

see, e.g., [40, Ch. 7], or to be uniformly band-limited in the sense of $\operatorname{supp}(\widehat{\chi_n}) \subseteq B_R(0)$, for all $n \in \mathbb{N}_0$, with an R independent of n (see, e.g., [41, Ch. 2.3]). The inequality (18) shows that we can explicitly control the amount of translation invariance via the sub-sampling factors R_n. Furthermore, the condition $\lim_{n \to \infty} R_1 \cdot R_2 \cdot \ldots \cdot R_n = \infty$ (easily met by taking $R_n > 1$, for all $n \in \mathbb{N}$) yields, thanks to (18), asymptotically exact translation invariance according to

$$\lim_{n \to \infty} |||\Phi_\Omega^n(T_t f) - \Phi_\Omega^n(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d),\, \forall t \in \mathbb{R}^d. \tag{20}$$

Finally, we note that in practice, translation *covariance* in the sense of $\Phi_\Omega^n(T_t f) = T_t \Phi_\Omega^n(f)$, for all $f \in L^2(\mathbb{R}^d)$, and all $t \in \mathbb{R}^d$, may also be desirable, e.g., in face pose estimation where translations of a given image correspond to different poses which the feature extractor Φ_Ω should reflect.

Corollary 1. *Let $\Omega = \big((\Psi_n, M_n, R_n)\big)_{n\in\mathbb{N}}$ be a weakly admissible module-sequence, and assume that the operators $M_n : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ commute with the translation operator T_t in the sense of* (15)*. If, in addition, there exists a constant $K > 0$ (that does not depend on n) such that the Fourier transforms $\widehat{\chi_n}$ of the output-generating atoms χ_n satisfy the decay condition* (17)*, then*

$$|||\Phi_\Omega^n(T_t f) - T_t\Phi_\Omega^n(f)||| \leq 2\pi|t|K\big|1/(R_1 \dots R_n) - 1\big|\|f\|_2, \quad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}^d.$$

Proof. The proof is given in Appendix J. □

Theorem 2 and Corollary 1 nicely show that having $1/(R_1 \dots R_n)$ large yields more translation invariance but less translation covariance and vice versa.

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

$$\sum_{\lambda_n\in\Lambda_n} \int_{\mathbb{R}^d} |\langle f, T_b I g_{\lambda_n}\rangle|^2 \mathrm{d}b = \sum_{\lambda_n\in\Lambda_n} \|f * g_{\lambda_n}\|_2^2 \leq B_n\|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d),$$

for some $B_n > 0$, which is equivalent to

$$\sum_{\lambda_n\in\Lambda_n} |\widehat{g_{\lambda_n}}(\omega)|^2 \leq B_n, \quad a.e.\ \omega \in \mathbb{R}^d, \tag{21}$$

see Proposition 2. Pre-specified unstructured filters [13], [14] and learned filters [12]–[15] are therefore covered by our theory as long as (21) *is satisfied. We emphasize that* (21) *is a simple boundedness condition in the frequency domain. In classical frame theory $A_n > 0$ guarantees completeness of the set $\Psi_n = \{T_b I g_{\lambda_n}\}_{b\in\mathbb{R}^d, \lambda_n\in\Lambda_n}$ for the signal space under consideration, here $L^2(\mathbb{R}^d)$. The absence of a frame lower bound $A_n > 0$ therefore translates into a lack of completeness of Ψ_n, which may result in $\Phi_\Omega(f)$ not containing all essential features of the signal f. This will, in general, have a (possibly significant) impact on classification performance in practice, which is why ensuring the entire frame property* (24) *is prudent.*

V. Relation to Mallat's results

To see how Mallat's wavelet-modulus feature extractor Φ_M defined in (1) is covered by our generalized framework, simply note that Φ_M is a feature extractor Φ_Ω based on the module-sequence

$$\Omega_M = \big((\Psi_{\Lambda_{DW}}, |\cdot|, 1)\big)_{n\in\mathbb{N}}, \tag{22}$$

where each layer is associated with the same module $(\Psi_{\Lambda_{DW}}, |\cdot|, 1)$ and thus with the same semi-discrete shift-invariant directional wavelet frame $\Psi_{\Lambda_{DW}} = \{T_b I \psi_\lambda\}_{b\in\mathbb{R}^d, \lambda\in\Lambda_{DW}}$ and the modulus non-linearity $|\cdot|$. Since Φ_M does not involve sub-sampling, we have $R_n = 1$, for all $n \in \mathbb{N}$, and the output-generating atom for all layers is taken to be the low-pass filter $\psi_{(-J,0)}$, i.e., $\chi_n = \psi_{(-J,0)}$, for all $n \in \mathbb{N}_0$. Owing to [19, Eq. 2.7], the set $\{\psi_\lambda\}_{\lambda\in\Lambda_{DW}}$ satisfies the equivalent frame condition (26) with $A = B = 1$, and $\Psi_{\Lambda_{DW}}$ therefore forms a semi-discrete shift-invariant Parseval frame for $L^2(\mathbb{R}^d)$, which implies $A_n = B_n = 1$, for all $n \in \mathbb{N}$. The modulus non-linearity $M_n = |\cdot|$ is Lipschitz-continuous with Lipschitz constant $L_n = 1$, satisfies $M_n f = |f| = 0$ for $f = 0$, and, as a pointwise (memoryless) operator, trivially commutes with the translation operator T_t in the sense of (15), see Appendix D for the corresponding formal arguments. The weak admissibility condition (13) is met according to

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

so that all the conditions required by Theorems 1 and 2 and Corollary 1 are satisfied.

Translation invariance. Mallat's horizontal translation invariance result (4),

$$\lim_{J\to\infty} |||\Phi_M(T_t f) - \Phi_M(f)||| = \lim_{J\to\infty} \Big(\sum_{n=0}^{\infty} |||\Phi_M^n(T_t f) - \Phi_M^n(f)|||^2\Big)^{1/2} = 0,$$

is asymptotic in the wavelet scale parameter J, and guarantees translation invariance in every network layer in the sense of

$$\lim_{J\to\infty} |||\Phi_M^n(T_t f) - \Phi_M^n(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d),\, \forall t \in \mathbb{R}^d,\, \forall n \in \mathbb{N}_0.$$

In contrast, our vertical translation invariance result (20) is asymptotic in the network depth n and is in line with observations made in the deep learning literature, e.g., in [12]–[14], [17], [18], where it is found that the network's outputs generated at deeper layers tend to be more translation-invariant.

We can easily render Mallat's feature extractor Φ_M vertically translation-invariant by substituting the

module sequence (22) by

$$\widetilde{\Omega}_M := \big((\Psi_{\Lambda_{DW}}, |\cdot|, R_n)\big)_{n\in\mathbb{N}},$$

and choosing the sub-sampling factors such that $\lim_{n\to\infty} R_1 \cdot \ldots \cdot R_n = \infty$. First, the weak admissibility condition (13) is met on account of

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

where we used the Lipschitz continuity of $M_n = |\cdot|$ with $L_n = 1$. Furthermore, $M_n = |\cdot|$ satisfies the commutation property (15), as explained above, and, by $|\psi_{(-J,0)}(x)| \leq C_1(1+|x|)^{-d-2}$ and $|\nabla\psi_{(-J,0)}(x)| \leq C_2(1+|x|)^{-d-2}$ for some $C_1, C_2 > 0$, see [19, p. 1336], it follows that $\|\psi_{(-J,0)}\|_1 < \infty$ and $\|\nabla\psi_{(-J,0)}\|_1 < \infty$ [59, Ch. 2.2.], and thus $\|\psi_{(-J,0)}\|_1 + \|\nabla\psi_{(-J,0)}\|_1 < \infty$. By (19) the output-generating atoms $\chi_n = \psi_{(-J,0)}$, $n \in \mathbb{N}_0$, therefore satisfy the decay condition (17).

Deformation stability. Mallat's deformation stability bound (5) applies to translation-like deformations of the form $f(x-\tau(x))$, while our corresponding bound (14) pertains to the larger class of time-frequency deformations of the form $e^{2\pi i\omega(x)} f(x-\tau(x))$.

Furthermore, Mallat's deformation stability bound (5) depends on the scale parameter J. This is problematic as Mallat's horizontal translation invariance result (4) requires $J \to \infty$, which, by $J\|D\tau\|_\infty \to \infty$ for $J \to \infty$, renders the deformation stability upper bound (5) void as it goes to ∞. In contrast, in our framework, the deformation stability bound and the conditions for vertical translation invariance are completely decoupled.

Finally, Mallat's deformation stability bound (5) applies to the space

$$H_M := \Big\{ f \in L^2(\mathbb{R}^d) \;\Big|\; \|f\|_{H_M} := \sum_{n=0}^{\infty} \Big(\sum_{q\in(\Lambda_{DW})_1^n} \|U[q]f\|_2^2 \Big)^{1/2} < \infty \Big\}, \tag{23}$$

where $(\Lambda_{DW})_1^n$ denotes the set of paths $q = (\lambda_1, \ldots, \lambda_n)$ of length n with $\lambda_k \in \Lambda_{DW}$, $k = 1, \ldots, n$ (see Definition 2). While [19, p. 1350] cites numerical evidence on the series $\sum_{q\in(\Lambda_{DW})_1^n} \|U[q]f\|_2^2$ being finite (for some $n \in \mathbb{N}$) for a large class of signals $f \in L^2(\mathbb{R}^d)$, it seems difficult to establish this analytically, let alone to show that $\sum_{n=0}^{\infty} \Big(\sum_{q\in(\Lambda_{DW})_1^n} \|U[q]f\|_2^2 \Big)^{1/2}$ is finite. In contrast, our deformation stability bound (14) applies *provably* to the space of R-band-limited functions $L_R^2(\mathbb{R}^d)$. Finally, the space H_M in (23) depends on the wavelet frame atoms $\{\psi_\lambda\}_{\lambda\in\Lambda_{DW}}$, and thereby on the underlying signal transform, whereas $L_R^2(\mathbb{R}^d)$ is, of course, completely independent of the module-sequence Ω.

Proof techniques. The techniques used in [19] to prove the deformation stability bound (5) and the horizontal translation invariance result (4) make heavy use of structural specifics of the wavelet transform, namely, isotropic scaling (see, e.g., [19, Appendix A]), a constant number $K \in \mathbb{N}$ of directional wavelets across scales (see, e.g., [19, Eq. E.1]), and several technical conditions such as a vanishing moment condition on the mother wavelet ψ (see, e.g., [19, p. 1391]). In addition, Mallat imposes the scattering admissibility condition (6). First of all, this condition depends on the underlying signal transform, more precisely on the mother wavelet ψ, whereas our weak admissibility condition (13) is in terms of the frame upper bounds B_n, the Lipschitz constants L_n, and the sub-sampling factors R_n. As the frame upper bounds B_n can be adjusted by simply normalizing the frame elements, and this normalization affects neither vertical translation invariance nor deformation stability, we can argue that our weak admissibility condition is independent of the signal transforms underlying the network. Second, Mallat's scattering admissibility condition plays a critical role in the proof of the horizontal translation invariance result (4) (see, e.g., [19, p. 1347]), as well as in the proof of the deformation stability bound (5) (see, e.g., [19, Eq. 2.51]). It is therefore unclear how Mallat's proof techniques could be generalized to arbitrary convolutional transforms. Third, to the best of our knowledge, no mother wavelet $\psi \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$, for $d \geq 2$, satisfying the scattering admissibility condition (6) has been reported in the literature. In contrast, our proof techniques are completely detached from the algebraic structures of the frames Ψ_n in the module-sequence $\Omega = \big((\Psi_n, M_n, R_n)\big)_{n\in\mathbb{N}}$. Rather, it suffices to employ (i) a module-sequence Ω that satisfies the weak admissibility condition (13), (ii) non-linearities M_n that commute with the translation operator T_t, (iii) output-generating atoms χ_n that satisfy the decay condition (17), and (iv) sub-sampling factors R_n such that $\lim_{n\to\infty} R_1 \cdot R_2 \cdot \ldots \cdot R_n = \infty$. All these conditions were shown above to be easily satisfied in practice.

APPENDIX

A. Appendix: Semi-discrete shift-invariant frames

This appendix gives a brief review of the theory of semi-discrete shift-invariant frames [41, Section 5.1.5]. A list of structured example frames that are 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 shift-invariant frames are instances of *continuous* frames [38], [39], and appear in the literature, e.g., in the context of translation-covariant signal decompositions [44], [53], [60], and as an intermediate step in the construction of various

fully-discrete frames [29], [61], [62]. We first collect some basic results on semi-discrete shift-invariant frames.

Definition 5. *Let $\{g_\lambda\}_{\lambda\in\Lambda} \subseteq L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ be a set of functions indexed by a countable set Λ. The collection*

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

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

$$A\|f\|_2^2 \leq \sum_{\lambda\in\Lambda} \int_{\mathbb{R}^d} |\langle f, T_b I g_\lambda\rangle|^2 \mathrm{d}b = \sum_{\lambda\in\Lambda} \|f * g_\lambda\|_2^2 \leq B\|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d). \tag{24}$$

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

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

$$S_\Lambda f := \sum_{\lambda\in\Lambda} \int_{\mathbb{R}^d} \langle f, T_b I g_\lambda\rangle (T_b I g_\lambda)\ \mathrm{d}b = \Big(\sum_{\lambda\in\Lambda} g_\lambda * I g_\lambda\Big) * f, \tag{25}$$

where $\langle f, T_b I g_\lambda\rangle = (f * g_\lambda)(b)$, $(\lambda, b) \in \Lambda \times \mathbb{R}^d$, are called the frame coefficients. S_Λ is a bounded, positive, and boundedly invertible operator [41, Theorem 5.11].

The reader might want to think of semi-discrete shift-invariant frames as shift-invariant frames [63], [64] with a continuous translation parameter, and of the countable index set Λ as labeling a collection of scales, directions, or frequency-shifts, hence the terminology *semi-discrete*. For instance, Mallat's scattering network is based on a semi-discrete shift-invariant wavelet frame, where the atoms $\{g_\lambda\}_{\lambda\in\Lambda_{DW}}$ are indexed by the set $\Lambda_{DW} := \big\{(-J, 0)\big\} \cup \big\{(j, k) \mid j \in \mathbb{Z} \text{ with } j > -J,\ k \in \{0, \ldots, K-1\}\big\}$ labeling a collection of scales j and directions k.

The following result gives a so-called Littlewood-Paley condition [65], [66] for the collection $\Psi_\Lambda = \{T_b I g_\lambda\}_{(\lambda,b)\in\Lambda\times\mathbb{R}^d}$ to form a semi-discrete shift-invariant frame.

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

$$A \leq \sum_{\lambda\in\Lambda} |\widehat{g_\lambda}(\omega)|^2 \leq B, \quad a.e.\ \omega \in \mathbb{R}^d. \tag{26}$$

Proof. The proof is standard and can be found, e.g., in [41, Theorem 5.11]. □

Remark 3. *What is behind Proposition 2 is a result on the unitary equivalence between operators [67, Definition 5.19.3]. Specifically, Proposition 2 follows from the fact that the multiplier* $\sum_{\lambda\in\Lambda}|\widehat{g_\lambda}|^2$ *is unitarily equivalent to the frame operator* S_Λ *in* (25) *according to*

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

where $\mathcal{F}:L^2(\mathbb{R}^d)\to L^2(\mathbb{R}^d)$ *denotes the Fourier transform. We refer the interested reader to [68], where the framework of unitary equivalence was formalized in the context of shift-invariant frames for* $\ell^2(\mathbb{Z})$.

The following proposition states normalization results for semi-discrete shift-invariant frames that come in handy in satisfying the weak admissibility condition (13) as discussed in Section III.

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

i) For $C>0$, *the family of functions*

$$\widetilde{\Psi}_\Lambda:=\big\{T_bI\widetilde{g_\lambda}\big\}_{(\lambda,b)\in\Lambda\times\mathbb{R}^d},\quad \widetilde{g_\lambda}:=C^{-1/2}g_\lambda,\ \forall\lambda\in\Lambda,$$

is a semi-discrete shift-invariant frame for $L^2(\mathbb{R}^d)$ *with frame bounds* $\widetilde{A}:=\frac{A}{C}$ *and* $\widetilde{B}:=\frac{B}{C}$.

ii) The family of functions

$$\Psi_\Lambda^\natural:=\big\{T_bIg_\lambda^\natural\big\}_{(\lambda,b)\in\Lambda\times\mathbb{R}^d},\quad g_\lambda^\natural:=\mathcal{F}^{-1}\Big(\widehat{g_\lambda}\Big(\sum_{\lambda'\in\Lambda}|\widehat{g_{\lambda'}}|^2\Big)^{-1/2}\Big),\ \forall\lambda\in\Lambda,$$

is a semi-discrete shift-invariant Parseval frame for $L^2(\mathbb{R}^d)$.

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

$$A\|f\|_2^2\le\sum_{\lambda\in\Lambda}\|f*g_\lambda\|_2^2\le B\|f\|_2^2,\quad\forall f\in L^2(\mathbb{R}^d).\tag{27}$$

With $g_\lambda=\sqrt{C}\widetilde{g_\lambda}$, for all $\lambda\in\Lambda$, in (27) we get $A\|f\|_2^2\le\sum_{\lambda\in\Lambda}\|f*\sqrt{C}\widetilde{g_\lambda}\|_2^2\le B\|f\|_2^2$, for all $f\in L^2(\mathbb{R}^d)$, which is equivalent to $\frac{A}{C}\|f\|_2^2\le\sum_{\lambda\in\Lambda}\|f*\widetilde{g_\lambda}\|_2^2\le\frac{B}{C}\|f\|_2^2$, for all $f\in L^2(\mathbb{R}^d)$, and hence establishes i). To prove statement ii), we first note that $\mathcal{F}g_\lambda^\natural=\widehat{g_\lambda}\big(\sum_{\lambda'\in\Lambda}|\widehat{g_{\lambda'}}|^2\big)^{-1/2}$, for all $\lambda\in\Lambda$, and thus $\sum_{\lambda\in\Lambda}|(\mathcal{F}g_\lambda^\natural)(\omega)|^2=\sum_{\lambda\in\Lambda}|\widehat{g_\lambda}(\omega)|^2\Big(\sum_{\lambda'\in\Lambda}|\widehat{g_{\lambda'}}(\omega)|^2\Big)^{-1}=1$, a.e. $\omega\in\mathbb{R}^d$. Application of Proposition 2 then establishes that $\Psi_\Lambda^\natural$ is a semi-discrete shift-invariant Parseval frame for $L^2(\mathbb{R}^d)$. □

B. Appendix: Examples of semi-discrete shift-invariant frames in 1-D

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

$$\Psi = \{T_b I g_k\}_{(k,b)\in\mathbb{Z}\times\mathbb{R}} \tag{28}$$

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

$$A \leq \sum_{k\in\mathbb{Z}} |\widehat{g_k}(\omega)|^2 \leq B, \quad a.e.\ \omega \in \mathbb{R}. \tag{29}$$

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 shift-invariant Weyl-Heisenberg (Gabor) frames: Weyl-Heisenberg frames [69]–[72] are well-suited to the extraction of sinusoidal features from signals [73], and have been applied successfully in various practical feature extraction tasks [50], [74]. A semi-discrete shift-invariant Weyl-Heisenberg frame for $L^2(\mathbb{R})$ is a collection of functions according to (28), where $g_m(x) := e^{2\pi imx} g(x)$, $m \in \mathbb{Z}$, with the prototype function $g \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. The atoms $\{g_m\}_{m\in\mathbb{Z}}$ satisfy the Littlewood-Paley condition (29) according to

$$A \leq \sum_{m\in\mathbb{Z}} |\widehat{g}(\omega - m)|^2 \leq B, \quad a.e.\ \omega \in \mathbb{R}. \tag{30}$$

A popular function $g \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ satisfying (30) is the Gaussian function [71].

Semi-discrete shift-invariant wavelet frames: Wavelets are well-suited to the extraction of signal features characterized by singularities [44], [66], and have been applied successfully in various practical feature extraction tasks [51], [52]. A semi-discrete shift-invariant wavelet frame for $L^2(\mathbb{R})$ is a collection of functions according to (28), where $g_j(x) := 2^j \psi(2^j x)$, $j \in \mathbb{Z}$, with the mother wavelet $\psi \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. The atoms $\{g_j\}_{j\in\mathbb{Z}}$ satisfy the Littlewood-Paley condition (29) according to

$$A \leq \sum_{j\in\mathbb{Z}} |\widehat{\psi}(2^{-j}\omega)|^2 \leq B, \quad a.e.\ \omega \in \mathbb{R}. \tag{31}$$

A large class of functions ψ satisfying (31) can be obtained through a multi-resolution analysis in $L^2(\mathbb{R})$ [41, Definition 7.1].

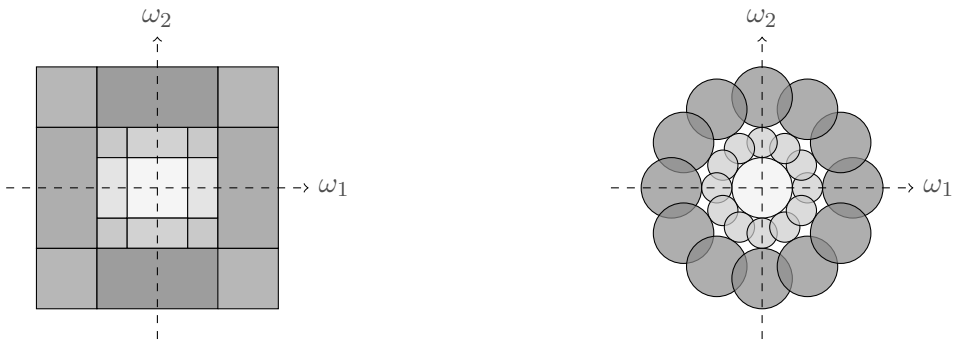

Fig. 3: Partitioning of the frequency plane $\mathbb{R}^2$ induced by (left) a semi-discrete shift-invariant tensor wavelet frame, and (right) a semi-discrete shift-invariant directional wavelet frame.

C. Examples of semi-discrete shift-invariant frames in 2-D

Semi-discrete shift-invariant 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 [75]), and have been applied successfully in various practical feature extraction tasks, e.g., in [16]–[18], [53]. Prominent families of two-dimensional wavelet frames are tensor wavelet frames and directional wavelet frames:

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

$$\Psi_{\Lambda_{TW}} := \{T_b I g_{(e,j)}\}_{(e,j)\in\Lambda_{TW}, b\in\mathbb{R}^2}, \qquad g_{(e,j)}(x) := 2^{2j}\psi^e(2^j x),$$

where $\Lambda_{TW} := \big\{((0,0),0)\big\} \cup \big\{(e,j) \mid e \in E\backslash\{(0,0)\},\ j \geq 0\big\}$, and $E := \{0,1\}^2$. Here, the functions $\psi^e \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$ are tensor products of a coarse-scale function $\phi \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ and a fine-scale function $\psi \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ according to

$$\psi^{(0,0)} := \phi \otimes \phi, \qquad \psi^{(1,0)} := \psi \otimes \phi, \qquad \psi^{(0,1)} := \phi \otimes \psi, \qquad \psi^{(1,1)} := \psi \otimes \psi.$$

The corresponding Littlewood-Paley condition (26) reads

$$A \leq \big|\widehat{\psi^{(0,0)}}(\omega)\big|^2 + \sum_{j\geq 0}\sum_{e\in E\backslash\{(0,0)\}} |\widehat{\psi^e}(2^{-j}\omega)|^2 \leq B, \qquad \text{a.e. } \omega \in \mathbb{R}^2. \tag{32}$$

A large class of functions ϕ, ψ satisfying (32) can be obtained through a multi-resolution analysis in $L^2(\mathbb{R})$ [41, Definition 7.1].

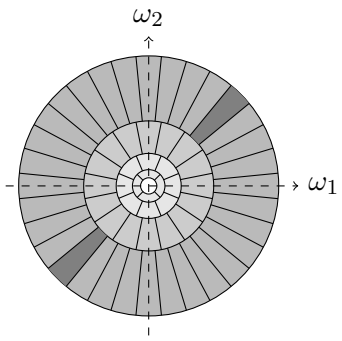

Fig. 4: Partitioning of the frequency plane $\mathbb{R}^2$ induced by a semi-discrete shift-invariant ridgelet frame.

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

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

with

$$g_{(-J,0)}(x) := 2^{-2J}\phi(2^{-J}x), \quad g_{(j,k)}(x) := 2^{2j}\psi(2^j R_{\theta_k} x),$$

where $\Lambda_{DW} := \big\{(-J,0)\big\} \cup \big\{(j,k) \mid j \in \mathbb{Z} \text{ with } j > -J,\ k \in \{0,\dots,K-1\}\big\}$, R_θ is a 2×2 rotation matrix defined as

$$R_\theta := \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}, \quad \theta \in [0, 2\pi), \tag{33}$$

and $\theta_k := \frac{2\pi k}{K}$, with $k = 0,\dots,K-1$, for a fixed $K \in \mathbb{N}$, are rotation angles. The functions $\phi \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$ and $\psi \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$ are referred to in the literature as coarse-scale wavelet and fine-scale wavelet, respectively. The integer $J \in \mathbb{Z}$ corresponds to the coarsest scale resolved and the atoms $\{g_{(j,k)}\}_{(j,k)\in\Lambda_{DW}}$ satisfy the Littlewood-Paley condition (26) according to

$$A \le |\widehat{\phi}(2^J\omega)|^2 + \sum_{j>-J}\sum_{k=0}^{K-1} |\widehat{\psi}(2^{-j}R_{\theta_k}\omega)|^2 \le B, \qquad \text{a.e. } \omega \in \mathbb{R}^2. \tag{34}$$

Prominent examples of functions ϕ, ψ satisfying (34) are the Gaussian function for ϕ and a modulated Gaussian function for ψ [41].

Semi-discrete shift-invariant ridgelet frames: Ridgelets, introduced in [26], [27], are well-suited to the extraction of signal features characterized by straight-line singularities (such as, e.g., straight edges in

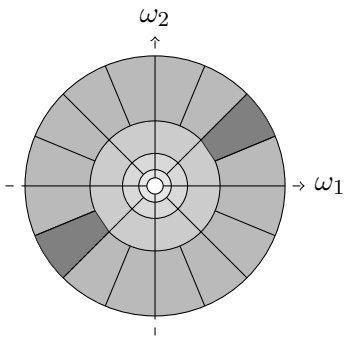

Fig. 5: Partitioning of the frequency plane $\mathbb{R}^2$ induced by a semi-discrete shift-invariant curvelet frame.

images), and have been applied successfully in various practical feature extraction tasks [54]–[56], [58].

A semi-discrete shift-invariant ridgelet frame for $L^2(\mathbb{R}^2)$ is a collection of functions according to

$$\Psi_{\Lambda_R} := \{T_b I g_{(j,l)}\}_{(j,l)\in\Lambda_R, b\in\mathbb{R}^2},$$

with

$$g_{(0,0)}(x) := \phi(x), \qquad g_{(j,l)}(x) := \psi_{(j,l)}(x),$$

where $\Lambda_R := \big\{(0,0)\big\} \cup \big\{(j,l) \mid j \geq 1,\ l = 1,\ldots,2^j-1\big\}$, and the atoms $\{g_{(j,l)}\}_{(j,l)\in\Lambda_R}$ satisfy the Littlewood-Paley condition (26) according to

$$A \leq |\widehat{\phi}(\omega)|^2 + \sum_{j=1}^{\infty}\sum_{l=1}^{2^j-1} |\widehat{\psi_{(j,l)}}(\omega)|^2 \leq B, \qquad \text{a.e. } \omega \in \mathbb{R}^2. \tag{35}$$

The functions $\psi_{(j,l)} \in L^1(\mathbb{R}^2)\cap L^2(\mathbb{R}^2)$, $(j,l) \in \Lambda_R\backslash\{(0,0)\}$, are designed to be constant in the direction specified by the parameter l, and to have a Fourier transform $\widehat{\psi}_{(j,l)}$ supported on a pair of opposite wedges of size $2^{-j} \times 2^j$ in the dyadic corona $\{\omega \in \mathbb{R}^2 \mid 2^j \leq |\omega| \leq 2^{j+1}\}$, see Figure 4. We refer the reader to [62, Section 2] for constructions of functions $\phi, \psi_{(j,l)}$ satisfying (35) with $A = B = 1$, see [62, Proposition 6].

Semi-discrete shift-invariant curvelet frames: Curvelets, introduced in [28], [29], 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 [57], [58].

A semi-discrete shift-invariant curvelet frame for $L^2(\mathbb{R}^2)$ is a collection of functions according to

$$\Psi_{\Lambda_C} := \{T_b I g_{(j,l)}\}_{(j,l)\in\Lambda_C, b\in\mathbb{R}^2},$$

with

$$g_{(-1,0)}(x) := \phi(x), \qquad g_{(j,l)}(x) := \psi_j(R_{\theta_{j,l}}x),$$

where $\Lambda_C := \big\{(-1,0)\big\} \cup \big\{(j,l) \mid j \geq 0,\ l = 0,\dots,L_j - 1\big\}$, $R_\theta \in \mathbb{R}^{2\times 2}$ is the rotation matrix defined in (33), and $\theta_{j,l} := \pi l 2^{-\lceil j/2 \rceil - 1}$, for $j \geq 0$, and $0 \leq l < L_j := 2^{\lceil j/2 \rceil + 2}$, are scale-dependent rotation angles. The functions $\phi \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$ and $\psi_j \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$ satisfy the Littlewood-Paley condition (26) according to

$$A \leq |\widehat{\phi}(\omega)|^2 + \sum_{j=0}^{\infty} \sum_{l=0}^{L_j - 1} |\widehat{\psi_j}(R_{\theta_{j,l}}\omega)|^2 \leq B, \qquad \text{a.e. } \omega \in \mathbb{R}^2. \tag{36}$$

The functions ψ_j, $j \geq 0$, are designed to have their Fourier transform $\widehat{\psi}_j$ supported on a pair of opposite wedges of size $2^{-j/2} \times 2^j$ in the dyadic corona $\{\omega \in \mathbb{R}^2 \mid 2^j \leq |\omega| \leq 2^{j+1}\}$, see Figure 5. We refer the reader to [29] for constructions of functions ϕ, ψ_j satisfying (36) with $A = B = 1$, see [29, Theorem 4.1].

Remark 4. *For further examples of interesting structured semi-discrete shift-invariant frames, we refer to [32], which discusses semi-discrete shift-invariant shearlet frames, and [30], which deals with semi-discrete shift-invariant α-curvelet frames.*

D. Appendix: 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 Corollary 1. Specifically, we need to verify the following:

(i) Lipschitz continuity: There exists a constant $L \geq 0$ such that

$$\|Mf - Mh\|_2 \leq L\|f - h\|_2, \qquad \forall f, h \in L^2(\mathbb{R}^d).$$

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

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

$$M : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d), \qquad (Mf)(x) = \rho(f(x)), \tag{37}$$

where $\rho : \mathbb{C} \to \mathbb{C}$. An immediate consequence of this property is that the operators M commute with the translation operator T_t:

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

Modulus: The modulus operator

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

has been applied successfully in the deep learning literature, e.g., in [13], [18], and most prominently in Mallat's scattering network [19]. Lipschitz continuity with $L = 1$ follows from

$$\||f| - |h|\|_2^2 = \int_{\mathbb{R}^d} ||f(x)| - |h(x)||^2 \mathrm{d}x \leq \int_{\mathbb{R}^d} |f(x) - h(x)|^2 \mathrm{d}x = \|f - h\|_2^2, \quad \forall f, h \in L^2(\mathbb{R}^d),$$

by the reverse triangle inequality. Furthermore, obviously $|f| = 0$ for $f = 0$, and finally $|\cdot|$ is pointwise as (37) is satisfied with $\rho(x) := |x|$.

Rectified linear unit: The rectified linear unit non-linearity (see, e.g., [34], [35]) is defined as

$$R : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d), \quad (Rf)(x) := \max\{0, \mathrm{Re}(f(x))\} + i \max\{0, \mathrm{Im}(f(x))\}.$$

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

$$\begin{aligned}
|(Rf)(x) - (Rh)(x)| &= \big| \max\{0, \mathrm{Re}(f(x))\} + i \max\{0, \mathrm{Im}(f(x))\} \\
&\quad - \big(\max\{0, \mathrm{Re}(h(x))\} + i \max\{0, \mathrm{Im}(h(x))\} \big) \big| \\
&\leq \big| \max\{0, \mathrm{Re}(f(x))\} - \max\{0, \mathrm{Re}(h(x))\} \big| \qquad (38) \\
&\quad + \big| \max\{0, \mathrm{Im}(f(x))\} - \max\{0, \mathrm{Im}(h(x))\} \big| \\
&\leq \big| \mathrm{Re}(f(x)) - \mathrm{Re}(h(x)) \big| + \big| \mathrm{Im}(f(x)) - \mathrm{Im}(h(x)) \big| \qquad (39) \\
&\leq \big| f(x) - h(x) \big| + \big| f(x) - h(x) \big| = 2|f(x) - h(x)|, \qquad (40)
\end{aligned}$$

where we used the triangle inequality in (38),

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

in (39), and the Lipschitz continuity (with $L = 1$) of $\mathrm{Re} : \mathbb{C} \to \mathbb{R}$ and $\mathrm{Im} : \mathbb{C} \to \mathbb{R}$ in (40). We therefore get

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

which establishes Lipschitz continuity of R with Lipschitz constant $L = 2$. Furthermore, obviously $Rf = 0$ for $f = 0$, and finally (37) is satisfied with $\rho(x) := \max\{0, \mathrm{Re}(x)\} + i \max\{0, \mathrm{Im}(x)\}$.

Hyperbolic tangent: The hyperbolic tangent non-linearity (see, e.g., [12]–[14]) is defined as

$$H : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d), \quad (Hf)(x) := \tanh(\mathrm{Re}(f(x))) + i \tanh(\mathrm{Im}(f(x))),$$

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

$$\begin{aligned}|(Hf)(x) - (Hh)(x)| &= \big| \tanh(\mathrm{Re}(f(x))) + i \tanh(\mathrm{Im}(f(x))) \\ &\quad - \big(\tanh(\mathrm{Re}(h(x))) + i \tanh(\mathrm{Im}(h(x))) \big) \big| \\ &\leq \big| \tanh(\mathrm{Re}(f(x))) - \tanh(\mathrm{Re}(h(x))) \big| \\ &\quad + \big| \tanh(\mathrm{Im}(f(x))) - \tanh(\mathrm{Im}(h(x))) \big|, \end{aligned} \tag{41}$$

where, again, we used the triangle inequality. In order to further upper-bound (41), 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 [76, Theorem 9.5.1]. □

Since $\tanh'(x) = 1 - \tanh^2(x)$, $x \in \mathbb{R}$, we have $\sup_{x \in \mathbb{R}} |\tanh'(x)| \leq 1$. By Lemma 1 we can therefore conclude that $\tanh$ is Lipschitz-continuous with $L = 1$, which when used in (41), yields

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

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

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

which proves that H is Lipschitz-continuous with $L = 2$. Since $\tanh(0) = 0$, we trivially have $Hf = 0$ for $f = 0$. Finally, (37) is satisfied with $\rho(x) := \tanh(\mathrm{Re}(x)) + i \tanh(\mathrm{Im}(x))$.

Shifted logistic sigmoid: The shifted logistic sigmoid non-linearity[9] (see, e.g., [36], [37]) is defined as

$$P : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d), \quad (Pf)(x) := \mathrm{sig}(\mathrm{Re}(f(x))) + i\mathrm{sig}(\mathrm{Im}(f(x))),$$

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

$$\begin{aligned}|(Pf)(x) - (Ph)(x)| =& \big| \mathrm{sig}(\mathrm{Re}(f(x))) + i\mathrm{sig}(\mathrm{Im}(f(x))) \\ & - \big(\mathrm{sig}(\mathrm{Re}(h(x))) + i\mathrm{sig}(\mathrm{Im}(h(x))) \big) \big| \\ \leq& \big| \mathrm{sig}(\mathrm{Re}(f(x))) - \mathrm{sig}(\mathrm{Re}(h(x))) \big| \\ & + \big| \mathrm{sig}(\mathrm{Im}(f(x))) - \mathrm{sig}(\mathrm{Im}(h(x))) \big|, \end{aligned} \tag{42}$$

where, again, we employed the triangle inequality. As before, to further upper-bound (42), we show that sig is Lipschitz-continuous. Specifically, we apply Lemma 1 with $\mathrm{sig}'(x) = \frac{e^{-x}}{(1+e^{-x})^2}$, $x \in \mathbb{R}$, and hence $\sup_{x \in \mathbb{R}} |\mathrm{sig}'(x)| \leq \frac{1}{4}$, to conclude that sig is Lipschitz-continuous with $L = \frac{1}{4}$. When used in (42) this yields (together with the Lipschitz continuity (with $L = 1$) of $\mathrm{Re} : \mathbb{C} \to \mathbb{R}$ and $\mathrm{Im} : \mathbb{C} \to \mathbb{R}$)

$$\begin{aligned}|(Pf)(x) - (Ph)(x)| \leq& \frac{1}{4} \Big| \mathrm{Re}(f(x)) - \mathrm{Re}(h(x)) \Big| + \frac{1}{4} \Big| \mathrm{Im}(f(x)) - \mathrm{Im}(h(x)) \Big| \\ \leq& \frac{1}{4} \Big| f(x) - h(x) \Big| + \frac{1}{4} \Big| f(x) - h(x) \Big| = \frac{1}{2} \Big| f(x) - h(x) \Big|. \end{aligned} \tag{43}$$

It now follows from (43) that

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

[9] 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 [36], [37]. We incorporated the offset $\frac{1}{2}$ in order to satisfy the requirement $Pf = 0$ for $f = 0$.

which establishes Lipschitz continuity of P with $L = \frac{1}{2}$. Since $\mathrm{sig}(0) = 0$, we trivially have $Pf = 0$ for $f = 0$. Finally, (37) is satisfied with $\rho(x) := \mathrm{sig}(\mathrm{Re}(x)) + i\mathrm{sig}(\mathrm{Im}(x))$.

E. Proof of Proposition 1

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

$$|||\Phi_\Omega(f)||| \leq \|f\|_2, \quad \forall f \in L^2(\mathbb{R}^d), \tag{44}$$

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 (10) and (13), we have $\|f_q\|_2 \leq \|f\|_2 < \infty$, and thus $f_q \in L^2(\mathbb{R}^d)$. We first write

$$|||\Phi_\Omega(f)|||^2 = \sum_{n=0}^{\infty} \sum_{q \in \Lambda_1^n} ||f_q * \chi_n||_2^2 = \lim_{N\to\infty} \sum_{n=0}^{N} \underbrace{\sum_{q \in \Lambda_1^n} ||f_q * \chi_n||_2^2}_{:=a_n}. \tag{45}$$

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

$$a_n \leq b_n - b_{n+1}, \quad \forall n \in \mathbb{N}_0, \tag{46}$$

with

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

and to use this result in a telescoping series argument according to

$$\begin{aligned} \sum_{n=0}^{N} a_n &\leq \sum_{n=0}^{N} (b_n - b_{n+1}) = (b_0 - b_1) + (b_1 - b_2) + \cdots + (b_N - b_{N+1}) = b_0 - \underbrace{b_{N+1}}_{\geq 0} \\ &\leq b_0 = \sum_{q \in \Lambda_1^0} \|f_q\|_2^2 = \|U[e]f\|_2^2 = \|f\|_2^2. \end{aligned} \tag{47}$$

By (45) this then implies (44). We start by noting that (46) reads

$$\sum_{q \in \Lambda_1^n} \|f_q * \chi_n\|_2^2 \leq \sum_{q \in \Lambda_1^n} ||f_q||_2^2 - \sum_{q \in \Lambda_1^{n+1}} \|f_q\|_2^2, \quad \forall n \in \mathbb{N}_0, \tag{48}$$

and proceed by examining the second term on the right hand side (RHS) of (48). Every path

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

of length $n+1$ can be decomposed into a path $q \in \Lambda_1^n$ of length n and an index $\lambda_{n+1} \in \Lambda_{n+1}$ according to $\tilde{q} = (q, \lambda_{n+1})$. Thanks to (9) we have $U[\tilde{q}] = U[(q, \lambda_{n+1})] = U_{n+1}[\lambda_{n+1}]U[q]$, which yields

$$\sum_{\tilde{q} \in \Lambda_1^{n+1}} \|f_{\tilde{q}}\|_2^2 = \sum_{q \in \Lambda_1^n} \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|U_{n+1}[\lambda_{n+1}] f_q\|_2^2. \tag{49}$$

Substituting the second term on the RHS of (48) by (49) now yields

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

which can be rewritten as

$$\sum_{q \in \Lambda_1^n} \Big(\|f_q * \chi_n\|_2^2 + \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|U_{n+1}[\lambda_{n+1}] f_q\|_2^2 \Big) \leq \sum_{q \in \Lambda_1^n} \|f_q\|_2^2, \quad \forall n \in \mathbb{N}_0. \tag{50}$$

Next, note that the second term inside the sum on the left hand side (LHS) of (50) can be written as

$$\begin{aligned}
\sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|U_{n+1}[\lambda_{n+1}] f_q\|_2^2 &= \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \int_{\mathbb{R}^d} |(U_{n+1}[\lambda_{n+1}] f_q)(x)|^2 \mathrm{d}x \\
&= \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \int_{\mathbb{R}^d} |(M_{n+1}(f_q * g_{\lambda_{n+1}}))(R_{n+1}x)|^2 \mathrm{d}x \\
&= R_{n+1}^{-d} \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \int_{\mathbb{R}^d} |(M_{n+1}(f_q * g_{\lambda_{n+1}}))(y)|^2 \mathrm{d}y \\
&= R_{n+1}^{-d} \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|M_{n+1}(f_q * g_{\lambda_{n+1}})\|_2^2, \quad \forall n \in \mathbb{N}_0.
\end{aligned} \tag{51}$$

We next use the Lipschitz property of M_{n+1}, i.e.,

$$\|M_{n+1}(f_q * g_{\lambda_{n+1}}) - M_{n+1}h\|_2 \leq L_{n+1} \|f_q * g_{\lambda_{n+1}} - h\|,$$

together with $M_{n+1}h = 0$ for $h = 0$, to upper-bound the terms inside the sum in (51) according to

$$\|M_{n+1}(f_q * g_{\lambda_{n+1}})\|_2^2 \leq L_{n+1}^2 \|f_q * g_{\lambda_{n+1}}\|_2^2, \quad \forall n \in \mathbb{N}_0. \tag{52}$$

Noting that $f_q \in L^2(\mathbb{R}^d)$, as established above, and $g_{\lambda_{n+1}} \in L^1(\mathbb{R}^d)$, by assumption, it follows that $(f_q * g_{\lambda_{n+1}}) \in L^2(\mathbb{R}^d)$ thanks to Young's inequality [59, Theorem 1.2.12]. Substituting the second term

inside the sum on the LHS of (50) by the upper bound resulting from insertion of (52) into (51) yields

$$\begin{aligned}&\sum_{q\in\Lambda_1^n}\Big(\|f_q*\chi_n\|_2^2+R_{n+1}^{-d}L_{n+1}^2\sum_{\lambda_{n+1}\in\Lambda_{n+1}}\|f_q*g_{\lambda_{n+1}}\|_2^2\Big)\\&\leq\sum_{q\in\Lambda_1^n}\max\{1,R_{n+1}^{-d}L_{n+1}^2\}\Big(\|f_q*\chi_n\|_2^2+\sum_{\lambda_{n+1}\in\Lambda_{n+1}}\|f_q*g_{\lambda_{n+1}}\|_2^2\Big),\quad\forall n\in\mathbb{N}_0.\end{aligned}\tag{53}$$

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

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

which, when used in (53) yields

$$\begin{aligned}&\sum_{q\in\Lambda_1^n}\Big(\|f_q*\chi_n\|_2^2+\sum_{\lambda_{n+1}\in\Lambda_{n+1}}\|U_{n+1}[\lambda_{n+1}]f_q\|_2^2\Big)\\&\leq\sum_{q\in\Lambda_1^n}\max\{1,R_{n+1}^{-d}L_{n+1}^2\}B_{n+1}\|f_q\|_2^2\\&=\sum_{q\in\Lambda_1^n}\max\{B_{n+1},B_{n+1}R_{n+1}^{-d}L_{n+1}^2\}\|f_q\|_2^2,\quad\forall n\in\mathbb{N}_0.\end{aligned}\tag{54}$$

Finally, invoking the assumption

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

in (54) yields (50) and thereby completes the proof.

F. Appendix: Proof of Theorem 1

The proof of the deformation stability bound (14) is based on two key ingredients. The first one, stated in Proposition 4 in Appendix G, establishes that the generalized feature extractor Φ_Ω is non-expansive, i.e.,

$$|||\Phi_\Omega(f)-\Phi_\Omega(h)|||\leq\|f-h\|_2,\quad\forall f,h\in L^2(\mathbb{R}^d),\tag{55}$$

and needs the weak admissibility condition (13) only. The second ingredient, stated in Proposition 5 in Appendix H, is an upper bound on the deformation error $\|f-F_{\tau,\omega}f\|_2$ given by

$$\|f-F_{\tau,\omega}f\|_2\leq C\big(R\|\tau\|_\infty+\|\omega\|_\infty\big)\|f\|_2,\quad\forall f\in L_R^2(\mathbb{R}^d),\tag{56}$$

and is valid under the assumptions $\omega \in C(\mathbb{R}^d, \mathbb{R})$ and $\tau \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ with $\|D\tau\|_\infty < \frac{1}{2d}$. We now show how (55) and (56) can be combined to establish the deformation stability bound (14). To this end, we first apply (55) with $h := F_{\tau,\omega} f = e^{2\pi i \omega(\cdot)} f(\cdot - \tau(\cdot))$ to get

$$|||\Phi_\Omega(f) - \Phi_\Omega(F_{\tau,\omega} f)||| \leq \|f - F_{\tau,\omega} f\|_2, \quad \forall f \in L^2(\mathbb{R}^d). \tag{57}$$

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

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

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

$$\frac{\mathrm{d}u}{\mathrm{d}x} = |\det(E - (D\tau)(x))| \geq 1 - d\|D\tau\|_\infty \geq 1/2, \quad \forall x \in \mathbb{R}^d. \tag{58}$$

The first inequality in (58) follows from:

Lemma 2. *[77, Corollary 1] Let $M \in \mathbb{R}^{d\times d}$ be such that $|M_{i,j}| \leq \alpha$, for all i, j with $1 \leq i, j \leq d$. If $d\alpha \leq 1$, then*

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

The second inequality in (58) is a consequence of the assumption $\|D\tau\|_\infty \leq \frac{1}{2d}$. The proof is finalized by replacing the RHS of (57) by the RHS of (56).

G. Appendix: Proposition 4

Proposition 4. *Let $\Omega = \big((\Psi_n, M_n, R_n)\big)_{n\in\mathbb{N}}$ be a weakly admissible module-sequence. The corresponding feature extractor $\Phi_\Omega : L^2(\mathbb{R}^d) \to (L^2(\mathbb{R}^d))^{\mathcal{Q}}$ is non-expansive, i.e.,*

$$|||\Phi_\Omega(f) - \Phi_\Omega(h)||| \leq \|f - h\|_2, \quad \forall f, h \in L^2(\mathbb{R}^d). \tag{59}$$

Remark 5. *Proposition 4 generalizes [19, Proposition 2.5], which shows that Mallat's wavelet-modulus feature extractor Φ_M is non-expansive. Specifically, our generalization allows for general semi-discrete shift-invariant frames, general Lipschitz-continuous operators, and sub-sampling operations, all of which can be different in different layers. Both, the proof of Proposition 4 stated below and the proof of [19, Proposition 2.5] employ a telescoping series argument.*

Proof. The key idea of the proof is—similarly to the proof of Proposition 1 in Appendix E—to judiciously employ telescoping series arguments. For ease of notation, we let $f_q := U[q]f$ and $h_q := U[q]h$,

for $f, h \in L^2(\mathbb{R}^d)$. Thanks to (10) and the weak admissibility condition (13), we have $\|f_q\|_2 \leq \|f\|_2 < \infty$ and $\|h_q\|_2 \leq \|h\|_2 < \infty$ and thus $f_q, h_q \in L^2(\mathbb{R}^d)$. We start by writing

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

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

$$a_n \leq b_n - b_{n+1}, \quad \forall n \in \mathbb{N}_0, \tag{60}$$

where here $b_n := \sum_{q \in \Lambda_1^n} \|f_q - h_q\|_2^2$, $\forall n \in \mathbb{N}_0$, and to note that, similarly to (47),

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

which then yields (59) according to

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

Writing out (60), it follows that we need to establish

$$\sum_{q \in \Lambda_1^n} \|f_q * \chi_n - h_q * \chi_n\|_2^2 \leq \sum_{q \in \Lambda_1^n} ||f_q - h_q||_2^2 - \sum_{q \in \Lambda_1^{n+1}} \|f_q - h_q\|_2^2, \quad \forall n \in \mathbb{N}_0. \tag{61}$$

We start by examining the second term on the RHS of (61) and note that, thanks to the decomposition

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

and $U[\tilde{q}] = U[(q, \lambda_{n+1})] = U_{n+1}[\lambda_{n+1}]U[q]$, by (9), we have

$$\sum_{\tilde{q} \in \Lambda_1^{n+1}} \|f_{\tilde{q}} - h_{\tilde{q}}\|_2^2 = \sum_{q \in \Lambda_1^n} \sum_{\lambda_{n+1} \in \Lambda_{n+1}} \|U_{n+1}[\lambda_{n+1}]f_q - U_{n+1}[\lambda_{n+1}]h_q\|_2^2. \tag{62}$$

Substituting (62) into (61) and rearranging terms, we obtain

$$\sum_{q\in\Lambda_1^n}\Big(\|f_q*\chi_n-h_q*\chi_n\|_2^2+\sum_{\lambda_{n+1}\in\Lambda_{n+1}}\|U_{n+1}[\lambda_{n+1}]f_q-U_{n+1}[\lambda_{n+1}]h_q\|_2^2\Big)$$
$$\leq\sum_{q\in\Lambda_1^n}\|f_q-h_q\|_2^2,\quad\forall n\in\mathbb{N}_0. \tag{63}$$

We next note that the second term inside the sum on the LHS of (63) satisfies

$$\sum_{\lambda_{n+1}\in\Lambda_{n+1}}\|U_{n+1}[\lambda_{n+1}]f_q-U_{n+1}[\lambda_{n+1}]h_q\|_2^2$$
$$\leq R_{n+1}^{-d}\sum_{\lambda_{n+1}\in\Lambda_{n+1}}\|M_{n+1}(f_q*g_{\lambda_{n+1}})-M_{n+1}(h_q*g_{\lambda_{n+1}})\|_2^2, \tag{64}$$

where we employed arguments similar to those leading to (51). Substituting the second term inside the sum on the LHS of (63) by the upper bound (64), and using the Lipschitz property of M_{n+1} yields

$$\sum_{q\in\Lambda_1^n}\Big(\|f_q*\chi_n-h_q*\chi_n\|_2^2+\sum_{\lambda_{n+1}\in\Lambda_{n+1}}\|U_{n+1}[\lambda_{n+1}]f_q-U_{n+1}[\lambda_{n+1}]h_q\|_2^2\Big)$$
$$\leq\sum_{q\in\Lambda_1^n}\max\{1,R_{n+1}^{-d}L_{n+1}^2\}\Big(\|(f_q-h_q)*\chi_n\|_2^2+\sum_{\lambda_{n+1}\in\Lambda_{n+1}}\|(f_q-h_q)*g_{\lambda_{n+1}}\|_2^2\Big), \tag{65}$$

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

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

which, when used in (65) yields

$$\sum_{q\in\Lambda_1^n}\Big(\|f_q*\chi_n-h_q*\chi_n\|_2^2+\sum_{\lambda_{n+1}\in\Lambda_{n+1}}\|U_{n+1}[\lambda_{n+1}]f_q-U_{n+1}[\lambda_{n+1}]h_q\|_2^2\Big)$$
$$\leq\sum_{q\in\Lambda_1^n}\max\{B_{n+1},B_{n+1}R_{n+1}^{-d}L_{n+1}^2\}\|f_q-h_q\|_2^2,\quad\forall n\in\mathbb{N}_0. \tag{66}$$

Finally, invoking the assumption

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

in (66) we get (63) and hence (60). This completes the proof. □

H. Appendix: Proposition 5

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

$$\|f - F_{\tau,\omega} f\|_2 \leq C\big(R\|\tau\|_\infty + \|\omega\|_\infty\big)\|f\|_2. \tag{67}$$

Remark 6. *A similar bound was derived by Mallat in [19, App. B], namely*

$$\|f * \psi_{(-J,0)} - F_\tau(f * \psi_{(-J,0)})\|_2 \leq C 2^{-J+d}\|\tau\|_\infty \|f\|_2, \quad \forall f \in L^2(\mathbb{R}^d), \tag{68}$$

where $\psi_{(-J,0)}$ is the low-pass filter of a semi-discrete shift-invariant directional wavelet frame for $L^2(\mathbb{R}^d)$, and $(F_\tau f)(x) = f(x - \tau(x))$. The techniques for proving (67) *and* (68) *are related in the sense of both employing Schur's Lemma [59, App. I.1] and a Taylor expansion argument [78, p. 411]. The major difference between our bound* (67) *and Mallat's bound* (68) *is that in* (67) *time-frequency deformations $F_{\tau,\omega}$ act on band-limited-functions $f \in L^2_R(\mathbb{R}^d)$, whereas in* (68) *translation-like deformations F_τ act on low-pass filtered functions $f * \psi_{(-J,0)}$.*

Proof. We first determine an integral operator

$$(Kf)(x) = \int_{\mathbb{R}^d} k(x,u) f(u) \mathrm{d}u \tag{69}$$

satisfying $Kf = F_{\tau,\omega} f - f$, for all $f \in L^2_R(\mathbb{R}^d)$, and then upper-bound the deformation error $\|f - F_{\tau,\omega} f\|_2$ according to

$$\|f - F_{\tau,\omega} f\|_2 = \|F_{\tau,\omega} f - f\|_2 = \|Kf\|_2 \leq \|K\|_{2,2}\|f\|_2, \quad \forall f \in L^2_R(\mathbb{R}^d).$$

Application of Schur's Lemma, stated below, then yields an upper bound on $\|K\|_{2,2}$ according to

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

which completes the proof.

Schur's Lemma. *[59, App. I.1] Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ be a locally integrable function satisfying*

$$(i)\ \sup_{x \in \mathbb{R}^d} \int_{\mathbb{R}^d} |k(x,u)| \mathrm{d}u \leq \alpha, \quad (ii)\ \sup_{u \in \mathbb{R}^d} \int_{\mathbb{R}^d} |k(x,u)| \mathrm{d}x \leq \alpha, \tag{70}$$

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

operator norm $\|K\|_{2,2} \leq \alpha$.

We start by determining the integral operator K in (69). To this end, consider $\eta \in S(\mathbb{R}^d, \mathbb{C})$ such that $\widehat{\eta}(\omega) = 1$, for all $\omega \in B_1(0)$. Setting $\gamma(x) := R^d \eta(Rx)$ yields $\gamma \in S(\mathbb{R}^d, \mathbb{C})$ and $\widehat{\gamma}(\omega) = \widehat{\eta}(\omega/R)$. Thus, $\widehat{\gamma}(\omega) = 1$, for all $\omega \in B_R(0)$, and hence $\widehat{f} = \widehat{f} \cdot \widehat{\gamma}$, so that $f = f * \gamma$, for all $f \in L^2_R(\mathbb{R}^d)$. Next, we define the operator $A_\gamma : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$, $A_\gamma f := f * \gamma$, and note that A_γ is well-defined, i.e., $A_\gamma f \in L^2(\mathbb{R}^d)$, for all $f \in L^2(\mathbb{R}^d)$, thanks to Young's inequality [59, Theorem 1.2.12] (since $f \in L^2(\mathbb{R}^d)$ and $\gamma \in S(\mathbb{R}^d, \mathbb{C}) \subseteq L^1(\mathbb{R}^d)$). Moreover, $A_\gamma f = f$, for all $f \in L^2_R(\mathbb{R}^d)$. Setting $K := F_{\tau,\omega} A_\gamma - A_\gamma$ we get $Kf = F_{\tau,\omega} A_\gamma f - A_\gamma f = F_{\tau,\omega} f - f$, for all $f \in L^2_R(\mathbb{R}^d)$, as desired. Furthermore, it follows from

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

that the integral operator $K = F_{\tau,\omega} A_\gamma - A_\gamma$, i.e., $(Kf)(x) = \int_{\mathbb{R}^d} k(x,u) f(u) \mathrm{d}u$, has the kernel

$$k(x,u) := e^{2\pi i \omega(x)} \gamma(x - \tau(x) - u) - \gamma(x - u). \tag{71}$$

Before we can apply Schur's Lemma to establish an upper bound on $\|K\|_{2,2}$, we need to verify that k in (71) is locally integrable, i.e., we need to show that for every compact set $S \subseteq \mathbb{R}^d \times \mathbb{R}^d$ we have $\int_S |k(x,u)| \mathrm{d}(x,u) < \infty$. To this end, let $S \subseteq \mathbb{R}^d \times \mathbb{R}^d$ be a compact set. Next, choose compact sets $S_1, S_2 \subseteq \mathbb{R}^d$ such that $S \subseteq S_1 \times S_2$. Thanks to $\gamma \in S(\mathbb{R}^d, \mathbb{C})$, $\tau \in C^1(\mathbb{R}^d, \mathbb{R}^d)$, and $\omega \in C(\mathbb{R}^d, \mathbb{R})$, all by assumption, the function $|k| : S_1 \times S_2 \to \mathbb{C}$ is continuous as a composition of continuous functions, and therefore also Lebesgue-measurable. We further have

$$\begin{aligned}
&\int_{S_1} \int_{S_2} |k(x,u)| \mathrm{d}x \mathrm{d}u \leq \int_{S_1} \int_{\mathbb{R}^d} |k(x,u)| \mathrm{d}x \mathrm{d}u \\
&\leq \int_{S_1} \int_{\mathbb{R}^d} |\gamma(x - \tau(x) - u)| \mathrm{d}x \mathrm{d}u + \int_{S_1} \int_{\mathbb{R}^d} |\gamma(x - u)| \mathrm{d}x \mathrm{d}u \\
&\leq 2 \int_{S_1} \int_{\mathbb{R}^d} |\gamma(y)| \mathrm{d}y \mathrm{d}u + \int_{S_1} \int_{\mathbb{R}^d} |\gamma(y)| \mathrm{d}y \ \mathrm{d}u = 3 \mu_L(S_1) \|\gamma\|_1 < \infty,
\end{aligned} \tag{72}$$

where the first term in (72) follows by the change of variables $y = x - \tau(x) - u$, together with

$$\frac{\mathrm{d}y}{\mathrm{d}x} = |\det(E - (D\tau)(x))| \geq 1 - d\|D\tau\|_\infty \geq 1/2, \quad \forall x \in \mathbb{R}^d. \tag{73}$$

The arguments underlying (73) were already detailed at the end of Appendix F. It follows that k is

locally integrable owing to

$$\int_S |k(x,u)|\mathrm{d}(x,u) \leq \int_{S_1\times S_2} |k(x,u)|\mathrm{d}(x,u) = \int_{S_1}\int_{S_2} |k(x,u)|\mathrm{d}x\mathrm{d}u < \infty, \tag{74}$$

where the first step in (74) follows from $S \subseteq S_1 \times S_2$, the second step is thanks to the Fubini-Tonelli Theorem [79, Theorem 14.2] noting that $|k| : S_1 \times S_2 \to \mathbb{C}$ is Lebesgue-measurable (as established above) and non-negative, and the last step is due to (72). Next, we need to verify conditions (i) and (ii) in (70) and determine the corresponding $\alpha > 0$. In fact, we seek a specific constant α of the form

$$\alpha = C\big(R\|\tau\|_\infty + \|\omega\|_\infty\big), \qquad \text{with } C > 0. \tag{75}$$

This will be accomplished as follows: For $x, u \in \mathbb{R}^d$, we parametrize the integral kernel in (71) according to $h_{x,u}(t) := e^{2\pi i t\omega(x)}\gamma(x - t\tau(x) - u) - \gamma(x-u)$. A Taylor expansion [78, p. 411] of $h_{x,u}(t)$ w.r.t. the variable t now yields

$$h_{x,u}(t) = \underbrace{h_{x,u}(0)}_{=0} + \int_0^t h'_{x,u}(\lambda)\mathrm{d}\lambda = \int_0^t h'_{x,u}(\lambda)\mathrm{d}\lambda, \quad \forall t \in \mathbb{R}, \tag{76}$$

where $h'_{x,u}(t) = (\frac{\mathrm{d}}{\mathrm{d}t}h_{x,u})(t)$. 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 (76) we get

$$|k(x,u)| = |h_{x,u}(1)| \leq \int_0^1 |h'_{x,u}(\lambda)|\mathrm{d}\lambda, \tag{77}$$

where

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

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

$$\begin{aligned} |h'_{x,u}(\lambda)| &\leq \big|\big\langle\nabla\gamma(x-\lambda\tau(x)-u),\tau(x)\big\rangle\big| + |2\pi\omega(x)\gamma(x-\lambda\tau(x)-u)| \\ &\leq |\tau(x)||\nabla\gamma(x-\lambda\tau(x)-u)| + 2\pi|\omega(x)||\gamma(x-\lambda\tau(x)-u)|. \end{aligned} \tag{78}$$

Now, using $|\tau(x)| \leq \sup\limits_{y\in\mathbb{R}^d}|\tau(y)| = \|\tau\|_\infty$ and $|\omega(x)| \leq \sup\limits_{y\in\mathbb{R}^d}|\omega(y)| = \|\omega\|_\infty$ in (78), together with (77), we get the upper bound

$$|k(x,u)| \leq \|\tau\|_\infty \int_0^1 |\nabla\gamma(x-\lambda\tau(x)-u)|\mathrm{d}\lambda + 2\pi\|\omega\|_\infty \int_0^1 |\gamma(x-\lambda\tau(x)-u)|\mathrm{d}\lambda. \tag{79}$$

Next, we integrate (79) w.r.t. u to establish (i) in (70):

$$\begin{aligned}
&\int_{\mathbb{R}^d} |k(x,u)| \mathrm{d}u \\
&\leq \|\tau\|_\infty \int_{\mathbb{R}^d} \int_0^1 |\nabla\gamma(x-\lambda\tau(x)-u)| \mathrm{d}\lambda \mathrm{d}u + 2\pi\|\omega\|_\infty \int_{\mathbb{R}^d} \int_0^1 |\gamma(x-\lambda\tau(x)-u)| \mathrm{d}\lambda \mathrm{d}u \\
&= \|\tau\|_\infty \int_0^1 \int_{\mathbb{R}^d} |\nabla\gamma(x-\lambda\tau(x)-u)| \mathrm{d}u \mathrm{d}\lambda + 2\pi\|\omega\|_\infty \int_0^1 \int_{\mathbb{R}^d} |\gamma(x-\lambda\tau(x)-u)| \mathrm{d}u \mathrm{d}\lambda \qquad (80) \\
&= \|\tau\|_\infty \int_0^1 \int_{\mathbb{R}^d} |\nabla\gamma(y)| \mathrm{d}y \mathrm{d}\lambda + 2\pi\|\omega\|_\infty \int_0^1 \int_{\mathbb{R}^d} |\gamma(y)| \mathrm{d}y \mathrm{d}\lambda \\
&= \|\tau\|_\infty \|\nabla\gamma\|_1 + 2\pi\|\omega\|_\infty \|\gamma\|_1, \qquad (81)
\end{aligned}$$

where (80) follows by application of the Fubini-Tonelli Theorem [79, 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 $\nabla\gamma = R^{d+1}\nabla\eta(R\cdot)$, $\|\gamma\|_1 = \|\eta\|_1$, and $\|\nabla\gamma\|_1 = R\|\nabla\eta\|_1$ in (81) yields

$$\begin{aligned}
\sup_{x\in\mathbb{R}^d} \int_{\mathbb{R}^d} |k(x,u)| \mathrm{d}u &\leq R\|\tau\|_\infty \|\nabla\eta\|_1 + 2\pi\|\omega\|_\infty \|\eta\|_1 \\
&\leq \max\big\{\|\nabla\eta\|_1, 2\pi\|\eta\|_1\big\}\big(R\|\tau\|_\infty + \|\omega\|_\infty\big), \qquad (82)
\end{aligned}$$

which establishes an upper bound of the form (i) in (70) that exhibits the desired structure for α. Condition (ii) in (70) is established similarly by integrating (79) w.r.t. x according to

$$\begin{aligned}
&\int_{\mathbb{R}^d} |k(x,u)| \mathrm{d}x \\
&\leq \|\tau\|_\infty \int_{\mathbb{R}^d} \int_0^1 |\nabla\gamma(x-\lambda\tau(x)-u)| \mathrm{d}\lambda \mathrm{d}x + 2\pi\|\omega\|_\infty \int_{\mathbb{R}^d} \int_0^1 |\gamma(x-\lambda\tau(x)-u)| \mathrm{d}\lambda \mathrm{d}x \\
&= \|\tau\|_\infty \int_0^1 \int_{\mathbb{R}^d} |\nabla\gamma(x-\lambda\tau(x)-u)| \mathrm{d}x \mathrm{d}\lambda + 2\pi\|\omega\|_\infty \int_0^1 \int_{\mathbb{R}^d} |\gamma(x-\lambda\tau(x)-u)| \mathrm{d}x \mathrm{d}\lambda \qquad (83) \\
&\leq 2\,\|\tau\|_\infty \int_0^1 \int_{\mathbb{R}^d} |\nabla\gamma(y)| \mathrm{d}y \mathrm{d}\lambda + 4\pi\|\omega\|_\infty \int_0^1 \int_{\mathbb{R}^d} |\gamma(y)| \mathrm{d}y \mathrm{d}\lambda \qquad (84) \\
&= 2\,\|\tau\|_\infty \|\nabla\gamma\|_1 + 4\pi\|\omega\|_\infty \|\gamma\|_1 \leq \max\big\{2\|\nabla\eta\|_1, 4\pi\|\eta\|_1\big\}\big(R\|\tau\|_\infty + \|\omega\|_\infty\big), \qquad (85)
\end{aligned}$$

which yields an upper bound of the form (ii) in (70) with the desired structure for α. Here, again, (83) follows by application of the Fubini-Tonelli Theorem [79, 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 a composition of continuous functions. The inequality (84) follows from a change of variables argument similar to the one in (72) and (73). Combining (82) and (85), we finally get (75) with

$$C := \max\big\{2\|\nabla\eta\|_1, 4\pi\|\eta\|_1\big\}. \tag{86}$$

This completes the proof. □

I. Appendix: Proof of Theorem 2

We start by proving i). The key step in establishing (16) is to show that the operator U_n, $n \in \mathbb{N}$, defined in (7) satisfies the relation

$$U_n[\lambda_n]T_t f = T_{t/R_n} U_n[\lambda_n] f, \quad \forall f \in L^2(\mathbb{R}^d),\, \forall t \in \mathbb{R}^d,\, \forall \lambda_n \in \Lambda_n. \tag{87}$$

With the definition of $U[q]$ in (9) this then yields

$$U[q]T_t f = T_{t/(R_1 R_2 \cdots R_n)} U[q] f, \quad \forall f \in L^2(\mathbb{R}^d),\, \forall t \in \mathbb{R}^d,\, \forall q \in \Lambda_1^n. \tag{88}$$

The identity (16) is then a direct consequence of (88) and the translation-covariance of the convolution operator:

$$\begin{aligned}
\Phi_\Omega^n(T_t f) &= \big\{\big(U[q]T_t f\big) * \chi_n\big\}_{q\in\Lambda_1^n} = \big\{\big(T_{t/(R_1R_2\cdots R_n)}U[q]f\big) * \chi_n\big\}_{q\in\Lambda_1^n} \\
&= \big\{T_{t/(R_1R_2\cdots R_n)}\big((U[q]f) * \chi_n\big)\big\}_{q\in\Lambda_1^n} = T_{t/(R_1R_2\cdots R_n)}\big\{(U[q]f) * \chi_n\big\}_{q\in\Lambda_1^n} \\
&= T_{t/(R_1R_2\cdots R_n)}\Phi_\Omega^n(f), \quad \forall f \in L^2(\mathbb{R}^d),\, \forall t \in \mathbb{R}^d.
\end{aligned}$$

To establish (87), we first define the operator $D_n : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$, $D_n f := f(R_n \cdot)$, and note that

$$\begin{aligned}
U_n[\lambda_n]T_t f &= (M_n((T_t f) * g_{\lambda_n}))(R_n\cdot) = D_n M_n((T_t f) * g_{\lambda_n}) \\
&= D_n M_n T_t (f * g_{\lambda_n}) = D_n T_t M_n (f * g_{\lambda_n}), \quad \forall f \in L^2(\mathbb{R}^d),\, \forall t \in \mathbb{R}^d,
\end{aligned} \tag{89}$$

where, in the last step, we employed $M_n T_t = T_t M_n$, for all $n \in \mathbb{N}$, and all $t \in \mathbb{R}$, which is by assumption. Next, using

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

in (89) yields

$$U_n[\lambda_n]T_t f = D_n T_t M_n (f * g_{\lambda_n}) = T_{t/R_n} D_n M_n (f * g_{\lambda_n}) = T_{t/R_n} U_n[\lambda_n] f,$$

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

Next, we prove ii). For ease of notation, again, we let $f_q := U[q]f$, for $f \in L^2(\mathbb{R}^d)$. Thanks to (10) and the weak admissibility condition (13), we have $\|f_q\|_2 \leq \|f\|_2 < \infty$, and thus $f_q \in L^2(\mathbb{R}^d)$. We first write

$$\begin{aligned} |||\Phi_\Omega^n(T_t f) - \Phi_\Omega^n(f)|||^2 &= |||T_{t/(R_1 \dots R_n)} \Phi_\Omega^n(f) - \Phi_\Omega^n(f)|||^2 && (90) \\ &= \sum_{q \in \Lambda_1^n} \|T_{t/(R_1 \dots R_n)}(f_q * \chi_n) - f_q * \chi_n\|_2^2 \\ &= \sum_{q \in \Lambda_1^n} \|M_{-t/(R_1 \dots R_n)}(\widehat{f_q * \chi_n}) - \widehat{f_q * \chi_n}\|_2^2, \quad \forall n \in \mathbb{N}, && (91) \end{aligned}$$

where in (90) we used (16), and in (91) we employed Parseval's formula [40, p. 189] (noting that $(f_q * \chi_n) \in L^2(\mathbb{R}^d)$ thanks to Young's inequality [59, Theorem 1.2.12]) together with the relation $\widehat{T_t f} = M_{-t}\widehat{f}$, for all $f \in L^2(\mathbb{R}^d)$, and all $t \in \mathbb{R}^d$. The key step is then to establish the upper bound

$$\|M_{-t/(R_1 \dots R_n)}(\widehat{f_q * \chi_n}) - \widehat{f_q * \chi_n}\|_2^2 \leq \frac{4\pi^2 |t|^2 K^2}{(R_1 \dots R_n)^2} \|f_q\|_2^2, \quad \forall n \in \mathbb{N}, \tag{92}$$

where $K > 0$ corresponds to the constant in the decay condition (17), and to note that

$$\sum_{q \in \Lambda_1^n} \|f_q\|_2^2 \leq \sum_{q \in \Lambda_1^{n-1}} \|f_q\|_2^2, \quad \forall n \in \mathbb{N}, \tag{93}$$

which follows from (46) thanks to

$$0 \leq \sum_{q \in \Lambda_1^{n-1}} ||f_q * \chi_{n-1}||_2^2 = a_{n-1} \leq b_{n-1} - b_n = \sum_{q \in \Lambda_1^{n-1}} \|f_q\|_2^2 - \sum_{q \in \Lambda_1^n} \|f_q\|_2^2, \quad \forall n \in \mathbb{N}.$$

Iterating on (93) yields

$$\sum_{q \in \Lambda_1^n} \|f_q\|_2^2 \leq \sum_{q \in \Lambda_1^{n-1}} \|f_q\|_2^2 \leq \cdots \leq \sum_{q \in \Lambda_1^0} \|f_q\|_2^2 = \|U[e]f\|_2^2 = \|f\|_2^2, \quad \forall n \in \mathbb{N}. \tag{94}$$

The identity (91) together with the inequalities (92) and (94) then directly imply

$$|||\Phi_\Omega^n(T_t f) - \Phi_\Omega^n(f)|||^2 \leq \frac{4\pi^2 |t|^2 K^2}{(R_1 \dots R_n)^2} \|f\|_2^2, \quad \forall n \in \mathbb{N}. \tag{95}$$

It remains to prove (92). To this end, we first note that

$$\begin{aligned}&\|M_{-t/(R_1\dots R_n)}(\widehat{f_q * \chi_n}) - \widehat{f_q * \chi_n}\|_2^2 \\ &= \int_{\mathbb{R}^d} |e^{-2\pi i\langle t,\omega\rangle/(R_1\dots R_n)} - 1|^2 |\widehat{\chi_n}(\omega)|^2 |\widehat{f}_q(\omega)|^2 \mathrm{d}\omega. \end{aligned} \tag{96}$$

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

$$|e^{-2\pi i\langle t,\omega\rangle/(R_1\dots R_n)} - 1|^2 \leq \frac{4\pi^2|\langle t,\omega\rangle|^2}{(R_1\dots R_n)^2} \leq \frac{4\pi^2|t|^2|\omega|^2}{(R_1\dots R_n)^2}, \tag{97}$$

where in the last step we employed the Cauchy-Schwartz inequality. Substituting (97) into (96) yields

$$\begin{aligned}&\|M_{-t/(R_1\dots R_n)}(\widehat{f_q * \chi_n}) - \widehat{f_q * \chi_n}\|_2^2 \\ &\leq \frac{4\pi^2|t|^2}{(R_1\dots R_n)^2} \int_{\mathbb{R}^d} |\omega|^2 |\widehat{\chi_n}(\omega)|^2 |\widehat{f}_q(\omega)|^2 \mathrm{d}\omega \\ &\leq \frac{4\pi^2|t|^2 K^2}{(R_1\dots R_n)^2} \int_{\mathbb{R}^d} |\widehat{f}_q(\omega)|^2 \mathrm{d}\omega \qquad (98) \\ &= \frac{4\pi^2|t|^2 K^2}{(R_1\dots R_n)^2} \|\widehat{f}_q\|_2^2 = \frac{4\pi^2|t|^2 K^2}{(R_1\dots R_n)^2} \|f_q\|_2^2, \quad \forall n \in \mathbb{N}, \qquad (99)\end{aligned}$$

where in (98) we employed the decay condition (17), and in the last step, again, we used Parseval's formula [40, p. 189]. This establishes (92) and thereby completes the proof of ii).

J. Appendix: Proof of Corollary 1

The key idea of the proof is—similarly to the proof of ii) in Theorem 2—to upper-bound the deviation from perfect covariance in the frequency domain. For ease of notation, again, we let $f_q := U[q]f$, for $f \in L^2(\mathbb{R}^d)$. Thanks to (10) and the weak admissibility condition (13), we have $\|f_q\|_2 \leq \|f\|_2 < \infty$, and thus $f_q \in L^2(\mathbb{R}^d)$. We first write

$$\begin{aligned}|||\Phi_\Omega^n(T_t f) - T_t\Phi_\Omega^n(f)|||^2 &= |||T_{t/(R_1\dots R_n)}\Phi_\Omega^n(f) - T_t\Phi_\Omega^n(f)|||^2 \qquad (100) \\ &= \sum_{q\in\Lambda_1^n} \|(T_{t/(R_1\dots R_n)} - T_t)(f_q * \chi_n)\|_2^2 \\ &= \sum_{q\in\Lambda_1^n} \|(M_{-t/(R_1\dots R_n)} - M_{-t})(\widehat{f_q * \chi_n})\|_2^2, \quad \forall n \in \mathbb{N}, \qquad (101)\end{aligned}$$

where in (100) we used (16), and in (101) we employed Parseval's formula [40, p. 189] (noting that $(f_q * \chi_n) \in L^2(\mathbb{R}^d)$ thanks to Young's inequality [59, Theorem 1.2.12]) together with the relation

$\widehat{T_t f} = M_{-t}\widehat{f}$, for all $f \in L^2(\mathbb{R}^d)$, and all $t \in \mathbb{R}^d$. The key step is then to establish the upper bound

$$\|(M_{-t/(R_1 \dots R_n)} - M_{-t})(\widehat{f_q * \chi_n})\|_2^2 \leq 4\pi^2 |t|^2 K^2 \big|1/(R_1 \dots R_n) - 1\big|^2 \|f_q\|_2^2, \tag{102}$$

where $K > 0$ corresponds to the constant in the decay condition (17). Arguments similar to those leading to (95) then complete the proof. It remains to prove (102):

$$\begin{aligned} &\|(M_{-t/(R_1 \dots R_n)} - M_{-t})(\widehat{f_q * \chi_n})\|_2^2 \\ &= \int_{\mathbb{R}^d} \left| e^{-2\pi i \langle t,\omega \rangle/(R_1 \dots R_n)} - e^{-2\pi i \langle t,\omega \rangle} \right|^2 |\widehat{\chi_n}(\omega)|^2 |\widehat{f_q}(\omega)|^2 \mathrm{d}\omega. \end{aligned} \tag{103}$$

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

$$\left| e^{-2\pi i \langle t,\omega \rangle/(R_1 \dots R_n)} - e^{-2\pi i \langle t,\omega \rangle} \right|^2 \leq 4\pi^2 |t|^2 |\omega|^2 \big|1/(R_1 \dots R_n) - 1\big|^2, \tag{104}$$

where, again, we employed the Cauchy-Schwartz inequality. Substituting (104) into (103), and employing arguments similar to those leading to (99), establishes (102) and thereby completes the proof.

ACKNOWLEDGMENTS

The authors would like to thank Rima Alaifari, Philipp Grohs, Gitta Kutyniok, and Michael Tschannen 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.

[7] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," in *Proc. of International Conference on Neural Information Processing Systems (NIPS)*, pp. 396–404, 1990.

[8] D. E. Rumelhart, G. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel distributed processing: Explorations in the microstructure of cognition* (J. L. McClelland and D. E. Rumelhart, eds.), pp. 318–362, MIT Press, 1986.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proc. of the IEEE*, pp. 2278–2324, 1998.

[10] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proc. of IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 253–256, 2010.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[12] F. J. Huang and Y. LeCun, "Large-scale learning with SVM and convolutional nets for generic object categorization," in *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 284–291, 2006.

[13] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?," in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, pp. 2146–2153, 2009.

[14] M. A. Ranzato, F. J. Huang, Y. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2007.

[15] M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun, "Efficient learning of sparse representations with an energy-based model," in *Proc. of International Conference on Neural Information Processing Systems (NIPS)*, pp. 1137–1144, 2006.

[16] N. Pinto, D. Cox, and J. DiCarlo, "Why is real-world visual object recognition hard," *PLoS Computational Biology*, vol. 4, no. 1, pp. 151–156, 2008.

[17] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," in *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 994–1000, 2005.

[18] J. Mutch and D. Lowe, "Multiclass object recognition with sparse, localized features," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 11–18, 2006.

[19] S. Mallat, "Group invariant scattering," *Comm. Pure Appl. Math.*, vol. 65, no. 10, pp. 1331–1398, 2012.

[20] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1872–1886, 2013.

[21] L. Sifre and S. Mallat, "Rigid-motion scattering for texture classification," *arXiv:1403.1687*, 2014.

[22] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4114–4128, 2014.

[23] J. P. Antoine, R. Murrenzi, P. Vandergheynst, and S. T. Ali, *Two-dimensional wavelets and their relatives*. Cambridge University Press, 2008.

[24] I. Selesnick, R. Baraniuk, and N. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 123–151, 2005.

[25] M. N. Do and M. Vetterli, "The contourlet transform: An efficient directional multiresolution image representation," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2091–2106, 2005.

[26] E. J. Candès, *Ridgelets: Theory and applications*. PhD thesis, Stanford University, 1998.

[27] E. J. Candès and D. L. Donoho, "Ridgelets: A key to higher-dimensional intermittency?," *Philos. Trans. R. Soc. London Ser. A*, vol. 357, no. 1760, pp. 2495–2509, 1999.

[28] E. J. Candès and D. L. Donoho, "New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities," *Comm. Pure Appl. Math.*, vol. 57, no. 2, pp. 219–266, 2004.

[29] E. J. Candès and D. L. Donoho, "Continuous curvelet transform: II. Discretization and frames," *Appl. Comput. Harmon. Anal.*, vol. 19, no. 2, pp. 198–222, 2005.

[30] P. Grohs, S. Keiper, G. Kutyniok, and M. Schaefer, "Cartoon approximation with α-curvelets," *arXiv:1404.1043*, 2014.

[31] K. Guo, G. Kutyniok, and D. Labate, "Sparse multidimensional representations using anisotropic dilation and shear operators," in *Wavelets and Splines* (G. Chen and M. J. Lai, eds.), pp. 189–201, Nashboro Press, 2006.

[32] G. Kutyniok and D. Labate, eds., *Shearlets: Multiscale analysis for multivariate data*. Birkhäuser, 2012.

[33] E. Oyallon and S. Mallat, "Deep roto-translation scattering for object classification," *arXiv:1412.8659*, 2014.

[34] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. of International Conference on Artificial Intelligence and Statistics*, pp. 315–323, 2011.

[35] V. Nair and G. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. of International Conference on Machine Learning (ICML)*, pp. 807–814, 2010.

[36] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 20, pp. 14–22, Jan. 2011.

[37] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. of International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

[38] S. T. Ali, J. P. Antoine, and J. P. Gazeau, "Continuous frames in Hilbert spaces," *Annals of Physics*, vol. 222, no. 1, pp. 1–37, 1993.

[39] G. Kaiser, *A friendly guide to wavelets*. Birkhäuser, 1994.

[40] W. Rudin, *Functional analysis*. McGraw-Hill, 2nd ed., 1991.

[41] S. Mallat, *A wavelet tour of signal processing: The sparse way*. Academic Press, 3rd ed., 2009.

[42] T. Lee, "Image representation using 2D Gabor wavelets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 10, pp. 959–971, 1996.

[43] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, 1986.

[44] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 7, pp. 710–732, 1992.

[45] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[46] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[47] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, 2010.

[48] S. Chen, C. Cowan, and P. M. Grant, "Orthogonal least squares learning algorithm for radial basis function networks," *IEEE Trans. Neural Netw.*, vol. 2, no. 2, pp. 302–309, 1991.

[49] P. P. Vaidyanathan, *Multirate systems and filter banks*. Prentice Hall, 1993.

[50] D. Ellis, Z. Zeng, and J. McDermott, "Classifying soundtracks with audio texture features," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5880–5883, 2011.

[51] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 10, no. 5, pp. 293–302, 2002.

[52] J. Lin and L. Qu, "Feature extraction based on Morlet wavelet and its application for mechanical fault diagnosis," *J. Sound Vib.*, vol. 234, no. 1, pp. 135–148, 2000.

[53] M. Unser, "Texture classification and segmentation using wavelet frames," *IEEE Trans. Image Process.*, vol. 4, no. 11, pp. 1549–1560, 1995.

[54] G. Y. Chen, T. D. Bui, and A. Krżyzak, "Rotation invariant pattern recognition using ridgelets, wavelet cycle-spinning and Fourier features," *Pattern Recognition*, vol. 38, no. 12, pp. 2314–2322, 2005.

[55] Y. L. Qiao, C. Y. Song, and C. H. Zhao, "M-band ridgelet transform based texture classification," *Pattern Recognition Letters*, vol. 31, no. 3, pp. 244–249, 2010.

[56] S. Arivazhagan, L. Ganesan, and T. S. Kumar, "Texture classification using ridgelet transform," *Pattern Recognition Letters*, vol. 27, no. 16, pp. 1875–1883, 2006.

[57] J. Ma and G. Plonka, "The curvelet transform," *IEEE Signal Process. Mag.*, vol. 27, no. 2, pp. 118–133, 2010.

[58] L. Dettori and L. Semler, "A comparison of wavelet, ridgelet, and curvelet-based texture classification algorithms in computed tomography," *Computers in Biology and Medicine*, vol. 37, no. 4, pp. 486–498, 2007.

[59] L. Grafakos, *Classical Fourier analysis*. Springer, 2nd ed., 2008.

[60] P. Vandergheynst, "Directional dyadic wavelet transforms: Design and algorithms," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 363–372, 2002.

[61] G. Kutyniok and D. Labate, "Introduction to shearlets," in *Shearlets: Multiscale analysis for multivariate data* [32], pp. 1–38.

[62] P. Grohs, "Ridgelet-type frame decompositions for Sobolev spaces related to linear transport," *J. Fourier Anal. Appl.*, vol. 18, no. 2, pp. 309–325, 2012.

[63] A. J. E. M. Janssen, "The duality condition for Weyl-Heisenberg frames," in *Gabor analysis: Theory and applications* (H. G. Feichtinger and T. Strohmer, eds.), pp. 33–84, Birkhäuser, 1998.

[64] A. Ron and Z. Shen, "Frames and stable bases for shift-invariant subspaces of $L^2(\mathbb{R}^d)$," *Canad. J. Math.*, vol. 47, no. 5, pp. 1051–1094, 1995.

[65] M. Frazier, B. Jawerth, and G. Weiss, *Littlewood-Paley theory and the study of function spaces*. American Mathematical Society, 1991.

[66] I. Daubechies, *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, 1992.

[67] A. W. Naylor and G. R. Sell, *Linear operator theory in engineering and science*. Springer, 1982.

[68] H. Bölcskei, F. Hlawatsch, and H. G. Feichtinger, "Frame-theoretic analysis of oversampled filter banks," *IEEE Trans. Signal Process.*, vol. 46, no. 12, pp. 3256–3268, 1998.

[69] A. J. E. M. Janssen, "Duality and biorthogonality for Weyl-Heisenberg frames," *J. Fourier Anal. Appl.*, vol. 1, no. 4, pp. 403–436, 1995.

[70] I. Daubechies, H. J. Landau, and Z. Landau, "Gabor time-frequency lattices and the Wexler-Raz identity," *J. Fourier Anal. Appl.*, vol. 1, no. 4, pp. 438–478, 1995.

[71] K. Gröchening, *Foundations of time-frequency analysis*. Birkhäuser, 2001.

[72] I. Daubechies, A. Grossmann, and Y. Meyer, "Painless nonorthogonal expansions," *J. Math. Phys.*, vol. 27, no. 5, pp. 1271–1283, 1986.

[73] K. Gröchenig and S. Samarah, "Nonlinear approximation with local Fourier bases," *Constr. Approx.*, vol. 16, no. 3, pp. 317–331, 2000.

[74] C. Lee, J. Shih, K. Yu, and H. Lin, "Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features," *IEEE Trans. Multimedia*, vol. 11, no. 4, pp. 670–682, 2009.

[75] G. Kutyniok and D. L. Donoho, "Microlocal analysis of the geometric separation problem," *Comm. Pure Appl. Math.*, vol. 66, no. 1, pp. 1–47, 2013.

[76] M. Searcóid, *Metric spaces*. Springer, 2007.

[77] R. P. Brent, J. H. Osborn, and W. D. Smith, "Note on best possible bounds for determinants of matrices close to the identity matrix," *Linear Algebra and its Applications*, vol. 466, pp. 21–26, 2015.

[78] W. Rudin, *Real and complex analysis*. McGraw-Hill, 2nd ed., 1983.

[79] E. DiBenedetto, *Real analysis*. Birkhäuser, 2002.