Wavelet Networks:
Scale Equivariant Learning From Raw Waveforms

David W. Romero
Vrije Universiteit Amsterdam
The Netherlands
d.w.romeroguzman@vu.nl

Erik J. Bekkers
Universiteit van Amsterdam
The Netherlands
e.j.bekkers@uva.nl

Jakub M. Tomczak
Vrije Universiteit Amsterdam
The Netherlands
j.m.tomczak@vu.nl

Mark Hoogendoorn
Vrije Universiteit Amsterdam
The Netherlands
m.hoogendoorn@vu.nl

Abstract

Inducing symmetry equivariance in deep neural architectures has resolved into improved data efficiency and generalization. In this work, we utilize the concept of scale and translation equivariance to tackle the problem of learning on time-series from raw waveforms. As a result, we obtain representations that largely resemble those of the wavelet transform at the first layer, but that evolve into much more descriptive ones as a function of depth. Our empirical results support the suitability of our *Wavelet Networks* which with a simple architecture design perform consistently better than CNNs on raw waveforms and on par with spectrogram-based methods.

1 Introduction

Symmetry rules our universe. In fact, it is due to the symmetricity of our universe that we can hope to extrapolate findings from our local region in time and space to predict the behaviour of galaxies lying billions of light-years away in the distant future. In order to construct effective and efficient statistical systems that extrapolate well to unseen data, it seems thus natural and essential to include prior knowledge about the symmetries that rule the space they act upon into their structure. A good example of this principle is the translation equivariance of the Convolutional Neural Network (CNN) [33], where the output of a translated input is equivalent to that of the original input up to an equivalent translation. In other words, it preserves the translational symmetry of physical objects encountered in the real world. Following the principle of symmetry preservation, additional approaches for computer vision tasks have been proposed to preserve additional symmetries in visual data, such as planar rotations [19, 45, 79, 76, 36, 8, 27, 4, 70, 35, 23], spherical rotations [13, 77, 75, 15, 65], scaling [46, 78, 62] and more general symmetry groups [11, 31, 74, 14, 3, 55, 71, 56]. Naturally, symmetries reach much further than only visual characteristics. Exemplarily, one can envisage the possible states of a dynamical system as symmetries of a single abstract system, which are equivariant to the total amount of energy in it. Similar intuitions have recently propelled the development of several deep learning approaches to model physical systems [54, 30, 73, 20]. These developments unavoidably lead to the question: Which symmetries are intrinsic to time-series? Zhang et al. [82] explored this direction for a particular symmetry encountered in audio: the vocal tract length. Speech recognition is known to be invariant to the vocal tract length of the speaker. Consequently, one can envisage a word pronounced by speakers with different vocal tract lengths as symmetries of that word, which are equivariant to its meaning. In their experiments, Zhang et al. [82] constructed a shallow network in which the input signal was perturbed by several vocal tract length warpings, which lead to enhanced generalization.

In this work, we take a general approach towards symmetry preservation in time-series. To this end, we start with the following question: *What is the simplest set of transformations that describes all possible symmetries underlying time-series?* The response to this question leads us to the symmetry group composed by translation and scaling, the dilation-translation group. Motivated by this observation, we propose to induce translation and scale equivariance in neural architectures for time-series, in a manner similar to Worrall and Welling [78], Sosnovik et al. [62] and [3] for visual data. Starting from this equivariance condition, we show that one unavoidably arrives at a very well-known tool of classical signal processing: the wavelet transform. Though the connection between the dilation-translation group and the wavelet transform has long been identified [24], we argue that making this connection explicit for machine learning applications could be of large value to the field, as future approaches could leverage insights from decades of research in spectrotemporal analysis, e.g., [60, 44, 17].

Our proposed theoretical framework allows for end-to-end learning on time-series in a translation and scale equivariant manner. To the best of our knowledge, we are the first to provide a general group theoretical approach for learning on time-series and second on the line of thought of symmetry preservation in this field (aside from Zhang et al. [82]). Due to the similarity of our approach with wavelet analysis, we discuss the adequacy of complementary time-frequency representations for time-series analysis from a symmetry preserving perspective. Our insights nicely connect with findings from neuroscience and psychology research on the *modus operandi* of the human auditory system. Based on these insights, we highlight important shortcomings of contemporarily widely used architectures, which are adequately addressed by our framework. As an instance of our theoretical framework, we propose the *Wavelet Networks*, a member of the group equivariant convolutional networks family which allows for end-to-end learning on raw time-series in a scale and translation equivariant manner. Our empirical results support the suitability of our approach which performs consistently better than CNNs on raw waveforms and on par with spectrogram-based methods on multiple tasks.

Contributions:

- We propose a theoretical framework for translation and scale equivariance in time-series.
- As an instance of this framework, we introduce our *Wavelet Networks*, which are equivariant to translation and scaling, and allow for end-to-end learning on raw waveform time-series.
- Our experiments support the adequacy of our Wavelet Networks for several audio applications, which perform consistently better than CNNs on raw waveforms and on par with spectrogram-based approaches based on strong hand-engineered representations (e.g., log-Mel).

2 Revisiting Classical Time-Frequency Signal Processing

Let us consider a general dictionary of atoms $\mathcal{D} = \{\phi_\gamma\}_{\gamma \in \Gamma}$, where γ might be a multi-index parameter. Linear time-frequency transforms correlate an input signal with a dictionary of waveforms concentrated in time and frequency called *time-frequency atoms*. In order to ensure invertibility and stability of the representation, the atoms ϕ_γ are assumed to have finite energy and unitary norm, i.e., $\phi_\gamma \in \mathbf{L^2}(\mathbb{R})$, $\|\phi_\gamma\|^2 = 1$. The linear time-frequency transform of a signal $f \in \mathbf{L^2}(\mathbb{R})$ is defined by:

$$\Phi[f](\gamma) = \int_{-\infty}^{+\infty} f(t)\phi_\gamma^*(t)\,\mathrm{d}t = \langle f, \phi_\gamma \rangle, \tag{1}$$

with ϕ_γ^* the complex conjugate of ϕ_γ. The prime time-frequency transform is the *Fourier transform* $\mathcal{F}$, in which f is described by a time-frequency dictionary of complex sinusoidal waves $\mathcal{D} = \{\mathrm{e}^{i\omega t}\}_{\omega \in \mathbb{R}}$:

$$\mathcal{F}[f](\omega) = \hat{f}(\omega) = \langle f, \mathrm{e}^{i\omega t} \rangle = \int_{-\infty}^{\infty} f(t)\,\mathrm{e}^{-i\omega t}\,\mathrm{d}t. \tag{2}$$

Motivated by quantum mechanics, Gabor [22] proposed to decompose signals over dictionaries of time-frequency atoms with minimal spread on the time-frequency plane. He showed that the joint time-frequency resolution of representations on the time-frequency plane is limited by a minimum surface, called the *Heisenberg rectangle*, $\sigma_{t,\gamma}\sigma_{\omega,\gamma} \geq \frac{1}{2}$, with $\sigma_{t,\gamma}$ and $\sigma_{\omega,\gamma}$ depicting the spread of a time-frequency atom ϕ_γ on time and frequency, respectively. As a result, it is not possible to construct time-frequency representations with arbitrary resolutions on time and frequency simultaneously (Fig. 1). This phenomenon is referred to as the *Heisenberg uncertainty principle* and plays a crucial role in the design of time-frequency representations. The two extremes of this resolution schema are given by signals $f(t)$ and their Fourier transform $\hat{f}(\omega)$, which behold exact time and frequency resolution, respectively, but *"null"* resolution in the complementary dimension.
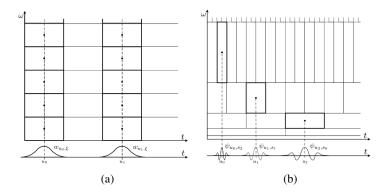
Figure 1: The tiling of the time-frequency plane for the short-time Fourier transform (Fig. 1a) and the wavelet transform (Fig. 1b). By modifying the form of the tiles (Heisenberg boxes), the wavelet transform is able to detect highly localized events with great precision both on time and frequency.

(a) (b)

2.1 Translation Equivariant Dictionaries

When analyzing structured data, it is advantageous to construct signal representations that commute with translations. That is, representations in which if a pattern is translated, its numerical descriptors are also translated, but not modified.[1] Formally, a representation is said to be *translation equivariant* if, for any $\phi_\gamma(t) \in \mathcal{D}$ and any shift $u \in \mathbb{R}$, the atom $\phi_\gamma(t + u)$ also belongs to $\mathcal{D}$. Such a dictionary is constructed by translating a family of generators $\{\phi_\gamma\}_{\gamma \in \Gamma}$ to form a dictionary $\mathcal{D} = \{\phi_{\gamma,u}\}_{\gamma \in \Gamma, u \in \mathbb{R}}$, with $\phi_{\gamma,u}(t) = \mathcal{L}_u[\phi_\gamma](t)$, and $\mathcal{L}_u[\phi_\gamma](t) := \phi_\gamma(t - u)$ a translation operator. Resultantly, the time-frequency transform of f, $\Phi[f](\gamma, u)$, can be expressed as a convolution:

$$\Phi[f](\gamma, u) = \langle f, \phi_{\gamma,u} \rangle = \int_{-\infty}^{+\infty} f(t) \mathcal{L}_u[\phi_\gamma^*](t) \, \mathrm{d}t = \int_{-\infty}^{+\infty} f(t) \phi_\gamma^*(t - u) \, \mathrm{d}t = [f \star \phi_\gamma](u). \quad (3)$$

Consequently, the translation equivariance condition can be formally described as:

$$[\mathcal{L}_u[f] \star \phi_\gamma](t) = \mathcal{L}_u[f \star \phi_\gamma](t). \quad (4)$$

In fact, being able to write $\Phi[f]$ as a convolution (Eq. 3) is both *necessary and sufficient*, as convolution (and reparametrizations thereof) is the *only* linear translation equivariant linear mapping [31, 14, 3].

The windowed Fourier transform. Gabor [22] introduced windowed Fourier atoms to measure frequency variations of sounds. To this end, a real, symmetric window $w(t) = w(-t)$ of local support and unitary norm is translated by u and modulated by the frequency ξ: $w_{u,\xi}(t) = e^{i\xi t} w(t - u)$. The windowed Fourier transform (also called *short-time Fourier transform*) $\mathcal{S}$ of a signal $f \in \mathbf{L}^2(\mathbb{R})$ is:

$$\mathcal{S}[f](u, \xi) = \langle f, w_{u,\xi} \rangle = \int_{-\infty}^{+\infty} f(t) \mathcal{L}_u[w](t) \, e^{-i\xi t} \, \mathrm{d}t = \int_{-\infty}^{+\infty} f(t) w(t - u) \, e^{-i\xi t} \, \mathrm{d}t. \quad (5)$$

Intuitively, the windowed Fourier transform localizes the Fourier integral (Eq. 2) in the neighbourhood of $t = u$. The windowed Fourier transform decomposes signals over waveforms that have the same time and frequency resolution. Consequently, it is effective as long as the signal f does not include structures having different time-frequency resolutions, some being very localized in time and other very localized in frequency. Unfortunately, this is the case for the vast majority of signals encountered in nature. Wavelets address this issue by changing the time and frequency resolution of their Heisenberg boxes at different scales (Fig. 1b).

Definition 2.1. *Wavelet.* A wavelet is a function $\psi \in \mathbf{L}^2(\mathbb{R})$ of compact support centered in the neighbourhood of $t = 0$, with unitary norm, $\|\psi\|^2 = 1$, and zero average, $\int_{-\infty}^{+\infty} \psi(t) \, \mathrm{d}t = 0$.

The Wavelet transform. Additional to the analysis of signal structures at different positions, it is advantageous to construct signal representations that allow for their analysis at different scales as well. To this end, it is necessary to construct time-frequency atoms with varying time supports. The *wavelet transform* accomplishes this by decomposing an input signal over dilated and translated wavelets. The dictionary of wavelet time-frequency atoms is obtained by translating and scaling a wavelet ψ as:

$$\mathcal{D} = \left\{ \psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t - u}{s}\right) \right\}_{u \in \mathbb{R}, s \in \mathbb{R}^+} = \left\{ \psi_{u,s}(t) = \frac{1}{\sqrt{s}} \mathcal{L}_u \mathcal{L}_s[\psi](t) \right\}_{u \in \mathbb{R}, s \in \mathbb{R}^+}, \quad (6)$$

---

[1]In classical signal processing, this is referred to as *translation invariance*, as the descriptors are not modified. We emphasize that these operators are rather *equivariant*, as they preserve the translation through the mapping.

with $\mathcal{L}_s[\psi](t) := \psi(s^{-1}t)$ a scaling operator. The wavelet $\psi_{0,1}(t) = \psi(t)$ from which the dictionary is constructed is called the *mother wavelet* and the *normalization coefficient* $(\sqrt{s})^{-1}$ is utilized to ensure that $\|\psi_{u,s}\|^2 = 1, \forall s \in \mathbb{R}^+$. The *Wavelet transform* $\mathcal{W}$ of a signal $f \in \mathbf{L}^2(\mathbb{R})$ is given by:

$$\mathcal{W}[f](u,s) = \langle f, \psi_{u,s} \rangle = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{s}} \mathcal{L}_u \mathcal{L}_s[\psi^*](t) \, \mathrm{d}t = \left[f \star \frac{1}{\sqrt{s}} \mathcal{L}_s[\psi^*] \right](u). \qquad (7)$$

Resultantly, $\mathcal{W}[f](u,s)$ measures the variation of f in a neighbourhood of u proportional to s. The wavelet transform is both *translation and scale equivariant*. This results from the fact that given a wavelet $\psi_{u,s} \in \mathcal{D}$ and arbitrary translations $\tilde{u} \in \mathbb{R}$ and scales $\tilde{s} \in \mathbb{R}^+$, the transformed wavelet $\mathcal{L}_{\tilde{u}}\mathcal{L}_{\tilde{s}}[\psi_{u,s}]$ remains a member of $\mathcal{D}$ (up to a multiplicative factor $\sqrt{\tilde{s}}$, c.f., Eq. 14):

$$\mathcal{L}_{\tilde{u}}\mathcal{L}_{\tilde{s}}[\psi_{u,s}](t) = \mathcal{L}_{\tilde{u}}\mathcal{L}_{\tilde{s}}\left[\frac{1}{\sqrt{s}} \mathcal{L}_u \mathcal{L}_s[\psi] \right](t) = \frac{1}{\sqrt{s}} \mathcal{L}_{\tilde{u}+\tilde{s}u}\mathcal{L}_{\tilde{s}s}[\psi](t) = \sqrt{\tilde{s}} \, \psi_{\tilde{u}+\tilde{s}u, \tilde{s}s}(t) \in \mathcal{D} \qquad (8)$$

In other words, if a pattern is either translated or scaled, its numerical descriptors are also translated or scaled, but not modified.

Spectrograms and scaleograms. Classical signal processing focuses on detecting the presence of frequency components of an input signal f. To this end, an energy density function is often defined on top of the time-frequency transform as $|\Phi[f](\gamma, u)|^2$. This density function, referred to as *spectrogram* $(|\mathcal{S}[f](u,\xi)|^2)$ or *scalogram* $(|\mathcal{W}[f](u,s)|^2)$ for the short-time Fourier and wavelet transform, respectively, measures the energy of f in a time-frequency neighbourhood determined by the spread of the time-frequency atom at a given location. Therefore, it is crucial to consider the role of input variations in position and scale on $|\cdot|^2$ as well. The effects of these variations on the time-frequency transforms outlined in this section are summarized in Thm. 2.1 and analyzed in detail in Appx. B.

Theorem 2.1. *Let $\mathcal{L}_{t_0}[f](t) = f(t-t_0)$ and $\mathcal{L}_{s_0}[f](t) = f(s_0^{-1}t)$, $t_0 \in \mathbb{R}$, $s_0 \in \mathbb{R}^+$, be a translated and scaled version of an arbitrary input signal $f \in \mathbf{L}^2(\mathbb{R})$, respectively. The relation among their time-frequency transforms is given by:*

- *Fourier Transform:*

$$\mathcal{F}[\mathcal{L}_{t_0}[f]](\omega) = \mathrm{e}^{-i\omega t_0}\mathcal{F}[f](\omega) \qquad \rightarrow |\mathcal{F}[\mathcal{L}_{t_0}[f]](\omega)|^2 = |\mathcal{F}[f](\omega)|^2 \qquad (9)$$

$$\mathcal{F}[\mathcal{L}_{s_0}[f]](\omega) = s_0 \mathcal{L}_{s_0^{-1}}[\mathcal{F}[f]](\omega) \qquad \rightarrow |\mathcal{F}[\mathcal{L}_{s_0}[f]](\omega)|^2 = |s_0|^2 |\mathcal{L}_{s_0^{-1}}[\mathcal{F}[f]](\omega)|^2 \qquad (10)$$

- *Short-Time Fourier Transform:*

$$\mathcal{S}[\mathcal{L}_{t_0}[f]](u,\xi) = \mathrm{e}^{-i\xi t_0}\mathcal{L}_{t_0}[\mathcal{S}[f]](u,\xi) \rightarrow |\mathcal{S}[\mathcal{L}_{t_0}[f]](u,\xi)|^2 = |\mathcal{L}_{t_0}[\mathcal{S}[f]](u,\xi)|^2 \qquad (11)$$

$$\mathcal{S}[\mathcal{L}_{s_0}[f]](u,\xi) \approx s_0 \, \mathcal{S}[f](s_0^{-1}u, s_0\xi) \quad \rightarrow |\mathcal{S}[\mathcal{L}_{s_0}[f]](u,\xi)|^2 \approx |s_0|^2 |\mathcal{S}[f](s_0^{-1}u, s_0\xi)|^2 \;^{(*)} \quad (12)$$

- *Wavelet Transform:*

$$\mathcal{W}[\mathcal{L}_{t_0}[f]](u,s) = \mathcal{L}_{t_0}[\mathcal{W}[f]](u,s) \qquad \rightarrow |\mathcal{W}[\mathcal{L}_{t_0}[f]](u,s)|^2 = |\mathcal{L}_{t_0}[\mathcal{W}[f]](u,s)|^2 \qquad (13)$$

$$\mathcal{W}[\mathcal{L}_{s_0}[f]](u,s) = \sqrt{s_0} \, \mathcal{L}_{s_0}[\mathcal{W}[f]](u,s) \quad \rightarrow |\mathcal{W}[\mathcal{L}_{s_0}[f]](u,s)|^2 = |\mathcal{L}_{s_0}[\mathcal{W}[f]](u,s)|^2 \qquad (14)$$

$^{(*)}$ *Eq. 12 only approximately holds for large windows (see Appx. B.2 for a detailed explanation).*

Proof. The derivation and interpretation of the aforementioned properties are provided in Appx. B.

3 The Problem of Learning 2D Conv. Filters on the Time-Frequency Plane

CNNs have been a major breakthrough in computer vision and have achieved startling results in numerous applications. Due to its success and the resemblance of time-frequency representations with image data, CNNs have been extensively utilized on top of spectrograms, e.g., [37, 61, 2], and (scarcely) scalograms, e.g., [52], for time-series learning. Unfortunately however, the underlying behaviour of images and time-frequency representations are of different natures and largely differ from one another. In fact, it has been shown that *Deep Priors* [68], which naturally emerge in CNNs for visual data, do not appear for audio [83]. These observations suggest that treating spectrograms as images and *directly* learning 2D convolutional filters on top of them is not adequate for proper time-series learning.

Differences between visual data and time-frequency representations. Two important differences between time-frequency representations and visual data exist that are universal to all time-frequency representations. Contrarily to images, sounds are *non-local* on these representations. Auditory signals are often composed of harmonic components that resonate at non-local frequencies called the *signal*

harmonics. As a result, contrarily to visual objects, the representation of an *auditory object* is able to (and often does) span a large part of the frequency axis in an sparse manner. Moreover, when considering complex auditory signals composed of several such auditory objects, their components often *superpose* at particular locations. This phenomenon is orthogonal to visual data, where information at a given location can be safely considered to belong to one exclusive object. When objects overlap, they occlude one another and the value at a certain position belongs to a single object. On the other hand, rather than occlude, audio signals *superpose*. This is why audio is considered to be *transparent*.

Analysis of the windowed Fourier transform. Research in psychology and neuroscience has shown that humans largely rely in the transient behaviour of auditory signals to distinguish auditory objects [9, 69, 6, 47]. The windowed Fourier transform, however, is not appropriate to handle transient signals. Furthermore, the time-frequency resolution of the windowed Fourier transform exhibits linear resolution on both axes, whereas the human auditory system possesses high spectral resolution at low-frequencies and high temporal resolution at higher frequencies [63, 59, 5]. For example, a difference of a semitone at the bottom of the piano scale ($\sim$30Hz) is of approximately 1.5Hz, whilst at the top of the musical scale ($\sim$5kHz) it is of approximately 200Hz. In fact, the spectrotemporal resolution of the human auditory system has motivated the development of improved audio representations on top of the windowed Fourier transform, e.g., log-Mel [63, 21] or constant-Q [7] spectra, and has lead to the development of several machine learning approaches, some of which are among the contemporarily most used, e.g., [67, 10, 38, 80]. These biologically inspired representations are incomplete, however, in the sense that they are not able to modify the temporal resolution of the underlying windowed Fourier transform. Interestingly, the wavelet transform directly and accurately describes the behaviour of the human auditory system (Fig. 1b). These observations suggest that, from a biological perspective, the wavelet transform might be better suited to represent auditory signals than the windowed Fourier transform and other signal representations building on top of them.

Analysis of the wavelet transform. Despite the previous observations, directly learning 2D convolutional filters on top of the wavelet transform does not resolve the issues related to the non-local, transparent nature of auditory signals outlined before. Though several approaches have been proposed to alleviate these problems, e.g., by defining filters than span large receptive fields along the frequency or time axis [49], or focusing learning on harmonic multiples of a particular frequency [83], a theoretically principled approach to this end is, to the best of our knowledge, yet missing.

Learning from raw waveforms. Several end-to-end learning approaches for time-series have recently emerged, e.g., [18, 19, 16, 53, 64]. In these works, the main research focus lies in how to obtain large parameter efficient receptive fields, which can easily span as long as 22.050 samples for a single second. The go-to solution relies in the usage of several *á-trous* convolution layers, with increasing dilation as a function of depth. In our work, by making use of multi-scale analysis similar to that of the wavelet transform at every layer, we obtain a virtual extremely large receptive field at every single layer of the network, which allows us to reason about information at several scales in one go (Fig. 2). Though approaches relying on layer-wise multi-scale representations exist [84, 43, 72, 25, 83], they are unable to preserve the same sense of scale across layers as they do not commute with scaling. Our theoretically principled, symmetry preserving approach alleviates this problem by utilizing group convolutions equivariant to translation and scaling directly from raw waveforms, and nicely lines up with several years of research in wavelet analysis, neuroscience and psychology [6, 44, 17, 41].

4 Wavelet Networks

In this section, we formalize our approach. Since it largely relies on the idea of symmetries, groups and group convolutions, we provide all the concepts required for proper understanding in Appx. A.

4.1 The Group Convolution and the Dilation-Translation Group

Let $f, \psi : G \to \mathbb{R}$ be a scalar signal and filter defined on a group G (Def. A.1) and $\mathcal{L}_g[\cdot]$ be the left-regular representation of G (Def. A.3). The group convolution ($\star_G$) is defined as:

$$[f \star_G \psi](g) = \langle f, \psi_g \rangle = \int_G f(\tilde{g})\mathcal{L}_g[\psi](\tilde{g}) \, \mathrm{d}\tilde{g} = \int_G f(\tilde{g})\psi(g^{-1}\tilde{g}) \, \mathrm{d}\tilde{g}. \tag{15}$$

This is a direct generalization of the convolution (Eq. 3) for domains defined by groups. A key property of the group convolution is that it generalizes equivariance to arbitrary symmetry transformations

The lifting group convolution:
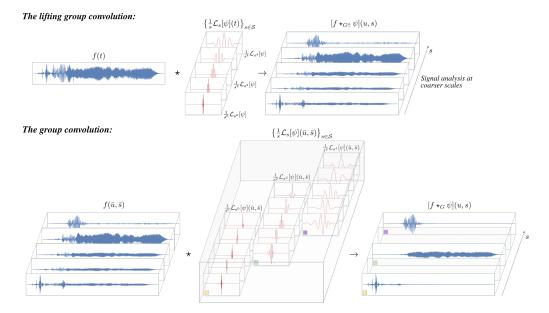
The group convolution:

Figure 2: The wavelet network structure. Wavelet networks are comprised of a starting lifting group convolution (Eq. 19), after which an arbitrary number of group convolutions (Eq. 17) can be concatenated. Their structure allows for multi-scale signal analysis at every layer of the network.

$g \in G$. Consequently, the equivariance constraint (Eq. 4) becomes:

$$[\mathcal{L}_{\bar{g}}[f] \star_G \psi](g) = \mathcal{L}_{\bar{g}}[f \star_G \psi](g), \quad \forall \bar{g}, g \in G. \tag{16}$$

Analogously, it holds that the group convolution is the *only* linear G-equivariant mapping [31, 14, 3].

Group convolution on affine groups. If G is affine (Def. A.2), the group convolution can be split by taking advantage of the group structure (Eq. 24) and the representation decomposition (Eq. 26) as:

$$[f \star_G \psi](g) = \langle f, \psi_g \rangle = \int_G f(\tilde{g}) \mathcal{L}_g[\psi](\tilde{g}) \, \mathrm{d}\tilde{g} = \int_H \int_{\mathbb{R}^d} f(\tilde{x}, \tilde{h}) \mathcal{L}_x \mathcal{L}_h[\psi](\tilde{x}, \tilde{h}) \frac{1}{|h|} \, \mathrm{d}\tilde{x} \, \mathrm{d}\tilde{h}, \tag{17}$$

where $g = (x, h)$, $\tilde{g} = (\tilde{x}, \tilde{h}) \in G$, $x, \tilde{x} \in \mathbb{R}^d$, $h, \tilde{h} \in H$, and $G = \mathbb{R}^d \rtimes H$. Here $|h|$ denotes the Jacobian of the action of h on $\mathbb{R}^d$, which appears as a normalizing factor when writing the Haar measure $\mathrm{d}\tilde{g} = \frac{1}{|h|} \mathrm{d}\tilde{x} \mathrm{d}\tilde{h}$ on G in terms of the spatial (Lebesgue) measure $\mathrm{d}\tilde{x}$ on $\mathbb{R}^d$ and the Haar measure $\mathrm{d}\tilde{h}$ on H. It then becomes apparent that the group convolution can be split into $|H|$ spatial convolutions of the input signal f and h-transformed filters $\mathcal{L}_h[\psi]$:

$$[f \star_G \psi](x, h) = \langle f, \frac{1}{|h|} \psi_{x,h} \rangle = \int_H \left[f \star \frac{1}{|h|} \mathcal{L}_h[\psi] \right](x, \tilde{h}) \, \mathrm{d}\tilde{h}. \tag{18}$$

In our work, we are interested in the dilation-translation group, which emerges from the semi-direct product ($\rtimes$) of the translation group $(\mathbb{R}^d, +)$ and the dilation group $H = (\mathbb{R}^+, \times)$ acting on $\mathbb{R}^d$.

The lifting group convolution. Up to now, we have considered convolutions between functions defined on a group G. However, predictive systems usually receive functions f defined in $\mathbb{R}^d$. As a result, it is crucial to rely on an operator to *lift* a function from $\mathbb{R}^d$ to G. The continuous wavelet transform is such an operator, which, in a group convolutional setting, can be regarded as a lifting group convolution (see Sec. 4.2). Let $f, \psi : \mathbb{R}^d \to \mathbb{R}$ be a scalar signal and filter function on $\mathbb{R}^d$, and $G = \mathbb{R}^d \rtimes H$ be an affine group. The lifting convolution ($\star_{G\uparrow}$) is defined as:

$$[f \star_{G\uparrow} \psi](x, h) = \langle f, \frac{1}{|h|} \psi_{x,h} \rangle = \int_{\mathbb{R}^d} f(\tilde{x}) \frac{1}{|h|} \mathcal{L}_x \mathcal{L}_h[\psi](\tilde{x}) \, \mathrm{d}\tilde{x} = \left\{ \left[f \star \frac{1}{|h|} \mathcal{L}_h[\psi] \right] \right\}_{h \in H} (x, h), \tag{19}$$

where $\frac{1}{|h|}$ is required to guarantee equivariance (Eq. 16) for non-unimodular affine groups such as $G = \mathbb{R}^d \rtimes \mathbb{R}^+$ [3]. The lifting convolution corresponds to the set obtained by spatially convolving the signal f with all h-transformed versions of the filter ψ, $\mathcal{L}_h[\psi]$, for all transformations in the group H.

Projecting functions from G to $\mathbb{R}^d$. Predictions must often be given in the same space as the input of the predictive system, $\mathbb{R}^d$, e.g., for segmentation tasks. To this end, one can simply project the

function from G to $\mathbb{R}^d$ by means of a pooling operation over H as:

$$\text{proj}_{\mathbb{R}^d} f = \text{pool}_{\tilde{h} \in H} f(x, \tilde{h}), \tag{20}$$

where the operation pool can be given by any pooling operation such as max or mean pooling.

4.2 The Wavelet Transform as Group Convolutions

The connection between the wavelet transform and group representations has long been identified. In one of the initial papers of wavelet analysis, Grossmann et al. [24] demonstrated that the continuous wavelet transform on $\mathbf{L}^2(\mathbb{R})$ as well as its inversion formula lie on a certain representation of the dilation-translation group acting on $\mathbb{R}$, $\mathbb{R} \rtimes \mathbb{R}^+$. In fact, the wavelet transform of a signal f (Eq. 7) is equivalent, up to a front factor, to a lifting convolution (Eq. 19) from $\mathbb{R}$ to $\mathbb{R} \rtimes \mathbb{R}^+$:

$$\mathcal{W}[f](u,s) = \left[f \star \tfrac{1}{\sqrt{s}} \mathcal{L}_s[\psi] \right](u) = \sqrt{s} \left[f \star \tfrac{1}{s} \mathcal{L}_s[\psi] \right](u) = \sqrt{s}\,[f \star_{G\uparrow} \psi](u,s), \tag{21}$$

with $u \in \mathbb{R}$, $s \in \mathbb{R}^+$. As outlined in Thm. 2.1 (Eq. 14) and Appx. B.3, the multiplicative factor $\sqrt{s}$ vanishes during the scalogram computation. This elucidates the fact that time-frequency transforms are optimized to represent energy density functions on the time-frequency plane (see Sec. 2, Appx. B).

In Appx. C we derive conditions on the parametrization of arbitrary convolutional filters ψ, such that their lifting convolution ($\psi : \mathbb{R}^d \to \mathbb{R}$) and group convolution ($\psi : \mathbb{R}^d \rtimes \mathbb{R}^+ \to \mathbb{R}$) with arbitrary input signals is exactly equivariant to scale and translation. We show that these filters are equivalent to wavelets up to a re-normalization factor. These properties are summarized in Thm. 4.1.

Theorem 4.1. *Let $G = \mathbb{R}^d \rtimes \mathbb{R}^+$ be the dilation-translation group and $\mathcal{L}_g[\cdot]$ its left regular representation. Let $f, \psi : \mathbb{R}^d \to \mathbb{R}$ be a scalar signal and filter defined on $\mathbb{R}^d$. A lifting convolution (Eq. 19) from $\mathbb{R}^d$ to $\mathbb{R}^d \rtimes \mathbb{R}^+$ is a transformation of the form $\langle f, \psi_{u,s} \rangle$ that satisfies the equivariance condition (Eq. 16) iff the set of transformed filters corresponds to a dictionary of re-normalized wavelet time-frequency atoms $\mathcal{D}$ given by:*

$$\mathcal{D} = \left\{ \psi_{u,s}(x) = \tfrac{1}{s^d} \mathcal{L}_u \mathcal{L}_s[\psi](x) \right\}_{u \in \mathbb{R}^d, s \in \mathbb{R}^+} = \left\{ \tfrac{1}{s^d} \psi\left(\tfrac{x-u}{s} \right) \right\}_{u \in \mathbb{R}^d, s \in \mathbb{R}^+}. \tag{22}$$

Let now $f(x, \bar{s})$, $\psi(x, \bar{s}) : \mathbb{R}^d \rtimes \mathbb{R}^+ \to \mathbb{R}$ be a scalar valued signal and filter defined on the dilation-translation group. The group convolution (Eq. 17) on $\mathbb{R}^d \rtimes \mathbb{R}^+$ is a transformation of the form $\langle f, \psi_{u,s} \rangle$, with $\langle f, \psi \rangle = \int_{\mathbb{R}^d} \int_{\mathbb{R}^+} f(x,s)\psi(x,s)\mathrm{d}x\mathrm{d}s$ the standard $\mathbf{L}^2$ inner product on $\mathbb{R}^d \times \mathbb{R}^+$, and it exactly satisfies the equivariance condition (Eq. 16) iff the set of filters corresponds to:[2]

$$\mathcal{D} = \left\{ \psi_{u,s}(x, \bar{s}) = \tfrac{1}{s^{d+1}} \mathcal{L}_u \mathcal{L}_s[\psi](x, \bar{s}) \right\}_{u \in \mathbb{R}^d, s \in \mathbb{R}^+} = \left\{ \tfrac{1}{s^{d+1}} \psi\left(\tfrac{x-u}{s}, s^{-1}\bar{s} \right) \right\}_{u \in \mathbb{R}^d, s \in \mathbb{R}^+}. \tag{23}$$

Proof. The derivation of the aforementioned conditions is provided in Appx. C.

4.3 Constructing Wavelet Networks

The input f of the predictive system is usually defined on $\mathbb{R}$. Consequently, wavelet networks (Fig. 2) are constructed starting with a layer of lifting group convolutions (Eq. 19), after which an arbitrary number of group convolution layers (Eq. 17) can be concatenated. In the last layer, the group function is projected back to $\mathbb{R}$ by means of a projective layer (Eq. 20). Though our derivations have been provided for scalar continuous functions, in practice, computations are performed on functions defined on a discrete grid. To this end, based on previous work [44, 39, 41, 78], we approximate the scale axis by means of a *dyadic dilation set* $\{s = 2^j\}_{j=j_{\min}}^{j_{\max}}$ with minimal and maximal scales defined by $2^{j_{\min}}$ and $2^{j_{\max}}$, respectively. Despite only using scalar functions in our derivations, our theory is also valid for vector-valued signals $f : \mathcal{X} \to \mathbb{R}^{N_c}$, $f = \{f_c\}_{c=1}^{N_c}$ on $\mathcal{X}$, e.g., stereo signals. In this case, a filter ψ of corresponding dimensionality is defined on the same domain $\mathcal{X}$, i.e., $\psi : \mathcal{X} \to \mathbb{R}^{N_c}$, $\psi = \{\psi_c\}_{c=1}^{N_c}$, and an additional sum along channels c is performed during the (group) convolution.

The usage of continuous smooth bases has proven advantageous for group convolutions [12, 76, 4, 75, 74, 13, 62, 3, 23]. This is due to the fact that the representation of the group H often imposes transformations not well-defined for discrete bases. As a result, interpolation is often required, which can introduce severe spurious artifacts, specially for small filters. The physiology of the

---

[2]On a practical note we remark that numerical integration over $\mathbb{R}^+$ with an exponential grid, e.g., a dyadic grid, leads to an effective front-factor of $\frac{1}{s^d}$. See Appx. D for details.

human auditory system has strongly motivated the usage of gammatone filters for auditory processing [28, 26, 40]. Based on the resemblance of gammatone filters with B^2-splines, we utilize the B-splines of Bekkers [3] as the basis of our filters. As a result, we not only avoid interpolation altogether, but allow for arbitrary approximations of the scale axis not only restricted to dyadic sets.

Classically, the filters ψ used in a wavelet transform are normalized to have unit norm and zero average (Def. 2.1). These constraints are required to preserve (stable) invertibility of the transform, which, however, is not necessary for most machine learning applications. Therefore, we release the $\|\psi\| = 1$ constraint, but still experiment with the zero-mean constraint as this ensures that the filters act as moving band-pass filters in the spectral domain with increasing scale [44]. We softly enforce this by adding the square of the average of each filter as an extra term to the training loss (WL).

5   Experiments

We validate our approach by comparing the performance of our wavelet networks (W-Nets) to convolutional baselines working of raw waveforms for a variety of tasks. For all our experiments, we replicate as close as possible the training regime of the corresponding baselines and utilize their implementation as baseline for the construction of our networks whenever possible. An extended version of this section encompassing detailed explanations of training regimes, utilized architectures and extended results is provided in Appx. D.[3]

UrbanSound8K. The UrbanSound8K (US8K) dataset [58] consists of 8732 audio clips of 4 seconds or less, with a total of 9.7 hours of audio uniformly drawn from 10 environmental sound classes. We compare the Mn-Nets of Dai et al. [16] and the 1DCNNs of Abdoli et al. [1] with similar W-Nets in terms of number of layers and parameters. Contrarily to Dai et al. [16] we sample audio at 22.05kHz as opposed to 8kHz, as early studies in the data indicated that some classes become indistinguishable for the human ear in this regime, e.g., drilling. We use the 50999-1DCNN [1] in our experiments, as it requires the less human engineering. Unfortunately, we were not able to replicate the results reported in [1] ($83 \pm 1{,}3\%$) in our experiments. In order to compare our results with other approaches in literature, we utilize 10-fold cross-validation by taking 8 subsets for training, one for validation and one for test. We take the $(n - 1) \bmod 10$ subset for validation when testing on the nth one. We emphasise that our training regime can be different from those used on other works, as they often do not disclose which subset is used for validation. Our results show that our wavelet networks consistently outperform the baselines and perform competitively to spectrogram-based approaches (Tab. 1). Furthermore approach benefits from encouraging zero-mean filters, which verifies the benefit of augmenting group convolutional networks to promote their similarity to the classical wavelet transform.

MagnaTagATune. The MagnaTagATune (MTAT) dataset [32] consists of 25879 audio clips with a total of 170 hours of audio, along with several per-song tags. Following Lee et al. [34], we extract the most frequently used 50 tags and trim the audios to 29.1 seconds at a sample-rate of 22.05kHz. Following the convention in literature, we use ROC-curve (AUC) and mean average precision (MAP) as performance metrics. We compare the (best-performing) 3^9-Net of Lee et al. [34], with a similar W3^9-Net in terms of number of layers and parameters. Our results show that our W3^9-Net consistently outperforms the baseline and performs competitively to spectrogram-based approaches (Tab. 1).

Discussion. Our results demonstrate that our proposed *Wavelet Networks* are a promising direction for learning on time-series. We demonstrate that symmetry preservation and equivariance to more general groups than translations are strong inductive biases outside computer vision tasks as well. The biggest shortcoming of our approach is related to the increment of memory requirements, which grows linearly as a function of the number of scales considered. In future work, we want to explore how to reduce these computational costs. Furthermore, we would like to replace group convolutions by their attentive counterparts [55, 56] to construct spectrotemporal attention maps indicating the importance of particular time-frequency atoms for a given task. Additionally, we would like to investigate the effect of defining and learning our filters in a complex-valued basis, as is usually done in classical signal processing, and evaluate the adequacy of W-Nets for tasks that require invertibility.

---

[3]Our code is publicly available at `https://github.com/dwromero/wavelet_networks`

Table 1: Experimental Results. Bold / underlined fonts mark the best per-section / absolute results.

URBANSOUND8K			
MODEL	ACC. (%) [34]	(%) 10-FOLD	PARAM.
M3-NET	54.48	-	220.67K
W3-NET	**63.08**	-	
W3-NET-WL	61.05	-	219.45K
M5-NET	69.89	-	558.08K
W5-NET	**74.55**	-	
W5-NET-WL	72.28	-	558.03K
M11-NET	74.43	-	1.784M
W11-NET	79.33	66.97 ± 5.178	
W11-NET-WL	<u>80.41</u>	<u>68.47 ± 4.914</u>	1.806M
M18-NET	69.65	-	3.680M
W18-NET	75.87	64.02 ± 4.645	
W18-NET-WL	**78.26**	**65.01 ± 5.431**	3.759M
M34-NET	75.15	-	3.978M
W34-NET	76.22	65.69 ± 5.780	
W34-NET-WL	**78.38**	**66.77 ± 4.771**	4.021M
1DCNN	-	62.00 ± 6.791	453.42K
W-1DCNN	-	62.47 ± 4.925	
W-1DCNN-WL	-	**62.64 ± 4.979**	458.61K

URBANSOUND8K - COMPARISON WITH OTHER APPROACHES			
MODEL	TYPE	(%) 10-FOLD	PARAM.
W11-NET-WL	RAW	68.47 ± 4.914	1.806M
PICZAKCNN [48]		73.7 / 73.1	26M
SB-CNN [57]	SPECT.	73	241K
VGG [49]		70.74	77M
ENVNET-V2 [66]	RAW (BAGGING)	<u>78</u>	101M

MAGNATAGATUNE					
MODEL	AVERAGE AUC		MAP		PARAM.
	PER-CLASS	PER-CLIP	PER-CLASS	PER-CLIP	
3^9-NET	0.893	0.936	0.385	0.700	2.394M
W3^9-NET	0.895	0.941	0.397	0.719	
W3^9-NET-WL	**0.899**	**0.943**	**0.404**	<u>**0.723**</u>	2.404M
PCNN [42]	0.9013	**0.9365**	<u>0.4267</u>	0.6902	-
RAW [50]*	0.8905	-	0.3492	-	11.8M
SPECT. [50]*	<u>0.9040</u>	-	0.3811	-	5M
SPECT. [51]	0.893	-	-	-	191K

* Reported results are obtained in a more difficult version of this dataset.

Broader Impact

Our work brings together intuitions and results from several years of research in time-frequency representations, group theory, neuroscience and psychology, to present a theoretically strong approach for learning in raw time-series. We believe this can impact two different types of stakeholders. First of all, scientists in the specific area of deep learning can exploit and build on top of our method and theory to further the field of deep learning for time-series data. Secondly, companies can utilize our novel technique in more practical settings, and can benefit from more accurate and insightful results. Examples of such practical settings include (but are not limited to) speech recognition, predictive maintenance based on sensory data, the analysis of sensory data from wearables, and also learning on time series collected through medical devices.

While we feel these approaches can bring great benefits, it is obvious that boosting the performance of automated systems through a novel algorithm can make automated systems more competitive with humans or even surpass their performance. Potentially, this can result in human labor being replaced with automation.

In case our system would fail, the impact largely depends on the application domain at hand. Certainly for more sensitive application domains such as the medical domain great caution is needed before using the approach for actual decision making as mistakes could have huge consequences.

Finally, the system could potentially be sensitive to bias in data and might exploit these biases to improve its learning performance. Therefore, great care is needed in constructing proper datasets before the application of our newly proposed technique to avoid the risk of such biases playing a pivotal role in the application at hand.

Acknowledgments and Disclosure of Funding

References

[1] Sajjad Abdoli, Patrick Cardinal, and Alessandro Lameiras Koerich. End-to-end environmental sound classification using a 1d convolutional neural network. *Expert Systems with Applications*, 136:252–263, 2019.

[2] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 international conference on platform technology and service (PlatCon)*, pages 1–5. IEEE, 2017.

[3] Erik J Bekkers. B-spline {cnn}s on lie groups. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1gBhkBFDH.

[4] Erik J Bekkers, Maxime W Lafarge, Mitko Veta, Koen AJ Eppenhof, Josien PW Pluim, and Remco Duits. Roto-translation covariant convolutional networks for medical image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 440–448. Springer, 2018.

[5] Gavin M Bidelman and Ameenuddin Syed Khaja. Spectrotemporal resolution tradeoff in auditory processing as revealed by human auditory brainstem responses and psychophysical indices. *Neuroscience letters*, 572:53–57, 2014.

[6] Albert S Bregman. Auditory scene analysis: The perceptual organization of sound. 1990.

[7] Judith C Brown. Calculation of a constant q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, 1991.

[8] Xiuyuan Cheng, Qiang Qiu, Robert Calderbank, and Guillermo Sapiro. Rotdcf: Decomposition of convolutional filters for rotation-equivariant deep networks. *arXiv preprint arXiv:1805.06846*, 2018.

[9] E Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979, 1953.

[10] Keunwoo Choi, George Fazekas, and Mark Sandler. Automatic tagging using deep convolutional neural networks. *arXiv preprint arXiv:1606.00298*, 2016.

[11] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999, 2016.

[12] Taco S Cohen and Max Welling. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016.

[13] Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *CoRR*, abs/1801.10130, 2018. URL http://arxiv.org/abs/1801.10130.

[14] Taco S Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant cnns on homogeneous spaces. In *Advances in Neural Information Processing Systems*, pages 9142–9153, 2019.

[15] Taco S Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral cnn. *arXiv preprint arXiv:1902.04615*, 2019.

[16] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das. Very deep convolutional neural networks for raw waveforms. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 421–425. IEEE, 2017.

[17] Ingrid Daubechies. *Fundamental papers in wavelet theory*. Princeton University Press, 2006.

[18] Sander Dieleman and Benjamin Schrauwen. End-to-end learning for music audio. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6964–6968. IEEE, 2014.

[19] Sander Dieleman, Jeffrey De Fauw, and Koray Kavukcuoglu. Exploiting cyclic symmetry in convolutional neural networks. *arXiv preprint arXiv:1602.02660*, 2016.

[20] Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. *arXiv preprint arXiv:2002.12880*, 2020.

[21] Sadaoki Furui. Speaker-independent isolated word recognition based on emphasized spectral dynamics. In *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 1991–1994. IEEE, 1986.

[22] Dennis Gabor. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, 93(26): 429–441, 1946.

[23] Simon Graham, David Epstein, and Nasir Rajpoot. Dense steerable filter cnns for exploiting rotational symmetry in histology images. *arXiv preprint arXiv:2004.03037*, 2020.

[24] Alex Grossmann, Jean Morlet, and T Paul. Transforms associated to square integrable group representations. i. general results. *Journal of Mathematical Physics*, 26(10):2473–2479, 1985.

[25] Eric Guizzo, Tillman Weyde, and Jack Barnett Leveson. Multi-time-scale convolution for emotion recognition from speech audio signals. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6489–6493. IEEE, 2020.

[26] Michael J Hewitt and Ray Meddis. A computer model of amplitude-modulation sensitivity of single units in the inferior colliculus. *The Journal of the Acoustical Society of America*, 95(4): 2145–2159, 1994.

[27] Emiel Hoogeboom, Jorn WT Peters, Taco S Cohen, and Max Welling. Hexaconv. *arXiv preprint arXiv:1803.02108*, 2018.

[28] PLM Johannesma. The pre-response stimulus ensemble of neurons in the cochlear nucleus. In *Symposium on Hearing Theory, 1972*. IPO, 1972.

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[30] Jonas Köhler, Leon Klein, and Frank Noé. Equivariant flows: sampling configurations for multi-body systems with symmetric energies. *arXiv preprint arXiv:1910.00753*, 2019.

[31] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. *arXiv preprint arXiv:1802.03690*, 2018.

[32] Edith Law, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie. Evaluation of algorithms using games: The case of music tagging. In *ISMIR*, pages 387–392, 2009.

[33] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[34] Jongpil Lee, Jiyoung Park, Keunhyoung Luke Kim, and Juhan Nam. Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms. *arXiv preprint arXiv:1703.01789*, 2017.

[35] Jan Eric Lenssen, Matthias Fey, and Pascal Libuschewski. Group equivariant capsule networks. In *Advances in Neural Information Processing Systems*, pages 8844–8853, 2018.

[36] Junying Li, Zichen Yang, Haifeng Liu, and Deng Cai. Deep rotation equivariant network. *Neurocomputing*, 290:26–33, 2018.

[37] Tom LH Li, Antoni B Chan, and Andy HW Chun. Automatic musical pattern feature extraction using convolutional neural network. *Genre*, 10:1x1, 2010.

[38] Thomas Lidy and Alexander Schindler. Cqt-based convolutional neural networks for audio scene classification. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, volume 90, pages 1032–1048, 2016.

[39] Tony Lindeberg. *Scale-space theory in computer vision*, volume 256. Springer Science & Business Media, 2013.

[40] Tony Lindeberg and Anders Friberg. Idealized computational models for auditory receptive fields. *PLoS one*, 10(3), 2015.

[41] Tony Lindeberg and Anders Friberg. Scale-space theory for auditory signals. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 3–15. Springer, 2015.

[42] Jen-Yu Liu, Shyh-Kang Jeng, and Yi-Hsuan Yang. Applying topological persistence in convolutional neural network for music audio signals. *arXiv preprint arXiv:1608.07373*, 2016.

[43] Xugang Lu, Peng Shen, Sheng Li, Yu Tsao, and Hisashi Kawai. Deep progressive multi-scale attention for acoustic event classification. *arXiv preprint arXiv:1912.12011*, 2019.

[44] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

[45] Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5048–5057, 2017.

[46] Diego Marcos, Benjamin Kellenberger, Sylvain Lobry, and Devis Tuia. Scale equivariance in cnns with vector fields. *arXiv preprint arXiv:1807.11783*, 2018.

[47] Brian CJ Moore and Hedwig E Gockel. Properties of auditory stream formation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1591):919–931, 2012.

[48] Karol J Piczak. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2015.

[49] Jordi Pons and Xavier Serra. Randomly weighted cnns for (music) audio classification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 336–340. IEEE, 2019.

[50] Jordi Pons, Oriol Nieto, Matthew Prockup, Erik Schmidt, Andreas Ehmann, and Xavier Serra. End-to-end learning for music audio tagging at scale. *arXiv preprint arXiv:1711.02520*, 2017.

[51] Jordi Pons, Olga Slizovskaia, Rong Gong, Emilia Gómez, and Xavier Serra. Timbre analysis of music audio signals with convolutional neural networks. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 2744–2748. IEEE, 2017.

[52] Zhao Ren, Kun Qian, Yebin Wang, Zixing Zhang, Vedhas Pandit, Alice Baird, and Bjorn Schuller. Deep scalogram representations for acoustic scene classification. *IEEE/CAA Journal of Automatica Sinica*, 5(3):662–669, 2018.

[53] Dario Rethage, Jordi Pons, and Xavier Serra. A wavenet for speech denoising. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5069–5073. IEEE, 2018.

[54] Danilo Jimenez Rezende, Sébastien Racanière, Irina Higgins, and Peter Toth. Equivariant hamiltonian flows. *arXiv preprint arXiv:1909.13739*, 2019.

[55] David W. Romero and Mark Hoogendoorn. Co-attentive equivariant neural networks: Focusing equivariance on transformations co-occurring in data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1g6ogrtDr.

[56] David W Romero, Erik J Bekkers, Jakub M Tomczak, and Mark Hoogendoorn. Attentive group equivariant convolutional networks. *arXiv preprint arXiv:2002.03830*, 2020.

[57] Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283, 2017.

[58] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044, 2014.

[59] Roberta Santoro, Michelle Moerel, Federico De Martino, Rainer Goebel, Kamil Ugurbil, Essa Yacoub, and Elia Formisano. Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS computational biology*, 10(1), 2014.

[60] Louis L Scharf. *Statistical signal processing*, volume 98. Addison-Wesley Reading, MA, 1991.

[61] Jan Schlüter and Sebastian Böck. Improved musical onset detection with convolutional neural networks. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 6979–6983. IEEE, 2014.

[62] Ivan Sosnovik, Michał Szmaja, and Arnold Smeulders. Scale-equivariant steerable networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJgpugrKPS.

[63] Stanley Smith Stevens, John Volkmann, and Edwin B Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3): 185–190, 1937.

[64] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185*, 2018.

[65] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor Field Networks: Rotation-and Translation-Equivariant Neural Networks for 3D Point Clouds. *arXiv preprint arXiv:1802.08219*, 2018.

[66] Yuji Tokozume and Tatsuya Harada. Learning environmental sounds with end-to-end convolutional neural network. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2721–2725. IEEE, 2017.

[67] Karen Ullrich, Jan Schlüter, and Thomas Grill. Boundary detection in music structure analysis using convolutional neural networks. In *ISMIR*, pages 417–422, 2014.

[68] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018.

[69] Leo Paulus Antonie Servatius van Noorden et al. *Temporal coherence in the perception of tone sequences*, volume 3. Institute for Perceptual Research Eindhoven, the Netherlands, 1975.

[70] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pages 210–218. Springer, 2018.

[71] Sai Raam Venkataraman, S. Balasubramanian, and R. Raghunatha Sarma. Building deep equivariant capsule networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJgNJgSFPS.

[72] Patrick von Platen, Chao Zhang, and Philip Woodland. Multi-span acoustic modelling using raw waveform signals. *arXiv preprint arXiv:1906.11047*, 2019.

[73] Rui Wang, Robin Walters, and Rose Yu. Incorporating symmetry into deep dynamics models for improved generalization. *arXiv preprint arXiv:2002.03061*, 2020.

[74] Maurice Weiler and Gabriele Cesa. General e (2)-equivariant steerable cnns. In *Advances in Neural Information Processing Systems*, pages 14334–14345, 2019.

[75] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco S Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. In *Advances in Neural Information Processing Systems*, pages 10381–10392, 2018.

[76] Maurice Weiler, Fred A Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 849–858, 2018.

[77] Daniel Worrall and Gabriel Brostow. Cubenet: Equivariance to 3d rotation and translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 567–584, 2018.

[78] Daniel E Worrall and Max Welling. Deep scale-spaces: Equivariance over scale. *arXiv preprint arXiv:1905.11697*, 2019.

[79] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5028–5037, 2017.

[80] Yong Xu, Qiuqiang Kong, Wenwu Wang, and Mark D Plumbley. Large-scale weakly supervised audio classification using gated convolutional neural network. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 121–125. IEEE, 2018.

[81] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[82] Chiyuan Zhang, Stephen Voinea, Georgios Evangelopoulos, Lorenzo Rosasco, and Tomaso Poggio. Discriminative template learning in group-convolutional networks for invariant speech representations. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[83] Zhoutong Zhang, Yunyun Wang, Chuang Gan, Jiajun Wu, Joshua B. Tenenbaum, Antonio Torralba, and William T. Freeman. Deep audio priors emerge from harmonic convolutional networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygjHxrYDB.

[84] Zhenyao Zhu, Jesse H Engel, and Awni Hannun. Learning multiscale features directly from waveforms. *arXiv preprint arXiv:1603.09509*, 2016.

Appendix

A   Concepts from Group Theory

Definition A.1. *Group.* A group is a tuple $(G, \cdot)$ consisting of a set G and a binary operation $\cdot : G \times G \to G$, referred to as the *group product*, that satisfies the following axioms: *Closure*: For all $h, g \in G$, $h \cdot g \in G$; *Identity:* There exists $e \in G$, such that $e \cdot g = g \cdot e = g$; *Inverse:* For all $g \in G$, there exists an element $g^{-1} \in G$, such that $g \cdot g^{-1} = g^{-1} \cdot g = e$; and *Associativity:* For all $g, h, k \in G$, $(g \cdot h) \cdot k = g \cdot (h \cdot g)$.

Definition A.2. *Semi-direct product and affine groups.* In practice, one is mainly interested in the analysis of data defined on $\mathbb{R}^d$. Consequently, groups of the form $G = \mathbb{R}^d \rtimes H$, resulting from the *semi-direct product* ($\rtimes$) between the translation group $\mathbb{R}^d$ and an arbitrary (Lie) group H that acts on $\mathbb{R}^d$, e.g., rotation, scaling or mirroring, are of main interest. This family of groups is referred to as *affine groups* and their group product is defined as:

$$g_1 \cdot g_2 = (x_1, h_1) \cdot (x_2, h_2) = (x_1 + h_1 \odot x_2, h_1 \cdot h_2), \tag{24}$$

with $g_1 = (x_1, h_1)$, $g_2 = (x_2, h_2) \in G$, $x_1, x_2 \in \mathbb{R}^d$ and $h_1, h_2 \in H$. The operator $\odot$ denotes the *action of $h \in H$ on $x \in \mathbb{R}^d$*, and it describes how a vector $x \in \mathbb{R}^d$ is modified by elements $h \in H$. The most relevant affine group for this work is the *dilation-translation group* $\mathbb{R}^d \rtimes \mathbb{R}^+$ acting on $\mathbb{R}^d$.

Definition A.3. *Group representation.* Let G be a group and $\mathbf{L}^2(X)$ be a space of functions defined on some vector space X. The *(left) regular group representation* of G is a linear transformation $\mathcal{L} : G \times \mathbf{L}^2(X) \to \mathbf{L}^2(X)$, $(g, f) \mapsto \mathcal{L}_g[f] := f(g^{-1} \odot x)$, that shares the group structure via:

$$\mathcal{L}_{g_1} \mathcal{L}_{g_2}[f] = \mathcal{L}_{g_1 g_2}[f] \tag{25}$$

for any $g_1, g_2 \in G$, $f \in \mathbb{L}_2(X)$. That is, concatenating two such transformations, parametrized by g_1 and g_2, is equivalent to a single transformation parametrized by $g_1 \cdot g_2 \in G$. Intuitively, the representation of G on a function f describes how the function as a whole, i.e., $f(x), \forall x \in X$, is transformed by the effect of group elements $g \in G$. If the group G is affine, the group representation $\mathcal{L}_g$ can be split as:

$$\mathcal{L}_g[f] = \mathcal{L}_x \mathcal{L}_h[f], \tag{26}$$

with $g = (x, h) \in G$, $x \in \mathbb{R}^d$ and $h \in H$.

B   Effect of Input Transformations on Time-Frequency Transformations

B.1   The Fourier Transform

Fourier analysis represents an arbitrary finite energy function $f \in \mathbf{L}^2(\mathbb{R})$ as a sum of complex sinusoidal waves $\mathrm{e}^{i\omega t} = \cos \omega t + i \sin \omega t$:

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega) \, \mathrm{e}^{i\omega t} \, \mathrm{d}\omega.$$

Here, $\hat{f}(\omega)$ depicts the amplitude of each sinusoidal wave $\mathrm{e}^{i\omega t}$ in f and can be understood as the "amount" of $\mathrm{e}^{i\omega t}$ in f. Consequently, the coefficients $\hat{f}(\omega)$ correspond to the correlation of f and $\mathrm{e}^{i\omega t}$ and is referred to as the *Fourier transform* $\mathcal{F}$:

$$\mathcal{F}[f](\omega) = \hat{f}(\omega) = \langle f, \mathrm{e}^{i\omega t} \rangle = \int_{-\infty}^{\infty} f(t) \, \mathrm{e}^{-i\omega t} \, \mathrm{d}t$$

This is equivalent to encoding f into a time-frequency dictionary $\mathcal{D} = \{\mathrm{e}^{i\omega t}\}_{\omega \in \mathbb{R}}$.

The effect of input translation. Let $\mathcal{L}_{t_0}[f](t) = f(t - t_0)$ be a translated version of f. Its Fourier transform is given by:

$$\mathcal{F}[\mathcal{L}_{t_0}[f]](\omega) = \int_{-\infty}^{\infty} f(t - t_0)\, \mathrm{e}^{-i\omega t}\, \mathrm{d}t \qquad \text{substitute } \tilde{t} = t - t_0 \ ; \ \mathrm{d}\tilde{t} = \mathrm{d}t$$

$$= \int_{-\infty}^{\infty} f(\tilde{t})\, \mathrm{e}^{-i\omega(\tilde{t}+t_0)}\, \mathrm{d}\tilde{t}$$

$$= \mathrm{e}^{-i\omega t_0} \int_{-\infty}^{\infty} f(\tilde{t})\, \mathrm{e}^{-i\omega\tilde{t}}\, \mathrm{d}\tilde{t}$$

$$= \mathrm{e}^{-i\omega t_0} \mathcal{F}[f](\omega) \tag{27}$$

In other words, a translation of t_0 in the time domain corresponds to a phase modulation of $\mathrm{e}^{-i\omega t_0}$ in the frequency domain.

The effect of input scaling. Let $\mathcal{L}_{s_0}[f](t) = f(s_0^{-1}t)$, $s_0 \in \mathbb{R}^+$, be a scaled version of f. Its Fourier transform is given by:

$$\mathcal{F}[\mathcal{L}_{s_0}[f]](\omega) = \int_{-\infty}^{\infty} f(s_0^{-1}t)\, \mathrm{e}^{-i\omega t}\, \mathrm{d}t \qquad \text{substitute } \tilde{t} = s_0^{-1}t \ ; \ \mathrm{d}\tilde{t} = s_0^{-1}\mathrm{d}t$$

$$= \int_{-\infty}^{\infty} f(\tilde{t})\, \mathrm{e}^{-i\omega(s_0\tilde{t})}\, \mathrm{d}(s_0\tilde{t})$$

$$= s_0 \int_{-\infty}^{\infty} f(\tilde{t})\, \mathrm{e}^{-i(s_0\omega)\tilde{t}}\, \mathrm{d}\tilde{t}$$

$$= s_0 \mathcal{F}[f](s_0\omega) = s_0 \mathcal{L}_{s_0^{-1}}[\mathcal{F}[f]](\omega) \tag{28}$$

In other words, a dilation (compression) on the time domain corresponds to a compression (dilation) in the frequency domain times the inverse of the "amount" of dilation (compression). Following the same derivation procedure, it can be shown that:

$$\mathcal{F}[\mathcal{L}_{s_0}\mathcal{L}_{t_0}[f]](\omega) = s_0 \mathrm{e}^{-i\omega t_0} \mathcal{F}[f](s_0\omega) = \mathrm{e}^{-i\omega t_0} s_0 \mathcal{L}_{s_0^{-1}}[\mathcal{F}[f]](\omega). \tag{29}$$

Transformation effects on the spectral density. The spectral density of a function $f \in \mathbf{L}^2(\mathbb{R})$ is defined as $|\mathcal{F}[f](\omega)|^2$. Consequently, Eqs. 27 and 28 are reduced to Eqs. 30 and 31, respectively:

$$|\mathcal{F}[\mathcal{L}_{t_0}[f]](\omega)|^2 = |\mathcal{F}[f](\omega)|^2 \tag{30}$$

$$|\mathcal{F}[\mathcal{L}_{s_0}[f]](\omega)|^2 = |s_0|^2 |\mathcal{L}_{s_0^{-1}}[\mathcal{F}[f]](\omega)|^2 \tag{31}$$

Equivariance and invariance properties. From Eq. 27 we can see that the Fourier transform is translation equivariant in the sense that it encodes translations of the input as a phase modulation of the output. Furthermore, we can also see (Eq. 28) that the Fourier transform is scale equivariant as well, as it encodes scaling in the input as a modulation of the frequency components in the output. One can prove that the Fourier transform is indeed equivariant to both transformations by proving that the output transformations $\mathrm{e}^{-i\omega t_0}$ and $s_0 \mathcal{L}_{s_0^{-1}}$ are *group representations* (Def. A.3) of the translation and scaling group, respectively, in the Fourier space. That is, it holds that the combination of any translations $t_0, t_1 \in \mathbb{R}$ or scalings $s_0, s_1 \in \mathbb{R}^+$ in the input domain, $\mathcal{L}_{t_1}[\mathcal{L}_{t_0}[f]] = \mathcal{L}_{t_1+t_0}[f]$, $\mathcal{L}_{s_1}[\mathcal{L}_{s_0}[f]] = \mathcal{L}_{s_1 s_0}[f]$, produces a transformation on the Fourier domain that preserves the group structure: $\mathrm{e}^{-i\omega t_1}\mathrm{e}^{-i\omega t_0}\mathcal{F}[f](\omega) = \mathrm{e}^{-i\omega(t_1+t_0)}\mathcal{F}[f](\omega)$, $s_1 \mathcal{L}_{s_1^{-1}}[s_0 \mathcal{L}_{s_0^{-1}}[\mathcal{F}[f]]](\omega) = (s_0 s_1)\mathcal{L}_{(s_0 s_1)^{-1}}[\mathcal{F}[f]](\omega)$. Unfortunately, the resulting group representations rapidly become cumbersome, specially in the presence of several input components.

The calculation of the spectral density leaves the scale equivariance property of the transformation unaffected. However, Eq. 30 shows that the spectral density calculation reduces translation equivariance to *translation invariance*. Consequently, the Fourier transform is commonly considered *not to carry positional information*. It is important to note that this lost of information destroys the invertibility property of the Fourier transform.

B.2 The Short-Time Fourier Transform

The windowed Fourier transform (also called *short-time Fourier transform*) of a signal $f \in \mathbf{L}^2(\mathbb{R})$ is given by:

$$\mathcal{S}[f](u, \xi) = \langle f, w_{u,\xi} \rangle = \int_{-\infty}^{+\infty} f(t) w(t-u) \, \mathrm{e}^{-i\xi t} \, \mathrm{d}t.$$

This is equivalent to encoding f into a time-frequency dictionary $\mathcal{D} = \{w_{u,\xi}\}_{(u,\xi) \in \mathbb{R}^2}$.

The effect of input translation. Let $\mathcal{L}_{t_0}[f](t) = f(t - t_0)$ be a translated version of f. Its short-time Fourier transform is given by:

$$\mathcal{S}[\mathcal{L}_{t_0}[f]](u, \xi) = \int_{-\infty}^{\infty} f(t - t_0) w(t - u) \, \mathrm{e}^{-i\xi t} \, \mathrm{d}t \qquad \text{substitute } \tilde{t} = t - t_0 \; ; \; \mathrm{d}\tilde{t} = \mathrm{d}t$$

$$= \int_{-\infty}^{\infty} f(\tilde{t}) w(\tilde{t} + t_0 - u) \, \mathrm{e}^{-i\xi(\tilde{t} + t_0)} \, \mathrm{d}\tilde{t}$$

$$= \mathrm{e}^{-i\xi t_0} \int_{-\infty}^{\infty} f(\tilde{t}) w(t - (u - t_0)) \, \mathrm{e}^{-i\xi \tilde{t}} \, \mathrm{d}\tilde{t}$$

$$= \mathrm{e}^{-i\xi t_0} \mathcal{S}[f](u - t_0, \xi) = \mathrm{e}^{-i\xi t_0} \mathcal{L}_{t_0}[\mathcal{S}[f]](u, \xi) \qquad (32)$$

In other words, a translation by t_0 in the time domain, corresponds to a shift of t_0 units along the time dimension of the time-frequency representation of $\mathcal{S}[f]$, and an additional phase modulation of $\mathrm{e}^{-i\xi t_0}$ within the window $\mathcal{L}_{(u-t_0)}[w](t)$, similar to that of the Fourier transform (Eq. 27).

The effect of input scaling. Let $\mathcal{L}_{s_0}[f](t) = f(s_0^{-1}t)$, $s_0 \in \mathbb{R}^+$, be a scaled version of f. Its short-time Fourier transform is given by:

$$\mathcal{S}[\mathcal{L}_{s_0}[f]](u, \xi) = \int_{-\infty}^{\infty} f(s_0^{-1}t) w(t - u) \, \mathrm{e}^{-i\xi t} \, \mathrm{d}t \qquad \text{substitute } \tilde{t} = s_0^{-1}t \; ; \; \mathrm{d}\tilde{t} = s_0^{-1}\mathrm{d}t$$

$$= \int_{-\infty}^{\infty} f(\tilde{t}) w(s_0\tilde{t} - u) \, \mathrm{e}^{-i\xi(s_0\tilde{t})} \, \mathrm{d}(s_0\tilde{t})$$

$$= s_0 \int_{-\infty}^{\infty} f(\tilde{t}) w(s_0\tilde{t} - u) \, \mathrm{e}^{-i(s_0\xi)\tilde{t}} \, \mathrm{d}\tilde{t} \qquad \text{substitute } u = s_0^{-1}s_0 u$$

$$= s_0 \int_{-\infty}^{\infty} f(\tilde{t}) w(s_0(\tilde{t} - s_0^{-1}u)) \, \mathrm{e}^{-i(s_0\xi)\tilde{t}} \, \mathrm{d}\tilde{t} \qquad \text{approximation } w(st) \approx w(t)$$

$$\approx s_0 \int_{-\infty}^{\infty} f(\tilde{t}) w(\tilde{t} - s_0^{-1}u) \, \mathrm{e}^{-i(s_0\xi)\tilde{t}} \, \mathrm{d}\tilde{t}$$

$$= s_0 \, \mathcal{S}[f](s_0^{-1}u, s_0\xi) \qquad (33)$$

In other words, a dilation in the time domain corresponds to a compression in the frequency domain, and fully corresponds to the phenomenon exhibited by the Fourier transform (Eq. 28). It is important to note that the approximation $w(x) \approx w(sx)$ generally does not hold. This approximation essentially implies spatial invariance of w, which can only roughly hold with increasing window sizes, i.e., when the short-term Fourier transform starts to approximate the (global) Fourier transform.

Transformation effects on the spectrogram. The spectrogram of a function $f \in \mathbf{L}^2(\mathbb{R})$ is defined as $|\mathcal{S}[f](u, \xi)|^2$. Consequently, Eqs. 32 and 33 are reduced to Eqs. 34 and 35, respectively:

$$|\mathcal{F}[\mathcal{L}_{t_0}[f]](u, \xi)|^2 = |\mathcal{L}_{t_0}[\mathcal{S}[f]](u, \xi)|^2 \qquad (34)$$

$$|\mathcal{F}[\mathcal{L}_{s_0}[f]](u, \xi)|^2 = |s_0|^2 |\mathcal{S}[f](s_0^{-1}u, s_0\xi)|^2 \qquad (35)$$

Equivariance and invariance properties. The short-time Fourier transform is approximately translation and scale equivariance in a manner similar to that of the Fourier transform. In comparison to the Fourier transform however, it "decomposes" input translations as a translation $u - t_0$ and a phase shift $\mathrm{e}^{-i\xi t_0}$ in the output space (Eq. 32). A shift in the input domain can thus be interpreted as the composition in the output space of a rough estimate $u - t_0$ signaling the position in which the window w is localized, and a fine grained localization given by the phase shift $\mathrm{e}^{-i\xi t_0}$ indicating the relative position of the pattern within the window $\mathcal{L}_{(u-t_0)}[w](t)$. Moreover, scale equivariance
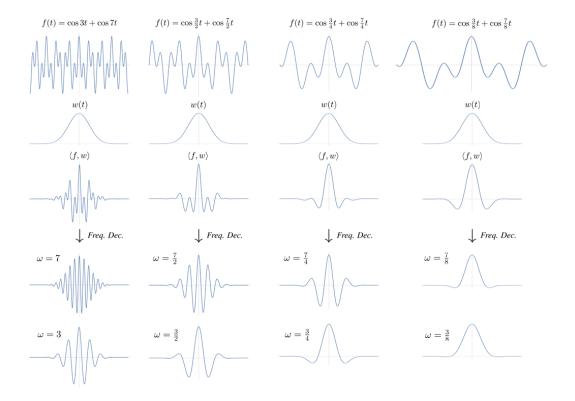
Figure 3: Scale equivariance in the windowed Fourier transform. Consider a function $f(t) = \cos \omega_1 t + \cos \omega_2 t$ composed of two frequencies $\omega_1 = 3$ and $\omega_2 = 7$, and a window function $w(t)$, with which the windowed Fourier transform is performed. For relatively high frequencies (left column), we see that the dot-product of f and w, $\langle f, w \rangle$, is able to capture sufficient spectral information from f, and the frequencies ω_1, ω_2 can perfectly be extracted from it. However, for dilated versions of the same signal f (right columns) obtained by reducing the frequency of the spectral components ω_1, ω_2 of f, we see that the capacity of the dot-product $\langle f, w \rangle$ to capture the spectral information in the input gradually degrades and, eventually, is entirely lost. We see thus that scale equivariance (approximately) holds for the set of scales for which *all* the spectral components of the signal of interest f lie within the range of the window w.

is analogous to the scale equivariance behaviour of the Fourier transform up to the fact that time and frequency are now jointly described. However, since the window itself does not scale with the sampled frequency (as is the case in wavelet transforms), exact equivariance does not hold. It is important to signalize that scale equivariance is now restricted to a subset of scales determined by the width of the window w utilized during the transformation (see Fig. 3 for a visual explanation).

The calculation of the spectrogram leaves the scale equivariance property of the transformation unaffected and is equivalent in a join manner to the scale equivariance property of the Fourier transform (Eq. 33). In contrast to the Fourier transform however, Eq. 34 shows that translation equivariance is partially preserved (only information about the phase shift within the window is lost). Resultantly, the short-time Fourier transform is considered to *carry positional information*.

B.3   The Wavelet Transform

The wavelet transform $\mathcal{W}[f]$ of a signal $f \in \mathbf{L}^2(\mathbb{R})$ is given by:

$$\mathcal{W}[f](u, s) = \langle f, \psi_{u,s} \rangle = \int_{-\infty}^{+\infty} f(t) \, \frac{1}{\sqrt{s}} \psi^* \left(\frac{t - u}{s} \right) \, \mathrm{d}t.$$

This is equivalent to encoding f into a time-frequency dictionary $\mathcal{D} = \{\psi_{u,s}\}_{u \in \mathbb{R}, s \in \mathbb{R}^+}$.

The effect of input translation. Let $\mathcal{L}_{t_0}[f](t) = f(t - t_0)$ be a translated version of f. Its wavelet transform is given by:

$$\mathcal{W}[\mathcal{L}_{t_0}[f]](u, s) = \int_{-\infty}^{\infty} f(t - t_0)(\sqrt{s}\,)^{-1}\psi^*\big(s^{-1}(t - u)\big)\, \mathrm{d}t \qquad \text{substitute } \tilde{t} = t - t_0\,;\, \mathrm{d}\tilde{t} = \mathrm{d}t$$

$$= \int_{-\infty}^{\infty} f(\tilde{t})(\sqrt{s}\,)^{-1}\psi^*\big(s^{-1}(\tilde{t} + t_0 - u)\big)\, \mathrm{d}\tilde{t}$$

$$= \int_{-\infty}^{\infty} f(\tilde{t})(\sqrt{s}\,)^{-1}\psi^*\big(s^{-1}(\tilde{t} - (u - t_0))\big)\, \mathrm{d}\tilde{t}$$

$$= \mathcal{W}[f](u - t_0, s) = \mathcal{L}_{t_0}\mathcal{W}[f](u, s) \tag{36}$$

In other words, a translation of t_0 in the input domain produces an equivalent translation in the wavelet domain.

The effect of input scaling. Let $\mathcal{L}_{s_0}[f](t) = f(s_0^{-1}t)$ be a scaled version of f. The corresponding wavelet transform is given by:

$$\mathcal{W}[\mathcal{L}_{s_0}[f]](u, s) = \int_{-\infty}^{\infty} f(s_0^{-1}t)(\sqrt{s}\,)^{-1}\psi^*\big(s^{-1}(t - u)\big)\, \mathrm{d}t \qquad \text{substitute } \tilde{t} = s_0^{-1}t\,;\, \mathrm{d}\tilde{t} = s_0^{-1}\mathrm{d}t$$

$$= \int_{-\infty}^{\infty} f(\tilde{t})(\sqrt{s}\,)^{-1}\psi^*\big(s^{-1}(s_0\tilde{t} - u)\big)\, \mathrm{d}(s_0\tilde{t}) \qquad \text{substitute } u = s_0^{-1}s_0 u$$

$$= \int_{-\infty}^{\infty} f(\tilde{t})(\sqrt{s}\,)^{-1}s_0\psi^*\big(s^{-1}s_0(\tilde{t} - s_0^{-1}u)\big)\, \mathrm{d}\tilde{t} \qquad \text{substitute } s_0 = \left(\sqrt{s_0^{-1}}\sqrt{s_0^{-1}}\right)^{-1}$$

$$= \sqrt{s_0}\int_{-\infty}^{\infty} f(\tilde{t})\left(\sqrt{s_0^{-1}s}\right)^{-1}\psi^*\big((s_0^{-1}s)^{-1}(\tilde{t} - s_0^{-1}u)\big)\, \mathrm{d}\tilde{t}$$

$$= \sqrt{s_0}\,\mathcal{W}[f](s_0^{-1}u, s_0^{-1}s) = \sqrt{s_0}\,\mathcal{L}_{s_0}\mathcal{W}[f](u, s) \tag{37}$$

In other words, a dilation of s_0 in the input domain produces an *equivalent dilation* in the wavelet domain on both components (u, s), up to a multiplicative factor $\sqrt{s_0}$. Following the same procedure, it can be shown that:

$$\mathcal{W}[f(s_0^{-1}(t - t_0)](u, s) = \sqrt{s_0}\,\mathcal{W}[f](s_0^{-1}(u - t_0), s_0^{-1}s) = \sqrt{s_0}\,\mathcal{L}_{t_0}\mathcal{L}_{s_0}\mathcal{W}[f](u, s) \tag{38}$$

Transformation effects on the scalogram. The scalogram of a function $f \in \mathbf{L}^2(\mathbb{R})$ is defined as $|\mathcal{W}[f](u, s)|^2$. Consequently, Eqs. 36 and 37 become Eqs. 39 and 40, respectively:

$$|\mathcal{W}[\mathcal{L}_{t_0}[f]](u, s)|^2 = |\mathcal{L}_{t_0}[\mathcal{W}[f]](u, s)|^2 \tag{39}$$

$$|\mathcal{W}[\mathcal{L}_{s_0}[f]](u, s)|^2 = |\mathcal{L}_{s_0}[\mathcal{W}[f]](u, s)|^2 \tag{40}$$

Equivariance and invariance properties. From Eq. 36, we can see that the wavelet transform is *exactly equivariant to translations* and the resulting group representation on the output space equals that of the input space. Furthermore, translation equivariance is preserved in the scalogram as well (Eq. 39). Similarly, scale equivariance is preserved both through the wavelet transform (Eq. 37) as well as the scalogram representation (Eq. 40).

Similarly to the translation equivariance case, the resulting scaling group representation on the output space resembles that of the input space. This behaviour leads to much more straightforward group representations than that exhibited by the Fourier or short-time Fourier transform, both for the translation and scaling group. Importantly, exact scale equivariance is *only* obtained on the scalogram (Eq. 40), whilst for the wavelet transform it is retained up to multiplicative factor (Eq. 37). This behaviour elucidates the fact that time-frequency transforms have been optimized for energy density representations rather than for the time-frequency representations themselves.

Derivation of scale equivariance on the scalogram. In this section and for the interested reader, we provide the derivation of scale equivariance on the scalogram (Eq. 40):

$$|\mathcal{W}[\mathcal{L}_{s_0}[f]](u,s)|^2 = \left| \int_{-\infty}^{\infty} f(s_0^{-1}t)(\sqrt{s}\,)^{-1}\psi^*(s^{-1}(t-u))\, \mathrm{d}t \right|^2$$

$$= \int_{-\infty}^{\infty} \left| f(s_0^{-1}t)(\sqrt{s}\,)^{-1}\psi^*(s^{-1}(t-u))\, \right|^2 \mathrm{d}t$$

$$= \int_{-\infty}^{\infty} |f(s_0^{-1}t)|^2 \, |(\sqrt{s}\,)^{-1}|^2 \, |\psi^*(s^{-1}(t-u))|^2 \, \mathrm{d}t \qquad \text{substitute } \tilde{t} = s_0^{-1}t \,;\, \mathrm{d}\tilde{t} = s_0^{-1}\mathrm{d}t$$

$$= \int_{-\infty}^{\infty} |f(\tilde{t})|^2 \, |(\sqrt{s}\,)^{-1}|^2 \, |\psi^*(s^{-1}(s_0\tilde{t}-u))|^2 \, \mathrm{d}(s_0\tilde{t}) \qquad \text{substitute } u = s_0^{-1}s_0 u$$

$$= \int_{-\infty}^{\infty} |f(\tilde{t})|^2 \, |(\sqrt{s}\,)^{-1}|^2 \, |\psi^*(s^{-1}s_0(\tilde{t}-s_0^{-1}u))|^2 \, \mathrm{d}(s_0\tilde{t}) \qquad \text{substitute } s_0 = \left| \sqrt{s_0^{-1}} \right|^2$$

$$= \int_{-\infty}^{\infty} |f(\tilde{t})|^2 \left| \left(\sqrt{s_0^{-1}s} \right)^{-1} \right|^2 \left| \psi^*\left((s_0^{-1}s)^{-1}(\tilde{t}-s_0^{-1}u) \right) \right|^2 \mathrm{d}\tilde{t}$$

$$= |\mathcal{W}[f](s_0^{-1}u, s_0^{-1}s)|^2 = |\mathcal{L}_{s_0}\mathcal{W}[f](u,s)|^2 \qquad (41)$$

C General Derivation of Equivariant Mappings for the Dilation-Translation Group $\mathbb{R}^d \rtimes \mathbb{R}^+$

In this section we provide the complete derivation of equivariant mappings starting from the equivariance constraint on $\mathbb{R}^d$. For completeness, let us first start with the conventional (spatial) convolution.[4]

The spatial convolution case. Let $f, \psi : \mathbb{R}^d \to \mathbb{R}$ be a scalar valued signal and filter defined on $\mathbb{R}^d$. The spatial convolution $(\star)$ is defined as:

$$[f \star \psi](u) = \int_{\mathbb{R}^d} f(x)\psi(x-u)\, \mathrm{d}x \,, u \in \mathbb{R}.$$

Now, let us consider a scaled translated version of the signal f, $\mathcal{L}_{y,z}[f](x) = f(z^{-1}(x-y)), y \in \mathbb{R}^d$, $z \in \mathbb{R}^+$. The correlation of $\mathcal{L}_{y,z}[f]$ with ψ is equal to:

$$[\mathcal{L}_{y,z}[f] \star \psi](u) = \int_{\mathbb{R}^d} f(z^{-1}(x-y))\psi(x-u)\, \mathrm{d}x$$

$$= \int_{\mathbb{R}^d} f(\tilde{x})\psi(z\tilde{x}+y-u)\, \mathrm{d}(z^d\tilde{x})$$

$$= \int_{\mathbb{R}^d} f(\tilde{x})\psi(z(\tilde{x}-z^{-1}(u-y))\, \mathrm{d}(z^d\tilde{x})$$

$$= z^d \int_{\mathbb{R}^d} f(\tilde{x})\mathcal{L}_{z^{-1}}[\psi](\tilde{x}-z^{-1}(u-y))\, \mathrm{d}\tilde{x}$$

$$= z^d[f \star \mathcal{L}_{z^{-1}}[\psi]](z^{-1}(u-y))$$

$$= z^d\mathcal{L}_{y,z}[f \star \mathcal{L}_{z^{-1}}[\psi]](u). \qquad (42)$$

We can see that, when comparing Eq. 42 with the equivariance condition:

$$[\mathcal{L}_{y,z}[f] \star_{\mathbb{R}^d} \psi](u) \stackrel{!}{=} \mathcal{L}_{y,z}[f \star_{\mathbb{R}^d} \psi](u),$$

we encounter some spurious coefficients: a multiplication by the correction factor of the Lebesgue measure via the scaling z, z^d, and a filter modification by $\mathcal{L}_{z^{-1}}$. We observe thus that spatial convolutions are *not* equivariant to scale transformations, unless $\forall_{z \in \mathbb{R}^+} : \mathcal{L}_z[\psi](x) = \psi(x)$, see e.g. [3, Thm. 1], but these are certainly not in $\mathbf{L}^2$.

It is well known, however, that equivariance to an arbitrary group can only be obtained via group convolutions (or reparametrizations thereof) defined in the corresponding group [31, 14, 3]. For our

---

[4]Formally this is a cross-correlation operation. However, we stick to the standard deep learning terminology.

particular case, this means that in order to obtain equivariance to the dilation-translation group acting on $\mathbb{R}^d$, $\mathbb{R}^d \rtimes \mathbb{R}^+$, we are necessarily restricted to group convolutions defined on it. Specifically, this means that it is sufficient for our study to find the constraints necessary for exact equivariance on lifting and group convolutions for the dilation-translation group, which we discuss next.

The lifting convolution case. Let us now define a lifting convolution from $\mathbb{R}^d$ to $\mathbb{R}^d \rtimes \mathbb{R}^+$ simply by changing the translation representation (for regular spatial convolutions) with the scale-translation representation:

$$[f \tilde{\star}_{G\uparrow} \psi](u, s) := \langle f, \mathcal{L}_{u,s}[\psi] \rangle = \int_{\mathbb{R}^d} f(\tilde{x}) \mathcal{L}_u \mathcal{L}_s[\psi](\tilde{x}) \, \mathrm{d}\tilde{x} = [f \star \mathcal{L}_s[\psi]](u).$$

The lifting convolution of $\mathcal{L}_{y,z}[f]$ with ψ is given by:

$$
\begin{aligned}
[\mathcal{L}_{y,z}[f] \tilde{\star}_{G\uparrow} \psi](u, s) &= \int_{\mathbb{R}^d} f(z^{-1}(x - y)) \mathcal{L}_s[\psi](x - u) \, \mathrm{d}x \\
&= \int_{\mathbb{R}^d} f(z^{-1}(x - y)) \psi(s^{-1}(x - u)) \, \mathrm{d}x \qquad \text{substitute } \tilde{x} = z^{-1}(x - y) \, ; \, \mathrm{d}\tilde{x} = z^{-d} \mathrm{d}x \\
&= \int_{\mathbb{R}^d} f(\tilde{x}) \psi(s^{-1}(z\tilde{x} + y - u)) \, \mathrm{d}(z^d \tilde{x}) \\
&= \int_{\mathbb{R}^d} f(\tilde{x}) \psi(s^{-1} z (\tilde{x} - z^{-1}(u - y))) \, \mathrm{d}(z^d \tilde{x}) \\
&= z^d \int_{\mathbb{R}^d} f(\tilde{x}) \mathcal{L}_{z^{-1}s}[\psi](\tilde{x} - z^{-1}(u - y)) \, \mathrm{d}\tilde{x} \\
&= z^d [f \star_{G\uparrow} \mathcal{L}_{z^{-1}s}[\psi]](z^{-1}(u - y)) \\
&= z^d [f \star_{G\uparrow} \psi](z^{-1}(u - y), z^{-1}s) \\
&= z^d \mathcal{L}_{y,z}[f \tilde{\star}_{G\uparrow} \psi](u, s).
\end{aligned}
\tag{43}
$$

We can see that, with such a lifting layer, we obtain equivariance to the dilation-translation group, up to a multiplicative factor z^d resulting from the correction of the Lebesgue measure via the scaling z. Since the equivariance condition must be valid for any input $f \in \mathbf{L^2}(\mathbb{R}^d)$, we must then find constraints on the parametrization of the filter ψ, such that exact equivariance is warranted. This is obtained by parametrizing the filter ψ as an scaled-normalized function $\psi_z = \frac{1}{z}\psi$. Since $z \in \mathbb{R}^+$, such parametrization is well-defined, and we can define the lifting convolution (Eq. 19) as:

$$[f \star_{G\uparrow} \psi](u, s) = [f \tilde{\star}_{G\uparrow} \psi_s](u, s) = \int_{\mathbb{R}^d} f(\tilde{x}) \frac{1}{s^d} \, \mathcal{L}_u \mathcal{L}_s[\psi](\tilde{x}) \, \mathrm{d}\tilde{x},$$

and consequently:

$$[\mathcal{L}_{y,z}[f] \star_{G\uparrow} \psi](u, s) = \mathcal{L}_{y,z}[f \star_{G\uparrow} \psi](u, s). \tag{44}$$

In other words, equivariance is obtained without any modifying front factors by using the regular representations $\mathcal{L}_{y,z}$ of the scale-translation group. Note that this parametrization is in fact equivalent to that of the wavelet transform (Sec. 2), up to an additional re-normalization factor $(\sqrt{s})^{-1}$. In conclusion, a lifting correlation from $\mathbb{R}^d$ to the dilation-translation group acting on $\mathbb{R}^d$ is (exactly) equivariant to the dilation-translation group via the left-regular representations if and only if the set of transformed filters corresponds to a dictionary of *re-normalized wavelet time-frequency atoms* $\mathcal{D}$ defined as:

$$\mathcal{D} = \left\{ \psi_{u,s}(x) = \frac{1}{s^d} \psi\left(\frac{x - u}{s}\right) \right\}_{u \in \mathbb{R}, s \in \mathbb{R}^+}. \tag{45}$$

Analogously to the wavelet transform, we define the wavelet $\psi_{0,1}(x) = \psi(x)$ as the *re-normalized mother wavelet*. Furthermore, it can be further proven that the scale in which the primary scale of the filter $\psi_{0,1}$ is fixated is not of relevance. This results from the fact that $(\mathbb{R}^+, \cdot)$ is a group and hence, for any $s_2, s_1 \in \mathbb{R}^+$, there always exists s, such that $s_2 = s \cdot s_1$. In other words, any value of $s \in \mathbb{R}^+$ can be obtained from any arbitrary fixated origin s_0.

The group correlation case. Let us now analyze the constraints required to obtain exact equivariance for group correlations. Let $f(x, \bar{s})$, $\psi(x, \bar{s}) : \mathbb{R}^d \rtimes \mathbb{R}^+ \to \mathbb{R}$ be a scalar valued signal and filter defined on the dilation-translation group acting on $\mathbb{R}^d$. Let us define a group correlation by simply

transforming the wavelet ψ via the representation of the scale translation group together with the Lebesgue inner product on $\mathbb{R}^d \times \mathbb{R}^+$:[5]

$$[f \tilde{\star}_G \psi](u,s) := \langle f, \mathcal{L}_{u,s}[\psi] \rangle = \int_{\mathbb{R}^d} \int_{\mathbb{R}^+} f(x,\bar{s}) \mathcal{L}_u \mathcal{L}_s[\psi](x,\bar{s}) \, \mathrm{d}\bar{s} \, \mathrm{d}x.$$

Similarly to the previous sections, let $\mathcal{L}_{y,z}[f](x,\bar{s}) = f(z^{-1}(x-y), z^{-1}\bar{s})$ be a scaled translated version of the signal f, $y \in \mathbb{R}^d$, $z \in \mathbb{R}^+$. The resulting group correlation is computed as:

$$
\begin{aligned}
[\mathcal{L}_{y,z}[f] \tilde{\star}_G \psi](u,s) &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^+} f(z^{-1}(x-y), z^{-1}\bar{s}) \mathcal{L}_{us}[\psi](x,\bar{s}) \, \mathrm{d}\bar{s} \, \mathrm{d}x \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^+} f(z^{-1}(x-y), z^{-1}\bar{s}) \psi(s^{-1}(x-u), s^{-1}\bar{s}) \, \mathrm{d}\bar{s} \, \mathrm{d}x \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^+} f(x^*, s^*) \psi(s^{-1}(zx^* + y - u), s^{-1}zs^*) \, \mathrm{d}(zs^*) \, \mathrm{d}(z^d x^*) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^+} f(x^*, s^*) \psi(s^{-1}z(x^* - z^{-1}(u-y)), s^{-1}zs^*) \, \mathrm{d}(zs^*) \, \mathrm{d}(z^d x^*) \\
&= z^{d+1} \int_{\mathbb{R}^d} \int_{\mathbb{R}^+} f(x^*, s^*) \mathcal{L}_{z^{-1}s} \psi(x^* - z^{-1}(u-y), s^*) \, \mathrm{d}s^* \, \mathrm{d}x^* \\
&= z^{d+1}[f \tilde{\star}_G \psi](z^{-1}(u-y), z^{-1}s) \\
&= z^{d+1} \mathcal{L}_{y,z}[f \tilde{\star}_G \psi](u,s).
\end{aligned}
\tag{46}
$$

Similarly to the lifting correlations case, we obtain equivariance, up to a multiplicative front factor resulting from the correction factor of the Lebesgue measure via the scaling z, z^{d+1}. Since, once again, the equivariance condition must be valid for any input $f \in \mathbf{L}^2(\mathbb{R}^d \rtimes \mathbb{R}^+)$, we must then find constraints on parametrization of the filter ψ, such that exact equivariance is warranted. This is obtained by parametrizing the (group) filter $\psi(x,\bar{s})$ as a scaled-normalized function $\psi_{\bar{s}}(x,\bar{s}) = \frac{1}{\bar{s}^{d+1}} \psi(x,\bar{s})$. Consequently, one can conclude that exact equivariance is obtained by the group correlation with respect to the left-regular representations $\mathcal{L}_g[\cdot]$ if and only if the set of transformed filters correspond to a dictionary of *re-normalized group wavelet time-frequency atoms* $\mathcal{D}$ defined as:

$$\mathcal{D} = \left\{ \psi_{u,s}(x,\bar{s}) = \frac{1}{s^{d+1}} \psi\left(\frac{x-u}{s}, s^{-1}\bar{s} \right) \right\}_{u \in \mathbb{R}, s \in \mathbb{R}^+}, \tag{47}$$

where the wavelet $\psi_{0,1}(x,\bar{s}) = \psi(u,\bar{s})$ can be regarded as the *re-normalized (group) mother wavelet*. Due to the similarity of this parametrization with wavelets, we refer to this parametrization as *re-normalized (group) wavelets*.

D Experimental Details

D.1 Implementation with a Discrete Scale Grid

A subtlety arises with respect to integrating over scale when implementing the continuous theory in a discrete setting that is suitable for numerical computations. The group convolutions include scale correction factors as part of the Haar measure, which makes the integration invariant to translations along the scale axis. That is, the integral of a signal $f(s)$ over scale is the same as that of the same signal $f(z^{-1})$, whose scale is shifted by a factor $z \in \mathbb{R}^+$:

$$\int_{\mathbb{R}^+} f(z^{-1}s) \frac{1}{s} \mathrm{d}s \stackrel{s \to zs}{=} \int_{\mathbb{R}^+} f(z^{-1}s) \frac{1}{zs} \mathrm{d}zs = \int_{\mathbb{R}^+} f(s) \frac{1}{s} \mathrm{d}s.$$

We can translate the scale integration to the discrete setting via Riemann integrals, where we sample the function on a grid and take the weighted sum of these values, such that each sampled point is

---

[5]When dealing with functions on groups one usually relies on left-invariant Haar measures for integration. Such measures have the property $\mathrm{d}(gh) = \mathrm{d}h$ for every $g, h \in G$. The Haar measure on the dilation-translation group in terms of the Lebesgue measure on $\mathbb{R}^d \rtimes \mathbb{R}^+$ is given by $\mathrm{d}g = \frac{1}{s^{d+1}} \mathrm{d}x \mathrm{d}s$, with $g = (x,s)$. Consequently, the factor $\frac{1}{s^{d+1}}$ is omitted when the group correlation is defined via inner products on the group, as is the case in Eq. 17, and equivariance via the left-regular representations is directly obtained [3, Thm. 1].
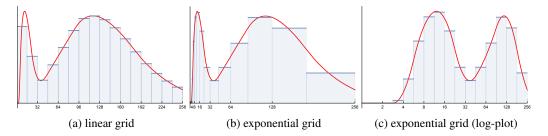
(a) linear grid	(b) exponential grid	(c) exponential grid (log-plot)

Figure 4: Riemann integration of functions on $\mathbb{R}^+$ using linear (a) and exponential grids (b,c).

weighted with the bin-width. That is:

$$\int_{\mathbb{R}^+} f(s) \tfrac{1}{s} \mathrm{d}s \approx \sum_{i}^{N} f(s_i) \tfrac{1}{s_i} \Delta_i.$$

When the scale grid is linear, the bin-widths Δ_i are constant, as depicted in Fig. 4a. When the scale grid is exponential, e.g., $s_i = b^{i-1}$ with b some base factor, the bin widths are proportional to the scale values at the grid points, i.e., $\Delta_i \propto s_i$. This is illustrated in Fig. 4b. In this setting, the factor $\frac{1}{s_i}$ cancels out (up to some constant) with the bin width Δ_i, and integration is simply done by summing the values sampled on the scale grid. *Hence, when working with an exponential grid along the scale axis, the factor in the group convolutions (Eq. 23) becomes $\frac{1}{s^d}$ instead of $\frac{1}{s^{d+1}}$.*

Finally, it is worth mentioning that working with an exponential grid is the natural thing to do when dealing with the dilation group. It is a multiplicative group with a natural distance between group elements $z, s \in \mathbb{R}^+$ defined by $\| \log z^{-1}s \|$. Then, on an exponential grid, the grid points are spaced uniformly with respect to this distance, as illustrated in Fig. 4c.

D.2   The Number of Scales per Layer

Wavelet Networks posses two additional hyperparameters in comparison to conventional CNNs: the amount of scales N_s to be considered at each layer and the base scale factor $b > 1 \in \mathbb{R}$. Importantly, the amount of scales one can consider is tightly related to the spatial dimension of the input itself. Consequently, one can leverage the spatial dimension of the signal as a proxy for the choice of the number of scales.

To illustrate this, consider a signal $f(z)$ which is sampled on a discrete grid $z \in [1, N_f] \subset \mathbb{Z}$ with N_f samples, i.e., $f : [1, N_f] \subset \mathbb{Z} \to \mathbb{R}$, and a convolutional filter ψ defined via a continuous parametrization $\psi : \mathbb{R} \to \mathbb{R}$, and sampled on a discrete grid $z \in [1, N_\psi] \subset \mathbb{Z}$ with N_ψ samples, i.e., $\psi : [1, N_\psi] \subset \mathbb{Z} \to \mathbb{R}$. Provided an adequate parametrization of ψ, when re-scaling it, one is restricted at the bottom of the scale axis by the Nyquist criterion, and at the top of the scale by the scale for which the filter becomes constant in an interval of N_f samples. The Nyquist criterion is required to avoid aliasing and intuitively restricts us to a compression factor on ψ such that it becomes as big as 2 grid samples. On the other hand, by having ψ re-scaled to an extreme to which it is constant in the support of the input signal f, its convolution will only perform blurring-like operations on the input. Importantly, an increased computational cost associated with the number of scales considered.[6] Consequently, we need to be selective about the scales considered in our model. We reason that selecting scales for which the sampled support of ψ is smaller than N_ψ is non-optimal, as any of these functions can be described (and learned) on the original support N_ψ of ψ as well. Consequently, we restrict our scale grid to a minimum scale of 1. On the other scale direction, we reason that utilizing scales for which the support of the filter overpasses that of the input, i.e., $N_f \leq N_\psi$, is non-optimal as well, as the values outside of the region $[1, N_f]$ are unknown and pattern matching becomes, in the best case scenario, only approximate. As a result, we consider the set of sensible scales to be given by the interval $[1, \frac{N_f}{N_\psi}]$.

In this work, we parametrize the kernels via B-splines, which allow for efficient implementations and which have finite support [3].[7] The spatial part of the filters are sampled on a zero-centered grid

---

[6]The memory requirements of our approach grow linearly with the number of scales considered.

[7]Other options can also be incorporated in our approach, such as the Hermite polynomials used in [62]

$[-\lfloor N_\psi/2 \rfloor, \lfloor N_\psi/2 \rfloor] \in \mathbb{Z}$. In the lifting and group convolutions, the filters are sampled at a range of scale factors on a grid that covers the support of the kernel. Thus, for a scale factor s, the function $\psi(z/s)$ is sampled on the grid $[-\lfloor sN_\psi/2 \rfloor, \lfloor sN_\psi/2 \rfloor] \in \mathbb{Z}$. Consequently, it is recommended to use a maximum scale factor such that the kernel is smaller than the input signal, i.e., $sN_\psi < N_f$.

Based on findings in several years of research in wavelet analysis, we select a dyadic dilation set as the basis b of our approach [44, 17]. In terms of the dyadic dilation set used in our experiments $\{2^j\}_{j=j_0}^{j_{max}}$, $j \in \mathbb{Z}$, this corresponds to scales given by the interval $[0, j_{max}$ s.t. $N_\psi\, 2^{j_{max}} \leq N_f]$. Unfortunately, this set is still very large and becomes computationally prohibitive at this point in time. Exemplarily, even the W-n Networks (Appx. D.3, Tab. 2), which utilize unconventionally large filters at the first layer, i.e., 79, require a dyadic set $[0, 10]$. This corresponds to a memory increment of 11 times that of the conventional convolutional layer. To alleviate this, we restrict the scales at the first (lifting) layer to 9 scales, and hence, to the set of scales $\{2^j\}_{j=0}^{8} = [1, 2, 4, \ldots, 256]$. Neural networks usually reduce the spatial dimension N_f of the input f as a function of depth. Following the rationale outlined before, we take advantage of this fact to reduce the number of scales as a function of the depth as well. To this end, we utilize the pooling factor as a proxy for the computation of the amount of scales that can be disregarded. As an example, a pooling of 8 can be interpreted as a reduction on the set of feasible scales by 2 in a dyadic set.[8] Consequently, the memory cost of the layers in our network is reduced as a function of depth.

We strongly believe that progress in computational infrastructures will allow for our system to be deployed with many more scales as those provided in our experiments. As a result, these developments could lead to enhanced results in our network structure without requiring an increment in the sample complexity of the model.

D.3 Training Regimes

Whenever possible, we utilized existing code for the baselines of our wavelet networks as an starting point for the general infrastructure of our model. Specifically, we utilize the PyTorch implementation provided in https://github.com/philipperemy/very-deep-convnets-raw-waveforms and https://github.com/kyungyunlee/sampleCNN-pytorch as baseline for the US8K experiments of Dai et al. [16], and the MTAT experiments of Lee et al. [34], respectively. By doing so, we aim to preserve the reproducibility of the experiments in the corresponding baseline papers during our own experiments, as some important training factors are not specified in the baseline papers, e.g., the learning rate utilized in Dai et al. [16]. Unfortunately, we could not find code online for Abdoli et al. [1], and thus we were forced to interpret some of the ambiguities in the paper, e.g., the pooling type utilized in the pooling layers and the implementation of the utilized loss. Any omitted parameters can safely be considered to be the default values in PyTorch 1.5.0. All our experiments are carried out in a Nvidia TITAN RTX GPU.

W-n Networks. We utilize a sampling rate of 22.05kHz as opposed to the 8kHz of Dai et al. [16], as early studies in the data indicated that some classes become indistinguishable for the human ear in this regime.[9] We zero-pad signals shorter than 4 seconds so that all input signals have a constant length of 80200 samples. Following the online implementation provided above, we utilize the Adam optimizer [29] with lr = 1e-2 and weight_decay = 1e-4, and perform training on the official first 9 folds and test on the 10th fold. Though the provided results are overoptimistic, one does see that the relative improvement of our approach is consistent over the baseline architectures (Tab. 1, column "Acc.(%) [16]"). During our experiments, we saw that reducing the learning rate from 1e-2 to 1e-3 further increased the performance of our W-Nets. Consequently, the reported results of the W-Net variants are obtained with this learning rate. We utilize mini-batches of size 16 and perform training for 400 epochs. The learning rate is reduced by half after 20 epochs of no improvement in the validation loss. Furthermore, we provide 10-cross fold validation results to provide fair comparisons with other network architectures (Tab. 1, column "(%) 10-Fold"). Our proposed wavelet network variants of the Mn-networks of [16] are provided in Table 2 (see [16, Tab. 1] for comparison).

W-1DCNN. Following Abdoli et al. [1], we utilize a sampling rate of 16kHz during our experiments. We zero-pad signals shorter than 4 seconds so that all input signals have a constant length of 64000 samples. Following the experimental description of the paper, we utilize the AdaDelta optimizer [81]

---

[8]This corresponds to 2^{3-1}. Note that trivial scale 1 is always preserved.

[9]Some examples of this behaviour are provided in experiments/UrbanSound8K/data_analysis.ipynb.

Table 2: Proposed wavelet network variants of the Mn-Networks of Dai et al. [16]. W3 (0.219M) denotes a 3-layered network with a total of 0.219M parameters. $[79/4, 150, 3]$ denotes a group convolutional layer with a nominal receptive field of 79 samples, 150 filters and 3 scales, with a stride of 4. Stride is omitted for stride 1 (e.g., $[3, 150, 3]$ has stride 1). Each convolutional layer uses batch normalization right after the convolution, after which ReLU is utilized as activation function. Following the findings of [56, Appx. C] on the influence of stride in the equivariance of the network, we replace strided convolutions by normal convolutions, followed by spatial pooling of the same size as the stride. $[\ldots] \times k$ denotes k stacked layers and double layers in brackets denote residual blocks as defined in [16, Fig. 1b]. In each of the levels of convolutional layers and residual blocks, the first convolution of the first block has scale 3 and the remaining convolutional layers at that level has scale 1.

W3 (0.219M)	W5 (0.558M)	W11 (1.806M)	W18 (3.759M)	W34 (4.021M)
INPUT: 80200X 1 TIME-DOMAIN WAVEFORM				
Lifting Layer (9 scales)				
$[79/4, 150]$	$[79/4, 74]$	$[79/4, 51]$	$[79/4, 57]$	$[79/4, 45]$
MAXPOOL: 4x1 (OUTPUT: 80200X 9X 1)				
$[3, 150, 3]$	$[3, 74, 3]$	$[3, 51, 3] \times 2$	$[3, 57, 3] \times 4$	$\begin{bmatrix} 3, 45 \\ 3, 45 \end{bmatrix} \times 3$
MAXPOOL: 4x1 (OUTPUT: 80200X 7X 1)				
	$[3, 148, 3]$	$[3, 102, 3] \times 2$	$[3, 114, 3] \times 4$	$\begin{bmatrix} 3, 90 \\ 3, 90 \end{bmatrix} \times 4$
MAXPOOL: 4x1 (OUTPUT: 80200X 5X 1)				
	$[3, 296, 3]$	$[3, 204, 3] \times 3$	$[3, 228, 3] \times 4$	$\begin{bmatrix} 3, 180 \\ 3, 180 \end{bmatrix} \times 6$
MAXPOOL: 4x1 (OUTPUT: 80200X 3X 1)				
		$[3, 408, 3] \times 2$	$[3, 456, 3] \times 4$	$\begin{bmatrix} 3, 360 \\ 3, 360 \end{bmatrix} \times 3$
GLOBAL AVERAGE POOLING (OUTPUT: 1 X N)				
SOFTMAX [110] (OUTPUT: 1 X N)				

with `lr = 1.0` and perform training in a 10-fold cross validation setting as described in Sec. 5. We utilize mini-batches of size 100 and perform training for 100 epochs. We utilize the 50999-1DCNN variant of [1], as it is the variant that requires the less human engineering.[10] Unfortunately, we were not able to replicate the results reported in Abdoli et al. [1], i.e., $83 \pm 1.3\%$ accuracy, in our experiments. Our experiments indicated a 10-cross fold accuracy of $62 \pm 1.3\%$, a difference of 21% mean accuracy in comparison to the reported results. We experiment with our interpretation of the mean squared logarithmic error (MSLE) loss function defined in [1, Eq. 4]. However, we find that the conventional cross-entropy loss delivers better results than the proposed MSLE loss. Consequently, all our reported results are based on training with this loss.[11] Despite this large drop in accuracy, we see that our wavelet networks are comparatively better than the baseline architecture. Our proposed wavelet network variant of the 50999-1DCNN of [1] is provided in Table 3 (see [1, Tab. 1] for comparison).

---

[10]The remaining architectures partition the input signal into overlapping windows after which the predictions of each windows are summarized via a voting mechanism (see Abdoli et al. [1] for details). Consequently, one could argue that the 50999-1DCNN is the only variant that truly receives the raw waveform signal. Nevertheless it is not clear from the paper how the input signal of 64000 samples is reduced to 50999 samples, which is the input dimension of the raw signal for this architecture type.

[11]The MSLE loss in [1, Eq. 4] is defined as $\frac{1}{N} \sum_{i=1}^{N} \log \frac{p_i+1}{a_i+1}^2$, where p_i, a_i and N are the predicted class, the actual class, and the number of samples respectively. Note, however, that obtaining the predicted class p_i, i.e., $p_i = \mathrm{argmax}_o f(x_i)$, where $f(x_i) \in \mathbb{R}^O$ is the output of the network for a classification problem with O classes and input x_i, is a non-differentiable function. Consequently, it is not possible to train the network based on the formulation provided in [1, Eq. 4]. In order to train our model with this loss, we re-formulate the MSLE loss as $\frac{1}{N} \sum_{i=1}^{N} \sum_{o=1}^{O} \log \frac{p_{i,o}+1}{a_{i,o}+1}^2$, where $\{a_{i,o}\}_{o=1}^{O}$ is a one-hot encoded version of the label a_i. In other words, our formulation of the loss measures the difference between the one-hot encoded label and the output of the network.

W-1DCNN (0.549M)
INPUT: 64000X 1
Lifting Layer (9 scales)
$[63/2, 12]$
MAXPOOL: 8x1
$[31/2, 24, 3]$
MAXPOOL: 8x1
$[15/2, 48, 3]$ $[7/2, 96, 3]$ $[3/2, 408, 3]$
MAXPOOL: 5x1
FLATTEN $196 \times 6 \rightarrow 1152$ FC: $[1152, 96]$ FC: $[96, 48]$ FC: $[48, 10]$
SOFTMAX

Table 3: Proposed wavelet network variant of the 50999-1DCNN of Abdoli et al. [1]. $[31/2, 24, 3]$ denotes a group convolutional layer with a nominal receptive field of 31 samples, 24 filters and 3 scales, with a stride of 2. FC: $[96, 48]$ denotes a fully-connected layer with 96 input channels and 48 output channels. Each convolutional layer uses batch normalization right after the convolution, after which ReLU is utilized as activation function. All fully connected layers (expect for the last one) utilize dropout with a dropout rate of 0.25 and ReLU as activation function. The last one does not use either dropout nor an activation function. Following the findings of [56, Appx. C] on the influence of stride in the equivariance of the network, we replace strided convolutions by normal convolutions, followed by spatial pooling of the same size as the stride. We note that the input size of our network is (presumably) different from that in [1]. Consequently, the last pooling layer utilizes a region of 5, in contrast to 4 as used in [1]. However, as it is not clear how the input dimension is reduced from 64000 to 50999, we stick to their original sampling procedure. Moreover, we interpret their poling layers as max-pooling ones.

W3^9-Network. For the experiments in the MTAT dataset, we utilize the PyTorch code provided by the authors. We use the same data and tag preprocessing as those used in [34]. We utilize the SGD optimizer with `lr=1e-2`, `weight_decay=1e-6` and `nesterov=True`. We utilize mini-batches of size 23 and perform training for 100 epochs. The learning rate is reduced by 5 after 3 epochs of no improvement in the validation loss. Early stopping is used whenever the learning rate drops under `1e-7`. Unfortunately, we were not able to replicate the per-class AUC results reported in [34], i.e., 0.9055 in our experiments. Our experiments indicated a per-class AUC of 0.893. Following the convention in literature, we report the per-class and per-sample average AUC as well as mean average precision (MAP). Our results show that our wavelet networks are consistently comparatively better than the baseline architecture. Our proposed wavelet network variant of the 3^9-Net of [34] is provided in Table 4 (see [34, Tab. 1] for comparison).

W-3^9 NET (2.404M)
INPUT: 59049X 1
Lifting Layer (9 scales)
$[3/3, 90]$
$[3/1, 90, 3]$, MP:3x1 $[3/1, 90, 1]$, MP:3x1 $[3/1, 180, 1]$, MP:3x1 $[3/1, 180, 3]$, MP:3x1 $[3/1, 180, 1]$, MP:3x1 $[3/1, 180, 1]$, MP:3x1 $[3/1, 180, 3]$, MP:3x1 $[3/1, 180, 1]$, MP:3x1 $[3/1, 360, 1]$, MP:3x1 $[3/1, 360, 3]$
FC: $[360, 50]$
SIGMOID

Table 4: Proposed wavelet network variant of the 3^9-Net of Lee et al. [34]. $[3/1, 90, 3]$ denotes a group convolutional layer with a nominal receptive field of 3 samples, 90 filters and 3 scales, with a stride of 1. MP:3x1 denotes max-pooling with on a region of 3. FC: $[360, 50]$ denotes a fully-connected layer with 360 input channels and 50 output channels. Each convolutional layer uses batch normalization right after the convolution, after which ReLU is utilized as activation function. Dropout with a dropout rate of 0.5 is used after the 6th and 11th layer. Following the findings of [56, Appx. C] on the influence of stride in the equivariance of the network, we replace strided convolutions by normal convolutions, followed by spatial pooling of the same size as the stride.

D.4 Additional Experiments

Bearing fault detection. We further validate our approach for the task of condition monitoring in induction motors. We aim to classify healthy and faulty bearings from raw data provided by Semiotic Labs BV.[12] The dataset consists of 246 clips each of which is 15 seconds long clips and sampled at 20kHz. This corresponds to a total of 1.03 hours of recording. The dataset is slightly unbalanced containing 155 healthy and 91 faulty recordings $[155, 91]$. The dataset is previously split into a training set of $[85, 52]$ and a test set of $[70, 39]$ samples, respectively. These splits are provided

---

[12] https://www.semioticlabs.com/

ensuring that measurements from the same motor are not included both in the train and the test set. We utilize 20% of the training set as a validation set for the selection of the best model. Each of the clips is composed of 6 channels measuring both current and voltage of the 3 poles of the motor. Based on the success of the M-11 and W-11 Networks in the US8K dataset, we utilize variants of these architectures in our experiments, equivalent to the original ones up to the first and last layer. Specifically, the lifting layer receives inputs with 6 channels and the output of the network consists of two channels.

We compare our wavelet network with the CNN baseline similar in the amount of layers and total parameters. Our results show that our wavelet network variant outperforms the CNN baseline for this task as well (Tab. 5). Consequently, we conclude that our approach seems to generalize to other kinds of data as well, and thus, that our approach seems to be a promising research direction for learning on general time-series.

Model	Acc.(%)	Param.
M11-Net	65.1376	1.806m
W11-Net	**68.8073**	1.823m

Table 5: Experimental results on condition monitoring. We see that our wavelet network variant outperforms the CNN baseline for this task as well.