On the Shift Invariance of Max Pooling Feature Maps in Convolutional Neural Networks
Hubert Leterme, Kévin Polisano, Valérie Perrier, Karteek Alahari

To cite this version:
Hubert Leterme, Kévin Polisano, Valérie Perrier, Karteek Alahari. On the Shift Invariance of Max Pooling Feature Maps in Convolutional Neural Networks. 2022. hal-03779434

HAL Id: hal-03779434
https://hal.archives-ouvertes.fr/hal-03779434
Preprint submitted on 16 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
On the Shift Invariance of Max Pooling Feature Maps in Convolutional Neural Networks

Hubert Leterme, Kévin Polisano, Valérie Perrier, and Karteek Alahari

Abstract—In this paper, we aim to improve the mathematical interpretability of convolutional neural networks for image classification. When trained on natural image datasets, such networks tend to learn parameters in the first layer that closely resemble oriented Gabor filters. By leveraging the properties of discrete Gabor-like convolutions, we prove that, under specific conditions, feature maps computed by the subsequent max pooling operator tend to approximate the modulus of complex Gabor-like coefficients, and as such, are stable with respect to certain input shifts. We then compute a probabilistic measure of shift invariance for these layers. More precisely, we show that some filters, depending on their frequency and orientation, are more likely than others to tend to approximate the modulus of complex Gabor-like coefficients, which is known to be nearly shift invariant.

We first consider an operator computing the modulus of real-Gabor-max-pooling (RPool) and a part of the above Gabor-like convolutions and computes their maximum value over a sliding discrete grid. We refer to this as complex-Gabor-modulus (CGMod). We then prove that, under additional conditions on the filter’s frequency and orientation, CGMod is stable with respect to small input shifts.

B. Proposed Approach

We first consider an operator which only keeps the real part of the above Gabor-like convolutions and computes their maximum value over a sliding discrete grid. We refer to this as a real-Gabor-max-pooling (RPool) extractor. We then prove that, under additional conditions on the filter’s frequency and orientation, CGMod and RPool produce comparable outputs. We deduce a measure of shift invariance for RPool operators, which benefit from the stability of CGMod.
Next, we show that, after training with ImageNet, the feature extractor formed by early layers of popular CNN architectures can approximately be reformulated as a stack of RGPool operators. Our framework therefore provides a theoretical grounding to study these networks.

We apply our theoretical results on the dual-tree complex wavelet packet transform (DT-CWPT), a particular case of discrete Gabor-like decomposition with perfect reconstruction properties [29], [30], possessing characteristics comparable to standard convolution layers. Finally, we verify our predictions on a deterministic setting based on DT-CWPT. Given an input image, we compute the mean discrepancy between the outputs of CGMod and RGPool, for each wavelet packet filter.\(^1\) We then observe that shift invariance, when measured on RGPool feature maps, is nearly achieved if they remain close to CGMod outputs. We therefore establish an invariance validity domain for RGPool operators.

Prior to this work, we presented a preliminary study [31], where we experimentally showed that an operator based on DT-CWPT can mimic the behavior of the first convolution layer with fewer parameters, while keeping the network’s predictive power. Our model was solely based on real-valued filters,\(^2\) which are known to be generally unstable [32]. Yet, we observed a limited but genuine form of shift invariance, compared to other models based on the standard, non-analytic wavelet packet transform. At the same time, we became aware of a preliminary work in Waldspurger’s PhD thesis [28, pp. 190–191], suggesting a potential connection between the combinations “real wavelet transform + max pooling” on the one hand and “complex wavelet transform + modulus” on the other hand. Following this idea, we decided to study whether invariance properties of complex moduli could somehow be captured by the max pooling operator. As shown in the present paper, Waldspurger’s work does not fully extend to discrete and subsampled convolutions. We address this issue by adopting a probabilistic point of view.

II. RELATED WORK

A. Wavelet Scattering Networks

These models, introduced by Bruna and Mallat [11], compute cascading wavelet convolutions followed by non-linear operations. They produce translation-invariant image representations which are stable to deformation and preserve high-frequency information. A variation has been proposed in [33] to improve stability with respect to small rotations. Wavelet scattering networks were later adapted to the discrete framework using the dual-tree complex wavelet transform [34], as well as functions defined on graphs [35].

Such networks, which are totally deterministic aside from the output classifier, achieve results on small image datasets but do not scale well to more complex ones. According to Oyallon et al. [36], [37], this is partly due to non-geometric sources of variability within classes. Instead, the authors proposed to use scattering coefficients as inputs to a CNN, showing that the network complexity can be reduced while keeping competitive performance. More recent work by Zarka et al. [38] proposed to sparsify wavelet scattering coefficients by learning a dictionary matrix, and managed to outperform AlexNet [10]. This was extended by the same team in [39], where the authors proposed to learn \(1 \times 1\) convolutions between feature maps of scattering coefficients and to apply soft-thresholding to reduce within-class variability. This model reached the classification accuracy of ResNet-18 on ImageNet.

Other work proposed architectures in which the scattering transform is no longer deterministic. Cotter and Kingsbury [40] built a learnable scattering network. In this model, feature maps of scattering coefficients are mixed together using trainable weights, to account for cross-channel filtering as implemented in CNNs. Their architecture outperformed VGG networks on small image datasets. Recently, Gauthier et al. [41] introduced parametric scattering networks, in which the scale, orientation and aspect ratio of each wavelet filter are adjusted during training. Their approach has proven successful when trained on limited dataset.

All these papers are driven by the purpose of building ad-hoc CNN-like feature extractors, implementing well defined mathematical operators specifically designed to meet a certain number of desired properties. By contrast, our work seeks evidence that such properties, which have been established for wavelet scattering networks, are—to some extent—embedded in existing CNN architectures, with no need to alter their behavior or introduce new features.

B. Invariance Studies in CNNs

Several papers analyze invariance properties in CNN-related feature extractors, including—but not limited to—wavelet scattering networks. Whereas extensive studies related to the original architecture are proposed by Mallat in [42], [43], more recent work tackle the question for various extensions of the model. In [44], [45], scattering networks based on uniform covering frames—i.e., frames splitting the frequency domain into windows of roughly equal size, much like Gabor frames—are studied. Besides, [8] considers a wide variety of feature extractors involving convolutions, Lipschitz-continuous nonlinearities and pooling operators. The paper shows that outputs become more translation invariant with increasing network depth. Finally, [46] shows that certain classes of CNNs are contained into the reproducing kernel Hilbert space (RKHS) of a multilayer convolutional kernel representation. As such, stability metrics are estimated, based on the RKHS norm which is difficult to control in practice.

In these studies, invariance properties are obtained for continuous signals. Whereas real-life CNNs can be mathematically described in the continuous framework, feature maps computed at their hidden and output layers are actually discrete sequences, which can be recovered by sampling the continuous signals. At each convolution and pooling layers, the sampling interval is increased (subsampling), resulting in a loss of information. Unfortunately, this may greatly affect shift invariance, as explained in section I. The current paper specifically addresses this issue.
III. Shift Invariance of Operators

The goal of this section is to theoretically establish conditions for near-shift invariance for the output of the first max pooling layer. We start by proving shift invariance of $\mathbb{R}$GPool and $\mathbb{C}$GMod produce closely related outputs. Finally, we derive a probabilistic measure of shift invariance for $\mathbb{R}$GPool.

A. Notations

The complex conjugate of any number $z \in \mathbb{C}$ is denoted by $\bar{z}$. For any $p \in \mathbb{R}_+^* \cup \{\infty\}$, $x \in \mathbb{R}^2$ and $r \in \mathbb{R}_+$, we denote by $B_p(x, r) \subset \mathbb{R}^2$ the closed $p$-ball with center $x$ and radius $r$. When $x = 0$, we write $B_p(r)$.

Continuous Framework: Given $p > 0$ and a measurable subset of $\mathbb{R}$ or $\mathbb{R}^2$ denoted by $E$, we consider $L^p(E)$ as the space of measurable complex-valued functions $f : E \to \mathbb{C}$ such that $\|f\|_{L^p} := \int_E |f(x)|^p \, dx < +\infty$. Whenever we talk about equality in $L^p(E)$ or inclusion in $E$, it shall be understood as “almost everywhere with respect to the Lebesgue measure”. Besides, we denote by $L^2_b(\mathbb{R}^2) \subset L^2(\mathbb{R}^2)$ the subset of real-valued functions. For any $f \in L^2(\mathbb{R}^2)$, $\hat{f}$ denotes its flipped version: $\hat{f}(x) := f(-x)$.

The 2D Fourier transform of any $f \in L^2(\mathbb{R}^2)$ is denoted by $\hat{f} \in L^2(\mathbb{R}^2)$, such that

$$\forall \nu \in \mathbb{R}^2, \; \hat{f}(\nu) := \int_{\mathbb{R}^2} f(x) e^{-i\langle \nu, x \rangle} \, dx. \quad (1)$$

For any $\varepsilon > 0$ and $\nu \in \mathbb{R}^2$, we denote by $V(\nu, \varepsilon) \subset L^2(\mathbb{R}^2)$ the set of functions whose Fourier transform is supported in a square region of size $\varepsilon \times \varepsilon$ centered in $\nu$:

$$V(\nu, \varepsilon) := \left\{ \psi \in L^2(\mathbb{R}^2) \mid \text{supp} \hat{\psi} \subset B(\nu, \varepsilon/2) \right\}. \quad (2)$$

For any $h \in \mathbb{R}^2$, we also consider the translation operator, denoted by $T_h$, defined by $T_h f : x \mapsto f(x-h)$.

Discrete Framework: We consider $l^2(\mathbb{Z}^d)$ as the space of $d$-dimensional sequences $X \in \mathbb{C}^{\mathbb{Z}^d}$ such that $\|X\|_2^2 := \sum_{n \in \mathbb{Z}^d} |X[n]|^2 < +\infty$. Indexing is made between square brackets: $\forall X \in l^2(\mathbb{Z}^d), \forall n \in \mathbb{Z}^d, X[n] \in \mathbb{C}$, and we denote by $l^2_b(\mathbb{Z}^d) \subset l^2(\mathbb{Z}^d)$ the subset of real-valued sequences. For any $X \in l^2(\mathbb{Z}^d)$, $\hat{X}$ denotes its flipped version: $\hat{X}[n] := X[-n]$. The subsampling operator is denoted by $\downarrow$: for any $X \in l^2(\mathbb{Z}^d)$ and any $m \in \mathbb{N}^*$, $\downarrow^m X[n] := X[mn]$.

2D images, feature maps and convolution kernels are considered as elements of $l^2(\mathbb{Z}^2)$, and are denoted by straight capital letters. Besides, arrays of 2D convolution kernels are denoted by bold straight capital letters, for instance: $X = (X_k)_{k \in \{0, K-1\}}$. Note that indexing starts at 0 to comply with practical implementations. We will also consider 1D sequences $x \in l^2(\mathbb{Z})$, denoted by straight lower-case letters.

The 2D discrete-time Fourier transform of any $X \in l^2(\mathbb{Z}^2)$ is denoted by $\hat{X} \in l^2([-\pi, \pi]^2)$, such that

$$\forall \xi \in [-\pi, \pi]^2, \; \hat{X}(\xi) := \sum_{n \in \mathbb{Z}^2} X[n] e^{-i\langle \xi, n \rangle}. \quad (3)$$

For any $\kappa \in [0, 2\pi]$ and $\xi \in B_\infty(\kappa)$, we denote by $G(\xi, \kappa) \subset l^2(\mathbb{Z}^2)$ the set of 2D sequences whose Fourier transform is supported in a square region of size $\kappa \times \kappa$ centered in $\xi$:

$$G(\xi, \kappa) := \left\{ W \in l^2(\mathbb{Z}^2) \mid \text{supp} \hat{W} \subset B_\infty(\xi, \kappa/2) \right\}. \quad (4)$$

Remark 1: The support $B_\infty(\xi, \kappa/2)$ actually lives in the quotient space $[-\pi, \pi]^2/(2\pi \mathbb{Z}^2)$. Consequently, when $\xi$ is close to an edge, a fraction of this region is located at the far end of the frequency domain. From now on, the choice of $\xi$ and $\kappa$ is implicitly assumed to avoid such a situation.

B. Intuition

In many CNNs for computer vision, input images are first transformed through subsampled—or strided—convolutions. For instance, in AlexNet, convolution kernels are of size $11 \times 11$ and the subsampling factor is equal to 4. Fig. 1 displays the corresponding kernels after training with ImageNet. This linear transform is generally followed by rectified linear unit (ReLU) and max pooling.

We can observe that many kernels display oscillating patterns with well-defined orientations. We denote by $V \in l^2_R(\mathbb{Z}^2)$ one of these “well-behaved” filters. Its Fourier spectrum roughly consists in two bright spots which are symmetric with respect to the origin. Now, we consider a complex-valued companion $W \in l^2(\mathbb{Z}^2)$ such that, for any $\xi = (\xi_1, \xi_2) \in [-\pi, \pi]^2$:

$$\hat{W}(\xi) := (1 + \text{sgn} \xi_1) \cdot \hat{V}(\xi). \quad (5)$$

We can show that $V$ is the real part of $W$, and that $W = V + iH(V)$, where $H$ denotes the two-dimensional Hilbert transform as introduced in [47]. As a consequence, $\hat{W}$ is equal to $2\hat{V}$ on one half of the Fourier domain, and 0 on the other half. Therefore, only one bright spot remains in the spectrum. It turns out that such complex filters with high frequency resolution produce stable signal representations, as we will see in section III-C. In the subsequent sections, we then wonder whether this property is kept when considering the max pooling of real-valued convolutions.

3Actually, the Fourier transform of any real-valued sequence is centrally symmetric: $\hat{\bar{X}}(\xi) = \hat{X}(-\xi)^\ast$. The specificity of well-oriented filters lies in the concentration of their power spectrum around two precise locations.

4$H(V)$ is defined such that $H(V)(\xi) := -i \text{sgn}(\xi_1) \hat{V}(\xi)$.
In what follows, \( W \) will be referred to as a discrete Gabor-like filter, and the coefficients resulting from the convolution with \( W \) will be referred to as discrete Gabor-like coefficients.

C. Shift Invariance of \( \mathcal{CGMod} \) Outputs

The aim of this section is to show that the modulus of discrete Gabor-like coefficients—i.e., the output of a \( \mathcal{CGMod} \) operator such as introduced in section I-B—is nearly shift-invariant (the meaning of shift invariance will be clarified). This result is hinted in [32] but not formally proven.

1) Continuous Framework: We introduce several results regarding functions defined on the continuous space \( \mathbb{R}^2 \). Near-shift invariance on discrete 2D sequences will then be derived from these results by taking advantage of sampling theorems.

Lemma 1: Given \( \varepsilon > 0 \) and \( \mathbf{v} \in \mathbb{R}^2 \), let \( \psi \in \mathcal{V}(\mathbf{v}, \varepsilon) \) denote a complex-valued filter such as defined in (2). Now, for any real-valued function \( f \in L^2_\mathbb{R}(\mathbb{R}^2) \), we consider the complex-valued function \( f_0 \in L^2(\mathbb{R}^2) \) defined by

\[
 f_0 : x \mapsto (f * \overline{\psi})(x) e^{i \langle \mathbf{v}, x \rangle}.
\]

Then \( f_0 \) is low-frequency, with \( \text{supp} \, \hat{f}_0 \subset B_\infty(\varepsilon/2) \).

\textbf{Proof:} See Appendix A.

On the other hand, the following proposition provides a shift invariance bound for low-frequency functions such as introduced above.

Proposition 1: For any \( f_0 \in L^2_\mathbb{R}(\mathbb{R}^2) \) such that \( \text{supp} \, f_0 \subset B_\infty(\varepsilon/2) \), and any \( h \in \mathbb{R}^2 \) satisfying \( \|h\|_1 \leq \pi / \varepsilon \),

\[
 \|T_h f_0 - f_0\|_{L^2} \leq \alpha(\varepsilon h) \|f_0\|_{L^2},
\]

where we have defined

\[
 \alpha : \tau \mapsto \frac{\|\tau\|_1}{2}.
\]

\textbf{Proof:} See Appendix B.

2) Adaptation to Discrete 2D Sequences: Given \( \kappa \in [0, 2\pi] \) and \( \xi \in B_\infty(\pi) \), let \( W \in \mathcal{G}(\xi, \kappa) \) denote a discrete Gabor-like filter such as defined in (4). For any image \( X \in L^2_\mathbb{Z}(\mathbb{Z}^2) \) with finite support and any subsampling factor \( m \in \mathbb{N}^+ \), we express \( (X * \overline{W}) \downarrow m \) using the continuous framework introduced above, and derive an invariance formula.

For any sampling interval \( s \in \mathbb{R}^+ \), let \( \mathcal{U}_s \) denote the space of 2D functions \( g \in L^2(\mathbb{R}^2) \) such that the support of \( \widehat{g} \) is included in \( B_\infty(\pi/s) \). We consider the following lemma.

Lemma 2: Let \( s > 0 \). For any \( g \in \mathcal{U}_s \) and any \( \omega \in B_\infty(\pi/s) \), we have

\[
 \widehat{g}(\omega) = s \hat{Y}(s \omega),
\]

where \( Y \in L^2(\mathbb{Z}^2) \) is defined such that \( Y[n] := s g(sn) \), for any \( n \in \mathbb{Z}^2 \). Besides, we have the following norm equality:

\[
 \|g\|_{L^2} = \|Y\|_2.
\]

\textbf{Proof:} See Appendix C.

We now consider \( \phi^{(s)} \in L^2_\mathbb{R}(\mathbb{R}^2) \) such that \( \hat{\phi^{(s)}} := s \mathbb{1}_{B_\infty(\pi/s)} \). For any \( n \in \mathbb{Z}^2 \), we denote by \( \phi^{(s)}(n) := T_{sn} \phi^{(s)} \) a shifted version of \( \phi^{(s)} \). According to Theorem 3.5 in [48, p. 68], \( \{\phi^{(s)}(n)\}_{n \in \mathbb{Z}^2} \) is an orthonormal basis of \( \mathcal{U}_s \). We then get the following proposition, which draws a bond between the discrete and continuous frameworks.

Proposition 2: Let \( X \in L^2_\mathbb{Z}(\mathbb{Z}^2) \) denote an input image with finite support, and \( W \in \mathcal{G}(\xi, \kappa) \). Considering a sampling interval \( s \in \mathbb{R}^+ \), we define \( f_X \in L^2_\mathbb{R}(\mathbb{R}^2) \) and \( \psi_W \in L^2(\mathbb{R}^2) \) such that

\[
 f_X := \sum_{n \in \mathbb{Z}^2} X[n] \phi^{(s)}(n) \quad \text{and} \quad \psi_W := \sum_{n \in \mathbb{Z}^2} W[n] \phi^{(s)}(n).
\]

Then, \( \psi_W \in \mathcal{V}(\xi/s, \kappa/s) \). Moreover, for all \( n \in \mathbb{Z} \),

\[
 X[n] = s f_X(sn) ; \quad W[n] = s \psi_W(sn),
\]

and, for a given subsampling factor \( m \in \mathbb{N}^+ \),

\[
 (X * \overline{W}) \downarrow m[n] = (f_X * \overline{\psi}_W)(msn).
\]

\textbf{Proof:} See Appendix D.

Proposition 2 introduces a latent subspace of \( L^2_\mathbb{R}(\mathbb{R}^2) \) from which input images are uniformly sampled. This allows us to define, for any \( u \in \mathbb{R}^2 \), a translation operator \( T_u \) on discrete sequences, even if \( u \) has non-integer values:

\[
 T_u X[n] := s T_{su} f_X(sn),
\]

where \( f_X \) is defined in (11). We can indeed show that this definition is independent from the choice of sampling interval \( s > 0 \). Besides, given \( X \in L^2_\mathbb{Z}(\mathbb{Z}^2) \), we have

\[
 \forall k \in \mathbb{Z}^2, T_k X[n] = X[n - k]; \\
 \forall u, v \in \mathbb{R}^2, T_u(T_v X) = T_{u+v} X,
\]

which shows that \( T_u \) corresponds to the intuitive idea of a shift operator. Expressions (15) and (16) are direct consequence of the following lemma, which bonds the shift operator in the discrete and continuous frameworks.

Lemma 3: For any \( X \in L^2_\mathbb{Z}(\mathbb{Z}^2) \) and any \( u \in \mathbb{R}^2 \),

\[
 f_{T_u X} = T_{su} f_X.
\]

\textbf{Proof:} See Appendix E.

We now consider the following corollary to Proposition 2.

Corollary 1: For any shift vector \( u \in \mathbb{R}^2 \), we have

\[
 (T_u X * \overline{W}) \downarrow m[n] = (T_{su} f_X * \overline{\psi}_W)(msn).
\]

\textbf{Proof:} Apply (13) in Proposition 2 with \( X \leftarrow T_u X \), and use Lemma 3 to conclude.
3) Shift Invariance in the Discrete Framework: We consider the following operator, for any $W \in \ell^2(\mathbb{Z}^2)$:

$$C_m[W] : X \mapsto |(X \ast \mathbf{W})|_m.$$  

(19)

When $W \in \mathcal{G}(\xi, \kappa)$, we refer to this as a CGMod operator. For the sake of concision, in what follows we will write $C_m$ instead of $C_m[W]$, when no ambiguity is possible.

We are now ready to state the main result about shift invariance of CGMod outputs.

Theorem 1: Let $W \in \mathcal{G}(\xi, \kappa)$ denote a discrete Gabor-like filter and $m \in \mathbb{N}^*$ denote a subsampling factor. If $\kappa \leq 2\pi/m$, then, for any input image $X \in \ell^2_{\mathbb{R}}(\mathbb{Z}^2)$ with finite support and any translation vector $u \in \mathbb{R}^2$ satisfying $|u|_1 \leq \pi/\kappa$, we have

$$\|C_m(\mathcal{T}_u X) - C_mX\|_2 \leq \alpha(\kappa u)\|C_mX\|_2,$$

where $\alpha$ has been defined in (8).

Proof: The proof of this theorem, which involves Lemmas 1-2, Propositions 1-2 and Corollary 1, is provided in Appendix F.

Interestingly, the reference value used in Theorem 1, i.e., $\|C_mX\|_2$, is fully shift-invariant, as stated in the following proposition.

Proposition 3: Let $W \in \mathcal{G}(\xi, \kappa)$ and $m \in \mathbb{N}^*$. Assuming $\kappa \leq 2\pi/m$, we have, for any $X \in \ell^2_{\mathbb{R}}(\mathbb{Z}^2)$ and any $u \in \mathbb{R}^2$,

$$\|C_m(\mathcal{T}_u X)\|_2 = \|C_mX\|_2.$$

(21)

Proof: See Appendix G.

D. From CGMod to RGPoool

Since CGMod operators are not found in classical CNN architectures, the above result does not applies straightforwardly. Instead, the first convolution layer contains real-valued kernels, and is generally followed by ReLU and max pooling. As shown in section IV, this process can be described as an operator parameterized by $W \in \ell^2(\mathbb{Z}^2)$, defined by

$$\mathcal{R}_{m, q}[W] : X \mapsto \text{MaxPool}_q\left( (X \ast \text{Re } \mathbf{W}) \downarrow m \right),$$

(22)

where MaxPool selects the maximum value over a sliding grid of size $(2q + 1) \times (2q + 1)$, with a subsampling factor of 2. More formally, for any $Y \in \ell^2_{\mathbb{R}}(\mathbb{Z}^2)$ and any $n \in \mathbb{Z}^2$,

$$\text{MaxPool}_q(Y)[n] := \max_{|k|_q \leq q} Y[2n + k].$$

(23)

As hinted in section III-B, an important number of trained convolution kernels exhibit oscillating patterns with various scales and orientations. In such a case, $W \in \mathcal{G}(\xi, \kappa)$ for a certain value of $\xi \in [-\pi, \pi]^2$ and $\kappa \in [0, 2\pi]$, and we refer to $\mathcal{R}_{m, q}[W]$ as an RGPool operator. For the sake of concision, from now on we write $\mathcal{R}_{m, q}$ instead of $\mathcal{R}_{m, q}[W]$, when no ambiguity is possible.

In what follows, we show that, under specific conditions on $\mathbf{W}$, RGPool and CGMod operators produce comparable outputs. We then provide a shift invariance bound for RGPool.

1) Continuous Framework: This paragraph, directly adapted from [28, pp. 190–191], provides an intuition about resemblance between RGPool and CGMod in the continuous framework. As will be highlighted later in this section III-D, adaptation to discrete 2D sequences is not straightforward and will require a probabilistic approach.

We consider an input function $f \in L^2_{\mathbb{R}}(\mathbb{R}^2)$ and a band-pass filter $\psi \in \mathcal{V}(\nu, \epsilon)$. Let us also consider

$$g : (x, h) \mapsto \cos \left( (\nu, h) - \eta(x) \right),$$

(24)

where $\eta$ denotes the phase of $f \ast \psi$. Lemma 1 introduced low-frequency functions $f_0$, with slow variations. Roughly speaking, since $\text{supp} f_0 \subset B_\infty(\epsilon/2)$, we can define a “minimal wavelength” $\lambda f_0 := 2\pi/\epsilon$. Then,

$$\|h\|_2 \ll \frac{2\pi}{\epsilon} \implies f_0(x + h) \approx f_0(x),$$

(25)

which leads to

$$(f \ast \text{Re } \psi)(x + h) \approx |(f \ast \psi)(x)| \cdot g(x, h).$$

(26)

On the other hand, we consider a continuous equivalent of the CGMod operator $C_m[W]$ as introduced in (19). Such an operator, denoted by $C[\psi]$, is defined, for any $f \in L^2_{\mathbb{R}}(\mathbb{R}^2)$, by

$$C[\psi] : x \mapsto |(f \ast \psi)(x)| .$$

(27)

On the other hand, we consider the continuous counterpart of RGPool as introduced in (22). It is defined as the maximum value of $f \ast \text{Re } \psi$ over a sliding spatial window of size $r > 0$. This is possible because $f$ and $\text{Re } \psi$ both belong to $L^2_{\mathbb{R}}(\mathbb{R}^2)$, and therefore $f \ast \text{Re } \psi$ is continuous. Such an operator, denoted by $\mathcal{R}_r[\psi]$, is defined, for any $f \in L^2_{\mathbb{R}}(\mathbb{R}^2)$, by

$$\mathcal{R}_r[\psi] : x \mapsto \max_{\|h\|_2 \leq r} (f \ast \text{Re } \psi)(x + h).$$

(28)

For the sake of concision, the parameter between square brackets is ignored from now on.

If $r < 2\pi/\epsilon$, then (26) is valid for any $h \in B_\infty(r)$. Then, using (27) and (28), we get

$$r \ll \frac{2\pi}{\epsilon} \implies \mathcal{R}_r[f(x) \approx C f(x) \max_{\|h\|_2 \leq r} g(x, h).$$

(29)

Using the periodicity of $g$, we can show that, if $r \geq \pi/\|\nu\|_2$, then $h \mapsto g(x, h)$ reaches its maximum value (1) on $B_\infty(r)$. We therefore get

$$\frac{\pi}{\|\nu\|_2} \leq r < \frac{2\pi}{\epsilon} \implies \mathcal{R}_r[f(x) \approx C f(x).$$

(30)

An exact quantification of the above approximation remains an open question. In the current paper, it will be provided as a conjecture, for the discrete framework.

2) Adaptation to Discrete 2D Sequences: We consider an input image $X \in \ell^2_{\mathbb{R}}(\mathbb{Z}^2)$, a subsampling factor $m \in \mathbb{N}^*$ and a grid half-size $q \in \mathbb{N}^*$. We seek a relationship between

$$Y_{pool} := \mathcal{R}_{m, q}X$$

and

$$Y_{mod} := C_{2m}X,$$

(31)

where $C_{2m}$ and $\mathcal{R}_{m, q}$ have been defined in (19) and (22), respectively. Note that, since max pooling also performs subsampling, both CGMod and RGPool operators, as defined in (31), have a subsampling factor equal to $2m$. 

LETERME et al., ON THE SHIFT INvariance OF MAX POOLING FEATURE MAPS
According to (22) and (23), we have
\[ Y(x, h) = C f(x, h) \]
with
\[ x, n \sim 2msn. \]
On the other hand, we postulate that
\[ R_m, q X[n] = R_m f(x, h) \]
for a certain value of \( q \). From (22) and (23), we have
\[ R_m, q X[n] = \max_{\|k\|_\infty \leq q} \text{Re} \left( (X \ast W) \downarrow m \right)[2n + k]. \]
Therefore, according to (13) in Proposition 2, we get
\[ R_m, q X[n] = \max_{\|k\|_\infty \leq q} (f_X \ast \text{Re} \psi_W)(x, h), \]
with
\[ x, n \sim 2msn \quad \text{and} \quad h_k \sim ms k. \]

By considering \( r_q := ms \left( q + \frac{1}{2} \right) \), we get a variant of (33) in which the maximum is evaluated on a discrete grid of \( (2q + 1)^2 \) elements, instead of the continuous region \( B_\infty(r_q) \). As a consequence, (29) is replaced in the discrete framework by
\[ q \ll 2\pi/(m\kappa) \implies R_m, q X[n] \approx C_{2m} X[n] \max_{\|k\|_\infty \leq q} g_X(x, h), \]
with
\[ g_X : (x, h) \mapsto \cos \left( (\nu, h) - \eta_X(x) \right), \]
with \( \nu := \xi/s, \) and where \( \eta_X \) denotes the phase of \( f_X \ast \psi_W \). Unlike the continuous case, even if the window size \( r_q \) is large enough, the existence of \( k \in \{-q \ldots q\}^2 \) such that \( g_X(x, h) = 1 \) is not guaranteed, as illustrated in Fig. 2 with \( q = 1 \). Instead, we can only seek a probabilistic estimation of the relative quadratic error between \( Y_{\text{pool}} \) and \( Y_{\text{mod}} \).

Approximation (37) implies
\[ q \ll 2\pi/(m\kappa) \implies ||C_{2m} X - R_{m, q} X||_2 \approx ||\delta_{m, q} X||_2 , \]
where \( \delta_{m, q} X \in l_2^2(\mathbb{Z}^2) \) is defined such that, for any \( n \in \mathbb{Z}^2, \)
\[ \delta_{m, q} X[n] := C_{2m} X[n] \left( 1 + \max_{\|k\|_\infty \leq q} g_X(x, h) \right) . \]
Expression (39) suggests that the ratio between the left and right terms can be bounded by a quantity which only depends on the product \( m\kappa \) (subsampling factor \( \times \) frequency localization) and the grid half-size \( q. \)

\[ ||C_{2m} X - R_{m, q} X||_2 \leq (1 + \beta_q(m\kappa)) ||\delta_{m, q} X||_2 , \]
for some function \( \beta_q : \mathbb{R}_+ \rightarrow \mathbb{R}_+ \) to be characterized. This is the goal of the following conjecture.

Conjecture 1: There exists
\[ \beta_q : \mathbb{R}_+ \rightarrow \mathbb{R}_+ \quad \text{satisfying} \quad \beta_q(t) = O(t), \]
independent from the characteristic frequency \( \xi \in [-\pi, \pi]^2 \), such that, for any \( X \in l_2^2(\mathbb{Z}^2) \), (41) is satisfied.

We now seek a probabilistic bound for \( ||\delta_{m, q} X||_2. \)
2) Random Variables: From now on, input $X$ is considered as discrete 2D stochastic processes. In order to “randomize” $f_X$ introduced in (11), we define a continuous stochastic process from $X$, denoted by $F_X$, such that

$$\forall x \in \mathbb{R}^2, F_X(x) := \sum_{n \in \mathbb{Z}^2} X[n] \phi_n^{(x)}(x).$$ (49)

Now, we consider the following stochastic processes, which are parameterized by $X$:

$$M_X := |F_X \ast \overline{\psi}|; \quad H_X := \angle(F_X \ast \overline{\psi}); \quad Z_X := e^{iH_X},$$ (50)

and, for any $k \in \{-q \ldots q\}$,

$$G_{X,k} := \text{Re}(Z_X^* z_k); \quad G_{\max X} := \max_{\|k\|_\infty \leq q} G_{X,k}.$$ (51)

For any $x \in \mathbb{R}^2$, $f_X(x)$ and $\eta_X(x)$, such as introduced in (11) and (38), are respectively drawn from $F_X(x)$ and $H_X(x)$. Then, $\chi_X(x)$ such as introduced in (45) is a realization of $Z_X(x)$. Consequently, according to (44), $g_X(x, h_n)$ is a realization of $G_{X,k}(x)$. Besides, according to the definition of $\text{CGMod}$ in (19) and $x_n$ in (36), Proposition 2 implies that

$$M_X(x_n) = C_{2m} X[n].$$ (52)

We remind that $\xi \in [-\pi, \pi]^2$ and $\kappa \in [0, 2\pi]$ respectively denote the center and size of the Fourier support of $W$, as introduced in section III-C. To compute the expected discrepancy between $Y_{\text{pool}}$ and $Y_{\text{mod}}$, we assume that

$$\|\xi\|_2 \gg 2\pi/M; \quad \|\xi\|_2 \gg \kappa,$$ (53) (54)

where $M \in \mathbb{N}^*$ denotes the support size of input images. These assumptions exclude low-frequency filters from the scope of our study. We then state the following hypotheses, for which a justification is provided in Appendix H.

Hypothesis 1: For any $x \in \mathbb{R}^2$, $Z_X(x)$ is uniformly distributed on $S^1$.

Hypothesis 2: For any $n \in \mathbb{N}^*$ and $x, y_1, \ldots, y_n \in \mathbb{R}^2$, the random variables $M_X(y_i)$ for $i \in \{1 \ldots n\}$ are jointly independent of $Z_X(x)$.

F. Expected Quadratic Error between $\mathbb{R}GPool$ and $\text{CGMod}$

In this section, we propose to estimate the expected value of the stochastic quadratic error $P_X$, defined such that

$$P_X := \|C_{2m}X - R_{m,q}X\|_2 / \|C_{2m}X\|_2.$$ (55)

According to (31), this is an estimation of the relative error between $Y_{\text{mod}}$ and $Y_{\text{pool}}$.

First, let us reformulate $\delta_{m,q}X$, introduced in (40), using the probabilistic framework. According to (44) and (51), we have, for any $n \in \mathbb{Z}^2$,

$$\delta_{m,q}X[n] := C_{2m}X[n] (1 - G_{\max X}(x_n)).$$ (56)

We now consider the stochastic process $Q_X := 1 - G_{\max X}$, and the random variable

$$\tilde{Q}_X := \|\delta_{m,q}X\|_2 / \|C_{2m}X\|_2.$$ (57)

The next steps are as follows: 1) at the pixel level, show that $\mathbb{E}[Q_X(x)^2]$ depends on the filter frequency $\xi$, and remains close to zero with some exceptions; 2) at the image level, show that the expected value of $Q_X^2$ is equal to the latter quantity; 3) use Conjecture 1, which states that $P_X \approx \tilde{Q}_X$, to deduce an upper bound on the expected value of $P_X$.

The first point is established in Proposition 4 below, and the two remaining ones are the purpose of Theorem 2. Proposition 4: Assuming Hypothesis 1, the expected value of $Q_X(x)^2$ is independent from the choice of $x \in \mathbb{R}^2$, and

$$\mathbb{E}[Q_X(x)^2] = \gamma_q(m\xi)^2,$$ (58)

where we have defined

$$\gamma_q : \xi \mapsto \sqrt{\frac{3}{2} + \frac{1}{4\pi} \sum_{i=0}^{N_q-1} \sin \omega_i(q)(\xi) - 8 \sin \frac{\omega_i(q)(\xi)}{2}},$$ (59)

with $\omega_i(q) : \mathbb{R}^2 \mapsto [0, 2\pi]$ being introduced in (48).

Proof: See Appendix J.

We consider an ideal scenario where the $z_k$ are evenly spaced on $S^1$. Then, an order 2 Taylor expansion yields $\gamma_q(\xi) = o(1/q^2)$, meaning that $Q_X$ quickly vanishes when the grid half-size $q$ increases. Fig. 4 displays $\xi \mapsto \gamma_q(m\xi)^2$ for $\xi \in [-\pi, \pi]^2$, with $m = 4$ and $q = 1$ as in AlexNet. We notice that, for the major part of the Fourier domain, $\gamma_q$ remains close to 0. However, we observe a regular pattern of dark regions, which correspond to pathological frequencies where the repartition of $z_k$ is unbalanced.

So far, we established a result at the pixel level. Before stating Theorem 2, we need the following intermediate result.

Proposition 5: We consider the random variable

$$\tilde{S}_X := \|C_{2m}X\|_2.$$ (60)

Under Hypothesis 2, for any $x \in \mathbb{R}^2$,

- $Z_X(x)$ is independent of $\tilde{S}_X$;
- $Z_X(x)$, $M_X(x)$ are conditionally independent given $\tilde{S}_X$.

Proof: See Appendix K.

Finally, Propositions 4 and 5 yield the following theorem. It provides an upper bound on the expected value of the relative quadratic error $P_X^2$, such as defined in (55).

Theorem 2: We assume that Conjecture 1 is true. Then, under Hypotheses 1 and 2, we have

$$\mathbb{E}[P_X^2] \leq (1 + \beta_q(m\kappa))^2 \gamma_q(m\xi)^2,$$ (61)

with $\beta_q$ and $\gamma_q$ being introduced in (42) and (59), respectively.

Proof: See Appendix L.

G. Shift Invariance of $\mathbb{R}GPool$ Outputs

In this section, we present the main theoretical claim of this paper. Based on the previous results, we provide a probabilistic measure of shift invariance for $\mathbb{R}GPool$ operators.

Theorem 3: Let $W \in \mathcal{G}(\xi, \kappa)$ denote a discrete Gabor-like filter, $m \in \mathbb{N}^*$ a subsampling factor and $q \in \mathbb{N}^*$ a grid half-size. We consider a stochastic process $X$ whose realizations
are elements of $I_2^2(\mathbb{Z}^2)$. We assume that Hypotheses 1 & 2 are satisfied.\textsuperscript{8} Besides, we assume Conjecture 1.

Given a translation vector $u \in \mathbb{R}^2$, we consider the following random variable:

$$\tilde{R}_X, u := \|R_{m, q}(T_u X) - R_{m, q}X\|_2 / \|C_{2m}X\|_2.$$  \hfill (62)

Then, under the following conditions:

$$\kappa \leq \pi/m \quad \text{and} \quad \|u\|_1 \leq \pi/\kappa,$$  \hfill (63)

we have

$$\mathbb{E}[\tilde{R}_X, u] \leq 2(1 + \beta_\gamma(m\kappa)) \gamma_\kappa(m\xi) + \alpha(\kappa u),$$  \hfill (64)

where $\alpha$, $\beta_\gamma$ and $\gamma_\kappa$ are defined in (8), (42) and (59).

\textbf{Proof:} See Appendix M.

If $\kappa$ is sufficiently small, then $\alpha(\kappa u)$ and $\beta_\gamma(m\kappa)$ become negligible with respect to $\gamma_\kappa(m\xi)$, and the bound given in (64) is roughly equal to $2\gamma_\kappa(m\xi)$. Theorem 3 therefore provides a validity domain for shift invariance of $\mathbb{R}GPool$ operators, as illustrated in Fig. 4 with $q = 1$.

Remark 3: The stochastic discrepancy introduced in (62) is estimated relatively to the $\mathbb{C}GMod$ output. This choice is motivated by the perfect shift invariance of its norm, as shown in Proposition 3.

Remark 4: In practice, most of the time max pooling is performed on a grid of size $3 \times 3$; therefore $q = 1$. For the sake of concision, in the remaining of this paper, we will drop $q$ in the notations, which implicitly means $q = 1$.

IV. STANDARD CNNS IN OUR FRAMEWORK

In this section, we show how the theoretical results established above on single-channel inputs can be applied to standard CNNs such as AlexNet or ResNet, which are usually designed for multichannel inputs—e.g., RGB images.

A. Background

We focus on the first layers of a classical CNN architecture, in which input images are propagated through a convolution layer followed by rectified linear unit and max pooling operators, as described below.

We denote by $K$ and $L \in \mathbb{N}^*$ the number of input and output channels, respectively. The convolution layer is characterized by a trainable weight tensor $V \in (I_2^2(\mathbb{Z}^2))^{L \times K}$ with a finite support, a bias vector $b \in \mathbb{R}^L$ and a subsampling factor $m \in \mathbb{N}^*$.

Considering an input image $X \in (I_2^2(\mathbb{Z}^2))^L$ with finite support, we denote by $A \in (I_2^2(\mathbb{Z}^2))^L$ the output of the max pooling layer. Then we have, for any $l \in \{0 \ldots L - 1\}$,

$$A_l := (\text{MaxPool} \circ \text{ReLU})\left(b_l + \sum_{k=0}^{K-1} (X_k \ast \nabla_{ik}) \downarrow m\right),$$  \hfill (65)

where MaxPool has been introduced in (23) with $q = 1$ (see Remark 4), and ReLU is defined such that, for any $X' \in I_2^2(\mathbb{Z}^2)$ and any $n \in \mathbb{Z}^2$, $\text{ReLU}(X')[n] := \max(0, X'[n])$.

Expression (65) also introduces a bias notation, defined such that $(b + X')[n] = b + X'[n]$ for any $n \in \mathbb{Z}^2$.

In many cases, input images are composed of three RGB channels; therefore $K = 3$. The other parameters depend on the chosen CNN architecture. For example, in AlexNet, the weight tensor $V$ is supported in a region of size $11 \times 11$ and the subsampling factor $m$ is equal to 4. ResNet models use kernels of size $7 \times 7$ with $m = 2$.

B. Making CNNs Compatible with our Theoretical Results

The bias and ReLU are outside the scope of our study. However, we can show that $A_l = \text{ReLU}(b_l + Y\text{pool}_l)$, where we have defined $Y\text{pool}_l := R_{m}^{(l)}X$, with

$$R_{m}^{(l)} : X \mapsto \text{MaxPool}\left(\sum_{k=0}^{K-1} (X_k \ast \nabla_{ik}) \downarrow m\right).$$  \hfill (66)

For the sake of our study, we therefore consider a strictly equivalent CNN architecture where the bias and ReLU are computed after the max pooling layer. We then focus on the intermediate output $Y\text{pool} \in (I_2^2(\mathbb{Z}^2))^L$.

Remark 5: In many architectures including ResNet, bias $b$ is replaced by a batch normalization layer with affine transformation [49]. Swapping such a layer with max pooling isn’t straightforward, but can nevertheless be done with caution. Therefore, we can once again focus on $Y\text{pool}$ such as introduced in (66).

In what follows, we assume that the network has been trained on a natural image dataset such as ImageNet. Let $\mathcal{L}$ denote the subset of output channels $l \in \{0 \ldots L - 1\}$ such that, for any $k \in \{0 \ldots K - 1\}$, $V_{lk}$ behaves like a band-pass filter (see Fig. 1). They are referred to as Gabor channels. We would like to apply the theoretical results obtained in section III on $Y\text{pool}_l$ for any $l \in \mathcal{L}$.

To do so, we need to show that $Y\text{pool}_l$ is the output of a $\mathbb{R}GPool$ operator such as introduced in (22), for some weight and input to be defined. This is in general not possible though, because the sum operator in (66) cannot be interchanged with max pooling. To solve this problem, we state the following hypothesis, applied on any Gabor channel $l \in \mathcal{L}$. It states that the trained kernels $V_{lk}$ for $k \in \{0 \ldots K - 1\}$ are identical, up to a multiplicative constant.

\textbf{Hypothesis 3:} Let $\tilde{V}_l := \frac{1}{K} \sum_{k=0}^{K-1} V_{lk}$ denote the mean kernel of the $l$-th output channel. Then, for any $l \in \mathcal{L}$, there exists $\mu_l \in \mathbb{R}^K$ such that

$$\forall k \in \{0 \ldots K - 1\}, \quad V_{lk} = \mu_k \tilde{V}_l.$$  \hfill (67)

Intuitively, when looking at Gabor-like kernels in Fig. 1, they roughly appear grayscale. This observation supports Hypothesis 3 with $\mu_k \approx 1$. A more accurate justification for this hypothesis is provided in Appendix I.

Then, considering the linear combination $\tilde{X}_l := \mu_l^\top X$, we apply Hypothesis 3 to (66):

$$\forall l \in \mathcal{L}, \quad Y\text{pool}_l = R_{m}^{(l)}X = R_{m}^{(l)}\tilde{X}_l,$$  \hfill (68)
TABLE I

<table>
<thead>
<tr>
<th>MODEL</th>
<th>NB CHANNELS</th>
<th>SIZE OF $\mathcal{L}$ $\times$ $\kappa$</th>
<th>MEAN RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALEXNET</td>
<td>64</td>
<td>26 $\times$ $\pi/4$</td>
<td>67%</td>
</tr>
<tr>
<td>RESNET-34</td>
<td>64</td>
<td>22 $\times$ $\pi/2$</td>
<td>76%</td>
</tr>
</tbody>
</table>

where $\tilde{W}_l \in l^2(\mathbb{Z}^2)$ denotes the complex-valued companion of $\tilde{V}_l$ satisfying (5). We also define $c_{2m}^{(l)}$ and its associated output feature map $Y_{mod,l}$, such that

$$\forall l \in \mathcal{L}, Y_{mod,l} := c_{2m}^{(l)}X = C_{2m}[\tilde{W}_l](\tilde{X}_l).$$

Besides, the following hypothesis states that, for any $l \in \mathcal{L}$, $\tilde{W}_l$ is a discrete Gabor-like filter for which the Fourier support size is shared among the Gabor channels.

**Hypothesis 4:** There exists $\kappa \leq \pi/m$ such that, for any Gabor channel $l \in \mathcal{L}$,

$$\exists \xi_l \in [-\pi, \pi]^2 : \tilde{W}_l \in \mathcal{G}(\xi_l, \kappa).$$

Let $l \in \mathcal{L}$. According to Hypotheses 3–4, $R_{m}^{(l)}$ and $c_{2m}^{(l)}$ such as defined in (68) and (69) can be respectively qualified as RGPool and CGMod operators. We now consider

$$\tilde{P}_X^{(l)} := \|c_{2m}^{(l)}X - R_{m}^{(l)}X\|_2^2/\|c_{2m}^{(l)}X\|_2^2;$$
$$\tilde{R}_{X,u}^{(l)} := \|R_{m}^{(l)}(T_uX) - R_{m}^{(l)}X\|_2^2/\|c_{2m}^{(l)}X\|_2^2.$$  

Then, under Hypotheses 1–4, we can apply Theorems 2 and 3 to the above random variables, thus providing a shift invariance measure for a subset of outputs in CNNs:

$$\mathbb{E}[\tilde{P}_X^{(l)}] \leq (1 + \beta(m\kappa))^2 \gamma(m\xi_l)^2;$$
$$\mathbb{E}[\tilde{R}_{X,u}^{(l)}] \leq 2(1 + \beta(m\kappa))^2 \gamma(m\xi_l) + \alpha(\kappa\xi_l).$$

where $\xi_l$ has been introduced in (70).

In practice, Hypothesis 4 cannot be exactly satisfied. This is because $\tilde{V}_l$ is finitely supported, and thus its power spectrum cannot be exactly zero on a region with non-zero measure. To evaluate how close we are to this ideal situation, we measured the maximum percentage of energy within a window of size $\kappa \times \kappa$ in the Fourier domain, with respect to the whole filter $W_l$. We then computed the mean percentage over all the Gabor channels $l \in \mathcal{L}$. The results are shown in Table I, for AlexNet and ResNet after training with ImageNet. The window size $\kappa$ has been set to its highest admissible value, i.e., $\pi/m$.

We notice that residual energy outside the window of interest is quite high, especially for AlexNet. Therefore, $\tilde{W}_l$ deviates from a “perfect” Gabor-like filter, which may lead to higher instabilities.

V. OPERATORS BASED ON THE DUAL-TREE WAVELET PACKET TRANSFORM

In order to validate the results established in section III, we consider the dual-tree wavelet packet transform (DT-CWPT), a fast algorithm which achieves subsampled convolutions with discrete Gabor-like filters. As explained below, DT-CWPT spawns a set of filters which tiles the whole frequency domain in a regular fashion. Furthermore, increasing the subsampling factor $m$ results in a decreased Fourier support size $\kappa = \pi/m$, therefore matching the condition stated in (63). This is also consistent with what is observed in CNNs such as AlexNet or ResNet. DT-CWPT thus provides a convenient framework to emulate the behavior of an actual CNN while testing Theorems 1–3 in a controlled environment.

A. Discrete Wavelet Packet Transform

This is a brief overview on the classical, real-valued 2D wavelet packet transform (WPT) algorithm [48, p. 377], which is the starting point for building the redundant, complex-valued DT-CWPT.

Given a pair of low- and high-pass 1D orthogonal filters $h, g \in l^2(\mathbb{R})$ satisfying a quadrature mirror filter (QMF) relation, we consider a separable 2D filter bank (FB), denoted by $G := (G_l)_{l \in \{0..3\}}$, defined by

$$G_0 = h \otimes h; \quad G_1 = h \otimes g; \quad G_2 = g \otimes h; \quad G_3 = g \otimes g.$$  

Let $X \in l^2(\mathbb{Z}^2)$. The decomposition starts with $X_0 = X$. Given $j \in \mathbb{N}$, suppose that we have computed $4^j$ sequences of wavelet packet coefficients at stage $j$, denoted by $X_{l,j}^{(l)} \in l^2(\mathbb{Z}^2)$ for each $l \in \{0..4^j-1\}$. They are referred to as feature maps.

At stage $j + 1$, we compute a new representation of $X$ with increased frequency resolution—and decreased spatial resolution. It is obtained by further decomposing each feature map $X_{l,j}^{(l)}$ into four sub-sequences, using subsampled (or strided) convolutions with kernels $G_k$, for each $k \in \{0..3\}$:

$$\forall k \in \{0..3\}, X_{4^j + k}^{(l)} = X_{l,j}^{(l)} * G_k.$$  

The algorithm stops after reaching the desired number of stages $J > 0$—referred to as decomposition depth. Then,

$$X^{(J)} := (X_{l,j}^{(l)})_{l \in \{0..4^j-1\}}.$$

constitutes a multichannel representation of $X$ in an orthonormal basis, from which the original image can be perfectly reconstructed.

The following proposition introduces an array of resulting kernels $V^{(J)}$ which is illustrated in Fig. 3 with $J = 2$.

**Proposition 6:** For any $l \in \{0..4^j-1\}$, there exists $V_{l,j}^{(l)} \in l^2(\mathbb{Z}^2)$ such that

$$X_{l,j}^{(l)} = (X * V_{l,j}^{(l)}) \downarrow 2^j.$$  

**Proof:** See Appendix N.

B. Dual-Tree Complex Wavelet Packet Transform

Despite having interesting properties such as sparse signal representation, WPT is unstable with respect to small shifts and suffers from a poor directional selectivity. To overcome this, N. Kingsbury designed a new type of discrete wavelet transform [29], where images are decomposed in a redundant frame of nearly-analytic, complex-valued waveforms. It was
Later extended to the wavelet packet framework [30]. The latter operation, referred to as dual-tree complex wavelet packet transform (DT-CWPT), is performed as follows.

Let \((h_0, g_0)\) and \((h_1, g_1)\) denote two pairs of QMFs as defined in section V-A, satisfying the half-sample delay condition:

\[
\forall \omega \in [-\pi, \pi], \quad \hat{h}_1(\omega) = e^{-i\omega/2} \hat{h}_0(\omega).
\]

Then, for any \(k \in \{0, 1\}\), we build a two-dimensional FB \(G_k := (G_{k,l})_{l \in \{0, 3\}}\) similarly to (75):

\[
G_{k,0} = h_i \otimes h_j; \quad G_{k,1} = h_i \otimes g_j; \quad G_{k,2} = g_i \otimes h_j; \quad G_{k,3} = g_i \otimes g_j,
\]

where \(i, j \in \{0, 1\}\) are defined such that \(k = 2 \times i + j\).

Let \(J > 0\) denote a decomposition depth. Using each of the four FBs \(G_{0,3}\) as defined above, we assume that we have decomposed an input image \(X\) into four multichannel WPT representations \(X_{0,3}\), each of which satisfies (76) and (77).

Then, for any \(l \in \{0 \ldots 4^J - 1\}\), the following complex feature maps are computed:

\[
\begin{pmatrix}
Z_{l}^{(J)} \\
Z_{l+4}^{(J)}
\end{pmatrix} =
\begin{pmatrix}
1 & i \\
1 & 1
\end{pmatrix}
\begin{pmatrix}
X_{0,l}^{(J)} \\
X_{3,l}^{(J)}
\end{pmatrix} +
\begin{pmatrix}
1 & 1 \\
1 & -1
\end{pmatrix}
\begin{pmatrix}
X_{2,l}^{(J)} \\
X_{1,l}^{(J)}
\end{pmatrix}.
\]

We denote by \(L_J := 2 \cdot 4^J\) the number of output feature maps. Then,

\[
Z^{(J)} := \{Z_{l}^{(J)}\}_{l \in \{0 \ldots L_J - 1\}}
\]

constitutes a complex-valued, four-time redundant multichannel representation of \(X\) from which the original image can be reconstructed.

As for standard WPT, the following proposition introduces an array of resulting (complex-valued) kernels \(W^{(J)}\) for which a graphical representation is provided in Fig. 3 with \(J = 2\).

**Proposition 7:** For any \(l \in \{0 \ldots L_J - 1\}\), there exists \(W^{(J)}_l \in \mathbb{C}^2\) such that

\[
Z_l^{(J)} = (X \ast W^{(J)}_l) \downarrow 2^J.
\]

**Proof:** See Appendix O.

\[
\Box
\]

C. Invariance Results Applied to the Dual-Tree Framework

We assume that \(h_0\) is a Shannon filter, such that \(\hat{h}_0(\omega) := \sqrt{2} \text{ if } \omega \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]\) and 0 otherwise. Let \(J \in \mathbb{N}^+\) denote the number of decomposition stages. We consider the resulting kernels \(W^{(J)}_l\) satisfying (84). The following hypothesis states that DT-CWPT tiles the frequency plane with a square grid.

**Hypothesis 5:** For any \(l \in \{0 \ldots L_J - 1\}\), there exists \(\sigma_l^{(J)} \in \{-2^J \ldots 2^J - 1\}\) such that

\[
W^{(J)}_l \in G(\xi^{(J)}_l, \kappa_l),
\]

where we have defined

\[
\xi^{(J)}_l := \left(\sigma_l^{(J)} + \frac{1}{2}\right) \frac{\pi}{2^J} \quad \text{and} \quad \kappa_l := \frac{\pi}{2^J}.
\]

It can be shown that Hypothesis 5 is an asymptotic result, when \(J \to \infty\). In reality, the Fourier support of \(W^{(J)}_l\) is contained in four symmetric square regions of size \(\kappa_l\). If the dual filter \(h_1\) satisfies the half-sample delay condition (79), then the energy of \(W^{(J)}_l\) goes to 0 in all but one of the four regions (relatively to the filter’s total energy). We nevertheless consider Hypothesis 5 as reasonable when \(J > 2\).

Then, according to (84) and (85), we can apply Theorems 1–3 to the dual-tree framework. More precisely, for any output channel \(l \in \{0 \ldots L_J - 1\}\), we consider

\[
Y^{(J)}_{\text{pool}_l} := R^{(J)}_l X \quad \text{and} \quad Y^{(J)}_{\text{modl}_l} := C^{(J)}_l X,
\]

where we have denoted, for the sake of readability,

\[
R^{(J)}_l := R_{m,l} \left[ W^{(J)}_l \right] \quad \text{and} \quad C^{(J)}_l := C_{2m,l} \left[ W^{(J)}_l \right],
\]

with \(m,l := 2^{J-1}\). We recall that \(R_m\) and \(C_m\) have been introduced in (31) with \(q \leftarrow 1\), for any \(m \in \mathbb{N}^*\).

Both \(R^{(J)}_l\) and \(C^{(J)}_l\) perform DT-CWPT with \(J\) decomposition stages. However, the max pooling operator requires an extra level of subsampling. To counterbalance this, when computing \(Y^{(J)}_{\text{pool}}\), the last stage of DT-CWPT decomposition must be performed without subsampling, at the cost of increased redundancy.\(^{11}\) This is why \(m,l := 2^{J-1}\) instead of \(2^J\).

In order Theorems 1 and 3 to be applicable, we need to check that conditions stated in (63) are satisfied. Using the DT-CWPT framework, they become

\[
kappa_J \leq \frac{\pi}{2^{J-1}} \quad \text{and} \quad \|u\|_1 \leq \frac{\pi}{\kappa_J} = 2^J.
\]

The first condition is always met, according to (86). As for the second one, it establishes a limit on \(\|u\|_1\) above which

\(^{10}\)This asymptotic result is not true for “edge filters”, i.e., when \(\|\xi^{(J)}_l\|_{\infty} = (1 - 2^{-(J+1)}) \pi\). In this case, a small fraction of the filter’s energy remains located at the far end of the Fourier domain [30]. However, this edge effect is ignored and we still consider Hypothesis 5 as reasonable.

\(^{11}\)This is similar to the concept of continuous wavelet transform [51].
shift invariance can no longer be estimated. Note however that shifting the input by $2^j$ pixels results in a 1-pixel output shift. Therefore, $R_l^{(J)}(T_{u} X)$ can always be compared with a shifted version of $R_l^{(J)} X$. We then get a partial measure of shift equivariance.

In the Shannon setting, $h_0[n]$, $g_0[n]$, $h_1[n]$ and $g_1[n]$ decay in $O(1/n)$, which makes them difficult to approximate in practice. It requires very large vectors to avoid numerical instabilities. Practical implementations use fast-decaying filters such as Meyer QMFs [52], or finite-length filters which approximate the half-sample delay condition [53]. Therefore, residual energy can be observed outside the Fourier windows introduced in Hypothesis 5. To counterbalance this, we relax this hypothesis by increasing the window size up to $\kappa$.

In the Shannon setting, $h_0[n]$, $g_0[n]$, $h_1[n]$ and $g_1[n]$ decay in $O(1/n)$. According to Hypothesis 5, these frequencies form a regular pattern of dark spots. More precisely, high discrepancies are expected for certain pathological frequencies, due to the search for a maximum value over a discrete grid—see Fig. 2.

VI. EXPERIMENTS AND RESULTS

In this section, we experimentally validate our theoretical results. To do so, we built operators based on DT-CWPT, as explained in section V. Using a dataset of natural images, we measured the mean discrepancy between $\mathbb{R}GPool$ and $\mathbb{C}GMod$ outputs, and evaluated the shift invariance of both models.

Dual-tree decompositions have been performed with Q-shift orthogonal filters of length 10 [50], which approximately meets the half-sample delay condition (79).

Our implementation is based on PyTorch. The models were evaluated on the validation set of ImageNet ILSVRC2012 [54], which contains $N_{\text{data}} = 50,000$ images.

A. Discrepancies between $\mathbb{R}GPool$ and $\mathbb{C}GMod$

Each image $n \in \{0 \ldots N_{\text{data}} - 1\}$ in the dataset was converted to grayscale, from which a center crop of size $224 \times 224$ was extracted. We denote by $X_n \in l^2_d(\mathbb{Z}^2)$ the resulting input feature map. For any $l \in \{0 \ldots L_J - 1\}$, we denote by $Y_{\text{pool}}^{(J)}$ and $Y_{\text{mod}}^{(J)}$ the outputs satisfying (87) with $X \leftarrow X_n$.

Then, the relative quadratic error between $Y_{\text{mod}}^{(J)}$ and $Y_{\text{pool}}^{(J)}$ was computed. It is defined by the square of

$$\rho_{nl}^{(J)} := \frac{\|Y_{\text{mod}}^{(J)} - Y_{\text{pool}}^{(J)}\|_2}{\|Y_{\text{mod}}^{(J)}\|_2}. \quad (90)$$

Finally, the for each output channel $l$, an empirical estimate for $\mathbb{E}[P_{X}^{2}]$, such as introduced in (55) was obtained by averaging $(\rho_{nl}^{(J)})^2$ over the whole dataset. We denote by $(\bar{\rho}_l^{(J)})^2$ the corresponding quantity.

Since $R_l^{(J)}$ and $C_l^{(J)}$ are parameterized by $W_l^{(J)}$, the value of $(\bar{\rho}_l^{(J)})^2$ depends on the filter’s characteristic frequency $\xi_l^{(J)}$. According to Hypothesis 5, these frequencies form a regular grid in the Fourier domain. This provides a visual representation of $(\bar{\rho}_l^{(J)})^2$, as shown in Fig. 5. We can observe a regular pattern of dark spots. More precisely, high discrepancies between max pooling and modulus seem to occur when the support of $W_l^{(J)}$ is located in a dark region of Fig. 4.

This result corroborates the theoretical study, which states that high discrepancies are expected for certain pathological frequencies, due to the search for a maximum value over a discrete grid—see Fig. 2.
B. Shift invariance

For each input image \( X_n \) previously converted to grayscale, two crops of size \( 224 \times 224 \) were extracted, such that the corresponding sequences \( X_n \) and \( X'_n \) are shifted by one pixel along the \( x \)-axis.

From these inputs, the following quantity was then computed:

\[
\rho_{\text{pool}}^{(J)} := \frac{\|Y^{(J)}_{\text{pool}} - Y^{(J)}_{\text{mod}}\|_2}{\|Y^{(J)}_{\text{mod}}\|_2},
\]

where \( Y^{(J)}_{\text{pool}} \) and \( Y^{(J)}_{\text{mod}} \) satisfy (87) with \( X \leftarrow X_n \) or \( X'_n \). Finally, for each output channel \( l \in \{0 \ldots L_J - 1\} \), an empirical estimate of the \( \rho_{\text{pool}}^{(J)} \) was obtained by averaging over the whole dataset. We denote by \( \check{\rho}_{\text{pool}}^{(J)} \), the corresponding quantity. We point out that shift invariance is measured relatively to the norm of the \( \text{CGMod} \) output, as explained in Remark 3.

On the other hand, the same procedure was applied on the \( \text{CGMod} \) operators:

\[
\check{\rho}_{\text{mod}}^{(J)} := \frac{\|Y^{(J)}_{\text{mod}} - Y^{(J)}_{\text{mod}}\|_2}{\|Y^{(J)}_{\text{mod}}\|_2},
\]

and \( \check{\rho}_{\text{mod}}^{(J)} \) was obtained as before by averaging \( \check{\rho}_{\text{mod}}^{(J)} \) over the whole dataset.

A visual representation of \( \check{\rho}_{\text{pool}}^{(J)} \) and \( \check{\rho}_{\text{mod}}^{(J)} \) are provided in Fig. 6. Two observations can be drawn here.

When the filter is horizontally oriented, the corresponding output is highly stable with respect to horizontal shifts. This can be explained by noticing that such kernels perform low-pass filtering along the \( x \) axis. The exact transposed phenomenon occurs for vertical shifts (not shown in this paper).

Elsewhere, we observe that high discrepancies between \( \text{RGPool} \) and \( \text{CGMod} \) outputs (Fig. 5) are correlated with shift instability of \( \text{RGPool} \) (Fig. 6, top). This is in line with (61) and (64) in Theorems 2–3. Note that \( \text{CGMod} \) outputs are nearly shift invariant regardless the characteristic frequency \( \xi_i^{(J)} \) (Fig. 6, bottom), as predicted by Theorem 1 (20).

VII. CONCLUSION

In this paper, we studied shift invariance properties captured by the max pooling operator, when applied on top of a convolution layer with Gabor-like kernels. More precisely, we established a validity domain for near-shift invariance, and confirmed our predictions with an experimental setting based on the dual-tree complex wavelet packet transform.

As shown in this paper, the \( \text{CGMod} \) operator acts like a proxy for \( \text{RGPool} \), extracting comparable features with higher stability. This result suggests a way to build an architecture sharing the structure and behavior of a standard network, except that shift invariance would be improved. This could be done by considering a DT-CWPT-based twin model such as introduced in our workshop paper [31], and replacing \( \text{RGPool} \) by \( \text{CGMod} \) operators as done above in a deterministic context.

Since CNNs generally implement successive blocks of convolution and pooling layers, another line of research would be to extend our results to a cascade of \( \text{RGPool} \) operators. Our work is thus an important step towards deeper understanding of popular networks, based on the wavelet theory.

APPENDIX A

PROOF OF LEMMA 1

Proof: We can show that, for any \( \omega \in \mathbb{R}^2 \),

\[
\hat{f}_0(\omega) = (f \ast \hat{\psi})(\omega - \nu) = T_\nu(f \ast \hat{\psi}^2)(\omega).
\]

By hypothesis on \( \psi \), we have \( \text{supp} \hat{\psi} \subset B_\infty(-\nu, \varepsilon/2) \), which yields the result.

APPENDIX B

PROOF OF PROPOSITION 1

Proof: Using the 2D Plancherel formula, we compute

\[
\|T_h f_0 - f_0\|_{L^2}^2 = \frac{1}{4\pi^2} \left\| \mathcal{T}_h f_0 - f_0 \right\|_{L^2}^2
\]

\[
= \frac{1}{4\pi^2} \int_{B_\infty(\varepsilon/2)} |f_0(\omega)|^2 |e^{-i(h, \omega)} - 1|^2 d^2 \omega
\]

\[
= \frac{1}{4\pi^2} \int_{B_\infty(\varepsilon/2)} |f_0(\omega)|^2 (2 - 2 \cos (h, \omega)) d^2 \omega.
\]

The integral is computed on a compact domain because, according to Lemma 1, \( \text{supp} \hat{f}_0 \subset B_\infty(\varepsilon/2) \). Now, we use the Cauchy-Schwarz inequality to compute:

\[
\forall \omega \in B_\infty(\varepsilon/2), \ |\langle h, \omega \rangle| \leq \|h\|_1 \cdot \|\omega\|_{\infty} \leq \varepsilon \frac{\|h\|_1}{2},
\]

By hypothesis on \( h, \varepsilon \|h\|_1 \leq \frac{\varepsilon}{2}, \) and thus

\[
\|T_h f_0 - f_0\|_{L^2}^2 \leq \left(2 - 2 \cos \varepsilon \frac{\|h\|_1}{2}\right) \|f_0\|_{L^2}^2.
\]

Finally, since \( \cos t \geq 1 - \frac{t^2}{2}, \) we get (7).

APPENDIX C

PROOF OF LEMMA 2

Proof: Expression (9) is obtained by adapting (3.2) and (3.3) in [48] to the 2D case. Then, by combining (9) with the Plancherel formula, we get

\[
\|g\|_{L^2}^2 = \frac{1}{4\pi^2} \|\hat{g}\|_{L^2}^2
\]

\[
= \frac{1}{4\pi^2} \int_{B_\infty(\pi/s)} |\hat{g}(\omega)|^2 d^2 \omega
\]

\[
= \frac{1}{4\pi^2} \int_{B_\infty(\pi/s)} |\hat{s} \hat{Y}(s \omega)|^2 d^2 \omega.
\]

The integral is performed on \( B_\infty(\pi/s) \) because \( g \in U_s \). Then, by applying the change of variable \( \omega' \leftarrow s \omega \), we get

\[
\|g\|_{L^2}^2 = \frac{1}{4\pi^2} \int_{B_\infty(\pi)} |\hat{Y}(\omega')|^2 d^2 \omega'
\]

\[
= \frac{1}{4\pi^2} \|\hat{Y}\|_{L^2}^2 = \|Y\|_{L^2}^2,
\]

hence (10), which concludes the proof.
Finally, plugging (94) into (93) concludes the proof.

Let $\{\phi_n^{(s)}\}_{n \in \mathbb{Z}^2}$ be an orthonormal basis of $\mathcal{U}_s$. Therefore, we get

$$C_mX[n] = |f_0(s'n)|; \quad C_m(\mathcal{T}_uX)[n] = |(\mathcal{T}_{su}f_0)(s'n)|. \quad (97)$$

According to (13) (Proposition 2), (18) (Corollary 1) and (96), we get

$$\|C_m(\mathcal{T}_uX) - C_mX\|_2^2 \leq \sum_{n \in \mathbb{Z}^2} |f_0(s'n) - f_0(s'n)|^2$$

$$= \sum_{n \in \mathbb{Z}^2} |g(s'n)|^2 = \frac{1}{s^2} \|Y\|_2^2,$$

where $\{\phi_n^{(s)}\}_{n \in \mathbb{Z}^2}$ are the reverse triangular inequality on (97) and (98).

$$\|C_m(\mathcal{T}_uX) - C_mX\|_2^2 \leq \frac{1}{s^2} \|Y\|_2^2 = \frac{1}{s^2} \|\mathcal{T}_u f_0 - f_0\|_{L_2}^2.$$

According to Proposition 1 with $\varepsilon \leftarrow \kappa/s$ and $h \leftarrow su$, we then get the following bound:

$$\|C_m(\mathcal{T}_uX) - C_mX\|_2^2 \leq \frac{(\alpha u)^2}{s^2} \|f_0\|_{L_2}^2.$$

(99)

Besides, according again to Lemma 2, $\|f_0\|_{L_2}^2 = \|X_0\|_2^2$, where $X_0[n] := s'f_0(s'n)$ for any $n \in \mathbb{Z}^2$. Therefore, according to (97),

$$\|C_mX\|_2 = \frac{1}{s} \|X_0\|_2 = \frac{1}{s} \|f_0\|_{L_2}.$$ \quad (100)

Finally, plugging (100) into (99) completes the proof. \hfill \blacksquare

### Appendix G

**Proof of Proposition 3**

Let $X \in L^2_0(\mathbb{Z}^2)$ and $s > 0$. We consider $f_0 \in L^2(\mathbb{R})$ as the “low-frequency” function satisfying (95). Again, we introduce $s' := ms$ and $X_0 \in L^2(\mathbb{Z}^2)$ such that $X_0[n] := s'f_0(s'n)$. Moreover, for any $Y \in L^2_0(\mathbb{Z}^2)$, we denote:

$$f_Y(s') := \sum_{n \in \mathbb{Z}^2} Y[n] \phi_n^{(s')}.$$

(101)

On the one hand, in the proof of Theorem 1, we already got (100). On the other hand, (98) can be rewritten

$$C_m(\mathcal{T}_uX)[n] = |\mathcal{T}_{su'}f_0(s'n)|,$$

(102)

with $u' := u/m$. Besides, according to the proof of Theorem 1, $f_0 \in \mathcal{U}_{s'}$. Thus, by definition of $X_0$, the Shannon-Whittaker theorem [48, p. 61] implies that $f_0 = f_{X_0}^{(s')}$ as defined in (101). Then, using Lemma 3 with $X \leftarrow X_0$, $u \leftarrow u'$ and $s \leftarrow s'$, we get

$$f_{\mathcal{T}_uX_0}(s') := \mathcal{T}_{su'}f_0(s'n).$$

(103)

Finally, using (12) in Proposition 2 with $X \leftarrow \mathcal{T}_uX_0$ and $s \leftarrow s'$, and inserting (103) into the result yields

$$\mathcal{T}_{su'}X_0[n] = s'f_{\mathcal{T}_uX_0}(s'n) = s'\mathcal{T}_{su'}f_0(s'n).$$

(104)
Therefore, (102) and (104) imply
\[ \|C_m(T_wX)\|_2 = \frac{1}{s} \|T_wX_0\|_2. \tag{105} \]
Moreover, since \( f_0 \in L_d \), and according to (104), we can use Lemma 2 with \( s \leftarrow s', g \leftarrow T_{s'}f_0 \) and \( Y \leftarrow T_{s'}X_0 \). We get
\[ \|T_{s'}X_0\|_2 = \|T_{s'}Y\|_2 = \|f_0\|_2. \tag{106} \]
Finally, (100), (105) and (106) imply \( \|C_m(T_wX)\|_2 = \|C_mX\|_2 \), which concludes the proof.

**Appendix H**

**Justification for Hypotheses 1 and 2**

Given \( n \in \mathbb{N}^* \), we define \( n \)-th order stationarity of a given stochastic process \( F \) as in [55, p. 152]: for any \( n' \in \{1 \ldots n\} \), \((x_1, \ldots, x_{n'}) \in (\mathbb{R}^2)^{n'} \) and \( h \in \mathbb{R}^2 \), the joint distribution of \( (F(x_1), \ldots, F(x_{n'})) \) is identical to the one of \( (F(x_1 + h), \ldots, F(x_{n'} + h)) \). Besides, strict-sense stationarity is defined as \( n \)-th order stationarity for any \( n \in \mathbb{N}^* \).

We recall that \( \nu := \xi/s \). We then state the following results.

**Proposition 8:** We assume that \( F \) is first-order stationary. If, for any \( x \in \mathbb{R}^2 \) and any \( h \in B_2(2\pi/\|\nu\|_2) \),
\[ (T_hF_X + \overline{\nu})(x) = e^{i(\nu, h)}(F_X + \overline{\nu})(x), \tag{107} \]
then Hypothesis 1 is satisfied.

**Proof:** We show that the probability measure of \( Z_X(x) \) is invariant with respect to phase shift. In other words, we show that, for any measurable set \( \mathcal{A} \subset \mathbb{S}^1 \),
\[ \forall \omega \in [0, 2\pi[, \mu(\mathcal{A}) = \mu(e^{i\omega}\mathcal{A}), \tag{108} \]
where we have denoted
\[ \mu : \mathcal{A} \mapsto \mathbb{P}\{Z_X(x) \in \mathcal{A}\}. \]
Let \( h \in B_2(2\pi/\|\nu\|_2) \). According to (107),
\[ Z_X(x) \in \mathcal{A} \iff T_hZ_X(x) \in e^{i(\nu, h)\mathcal{A}}. \]
Therefore,
\[ \mathbb{P}\{Z_X(x) \in \mathcal{A}\} = \mathbb{P}\{T_hZ_X(x) \in e^{i(\nu, h)\mathcal{A}}\}. \]
Since \( F_X \) is first-order stationary, \( Z_X(x) \) and \( T_hZ_X(x) \) have the same probability distribution. Thus we get
\[ \mathbb{P}\{Z_X(x) \in \mathcal{A}\} = \mathbb{P}\{Z_X(x) \in e^{i(\nu, h)\mathcal{A}}\}. \]
Let \( \omega \in [0, 2\pi] \). Considering \( h := \omega\nu/\|\nu\|_2 \), we have \( h \in B_2(2\pi/\|\nu\|_2) \) and \( \langle \nu, h \rangle = \omega \). Therefore,
\[ \forall \omega \in [0, 2\pi], \mathbb{P}\{Z_X(x) \in \mathcal{A}\} = \mathbb{P}\{Z_X(x) \in e^{i\omega}\mathcal{A}\}, \]
which yields (108).

Any probability measure defined on \( \mathbb{S}^1 \) is a Radon measure. Therefore, according to Haar’s theorem [56], there exists a unique probability measure on \( \mathbb{S}^1 \) satisfying (108). Since the uniform probability measure is also invariant to phase shift, we deduce that \( Z_X(x) \) is uniformly distributed on \( \mathbb{S}^1 \).

**Proposition 9:** We assume the conditions of Proposition 8 are met. If, moreover, \( F_X \) is strict-sense stationary, then Hypothesis 2 is satisfied.

**Proof:** Let \( n \in \mathbb{N}^* \) and \( x, y_1, \ldots, y_n \in \mathbb{R}^2 \). To alleviate notations, we consider the random vector \( M = (M_X(y_i))_{i \in \{1 \ldots n\}} \) with outcomes in \( \mathbb{R}_+^n \).

The proof is organized as follows. Using a similar reasoning as Proposition 8, we show that \( Z_X \) follows a uniform conditional probability distribution given \( M \). Since we already know that \( Z_X \) follows a uniform (unconditional) distribution, we deduce that \( Z_X \) and \( M \) are independent.

Let \( \mathfrak{A} \subset \mathbb{S}^1 \) and \( \mathfrak{E} := (\mathfrak{E}_i)_{i \in \{1 \ldots n\}} \subset \mathbb{R}_+^n \) denote measurable sets. Let \( h \in B_2(2\pi/\|\nu\|_2) \). According to (107),
\[ Z_X(x) \in \mathfrak{A} \iff T_hZ_X(x) \in e^{i(\nu, h)\mathfrak{A}}; \]
\[ M_X(y_i) \in \mathfrak{E}_i \iff T_hM_X(y_i) \in \mathfrak{E}_i \quad \forall i \in \{1 \ldots n\}. \]
Therefore,
\[ \mathbb{P}\{(Z_X(x) \in \mathfrak{A}) \& (M \in \mathfrak{E})\} = \mathbb{P}\{(T_hZ_X(x) \in e^{i(\nu, h)\mathfrak{A}}) \& (T_hM \in \mathfrak{E})\}. \]

Since \( F_X \) is strict-sense stationary, the joint probability density of \( T_hZ_X(x), T_hM_X(y_1), \ldots, T_hM_X(y_n) \) is identical to the one of \( Z_X(x), M_X(y_1), \ldots, M_X(y_n) \). Therefore we get
\[ \mathbb{P}\{(Z_X(x) \in \mathfrak{A}) \& (M \in \mathfrak{E})\} = \mathbb{P}\{(Z_X(x) \in e^{i(\nu, h)\mathfrak{A}}) \& (M \in \mathfrak{E})\}. \]

We assume that \( \mathbb{P}(M \in \mathfrak{E}) > 0 \). According to the above expression, and similarly to the proof of Proposition 8, we get,
\[ \forall \omega \in [0, 2\pi], \mathbb{P}\{Z_X(x) \in \mathfrak{A} \mid M \in \mathfrak{E}\} = \mathbb{P}\{Z_X(x) \in e^{i\omega}\mathfrak{A} \mid M \in \mathfrak{E}\}. \]
Then, the above conditional probability measure satisfies phase shift invariance (108). Therefore, as in the proof of Proposition 8, Haar’s theorem implies that \( Z_X(x) \) follows a uniform conditional distribution given \( M \in \mathfrak{E} \).

Moreover, strict-sense implies first-order stationarity, and thus, according to Proposition 8, \( Z_X(x) \) follows a uniform (unconditional) distribution. Therefore we get, for any measurable sets \( \mathfrak{A} \subset \mathbb{S}^1 \) and \( \mathfrak{E} \subset \mathbb{R}_+^n \) such that \( \mathbb{P}(M \in \mathfrak{E}) > 0 \),
\[ \mathbb{P}\{Z_X(x) \in \mathfrak{A} \mid M \in \mathfrak{E}\} = \mathbb{P}(Z_X \in \mathfrak{A}), \]
which proves independence between \( Z_X(x) \) and \( M \).

Strict-sense stationarity suggests that any translated version of a given image is equally likely. In reality, this statement is too strong, for several reasons. First, by construction, \( X \) has all its realizations in \( L_d^2(\mathbb{R}^2) \). In that context, a stationary process yields outcomes which are zero almost everywhere. Besides, depending on which category the image belongs to, the pixel distribution is likely to vary across various regions. For instance, we can expect the main subject to be located at the center of the image. More details on statistical properties of natural images can be found in [57]. Nevertheless, this hypothesis will be considered as a reasonable approximation if the shift is much smaller than the image “characteristic” size in the continuous domain; i.e., if \( \|h\|_2 \ll sM \), where, as a reminder, \( M \) denotes the support size of input images.\footnote{We refer readers to [58] for a related notion of local stationarity.}
As it turns out, the proofs of Propositions 8 and 9 only requires shifts with \( \|h\|_2 \leq 2\pi/\|\nu\|_2 \). Then, according to (53), \( \|h\|_2 \ll sM \), and the stationarity hypothesis holds.

Finally, to justify (107), we consider \( \varphi_W : x \mapsto \psi_W(x) e^{-i\langle \nu, x \rangle} \). Similarly to Lemma 1, we can show that \( \varphi_W \) is a low-pass filter, with \( \text{supp} \varphi_W \subseteq B_\infty(\varepsilon/2) \). For all \( h \in \mathbb{R}^2 \) such that \( \|h\|_2 \leq 2\pi/\|\nu\|_2 \), we have

\[
(T_h F_X + \varphi_W)(x) = \int \int \mathbb{T} h F_X(x - y) \varphi(y) e^{-i\langle \nu, y \rangle} \, d^2 y = e^{-i\langle \nu, h \rangle} \int \int F_X(x - y') \varphi(y' - h) e^{-i\langle \nu, y' \rangle} \, d^2 y'.
\]

Since \( \text{supp} \varphi_W \subseteq B_\infty(\varepsilon/2) \), we can define a “minimal wavelength” \( \lambda_{\psi W} := 2\pi \varepsilon/\kappa \). Then, if \( \|h\|_2 \ll \lambda_{\psi W} \), we can approximate \( \varphi(y' - h) \approx \varphi(y') \). This sufficient condition is actually met, because \( \|h\|_2 \leq 2\pi/\|\nu\|_2 \) and, according to (54), \( \|\nu\|_2 \gg \kappa/\varepsilon \). Therefore,

\[
(T_h F_X + \varphi_W)(x) \approx e^{i\langle \nu, h \rangle} (F_X + \varphi_W)(x). \tag{109}
\]

As a result, the conditions for Propositions 8 and 9 are approximately satisfied. We will therefore consider Hypotheses 1 and 2 as reasonable.

**Appendix I**

**Justification for Hypothesis 3**

We consider, for any \( l \in \mathcal{L} \) and any \( k \in \{0 \ldots K - 1\} \), the value of \( \mu \in \mathbb{R} \) minimizing \( \|\mu \hat{\nu}_l - V_{lk}\|_2 \), denoted by \( \mu_{lk} \). We then denote by \( \delta_{lk} := \frac{\|\mu_{lk} \hat{V}_l - V_{lk}\|_2}{\|V_{lk}\|_2} \) the relative quadratic error between \( V_{lk} \) and its projection on \( \mathbb{R} \hat{V}_l \). We get

\[
\delta_{lk} = 1 - \frac{\langle \hat{V}_l, V_{lk} \rangle^2}{\|\hat{V}_l\|_2^2 \|V_{lk}\|_2^2}. \tag{110}
\]

Expression (67) holds if and only if \( \delta_{lk} = 0 \) for any \( k \in \{0 \ldots K - 1\} \). In the case of AlexNet, when \( l \in \mathcal{L} \), \( \delta_{lk} \) do not exceed \( 10^{-4} \), and its mean value is around \( 10^{-2} \). Therefore \( \delta_{lk} \ll 1 \), and Hypothesis 3 can be considered as a reasonable assumption for any Gabor channel \( l \in \mathcal{L} \). Similar observations can be drawn for ResNet.

**Appendix J**

**Proof of Proposition 4**

**Proof**: We consider the Borel \( \sigma \)-algebra on \( \mathbb{S}^1 \) generated by \( \{\{z, z'\}_{\mathbb{S}^1} : z, z' \in \mathbb{S}^1\} \cup \{\mathbb{S}^1\} \), on which we have defined the angular measure \( \vartheta \) such that \( \vartheta(S^1) := 2\pi \), and

\[
\forall z, z' \in \mathbb{S}^1, \vartheta([z, z']_{\mathbb{S}^1}) := \angle(z^\ast z').
\]

For any \( p \in \mathbb{N}^* \), we compute the \( p \)-th moment of \( G_{\max x}(x) \) defined in (51). By considering

\[
h_{\max} : \mathbb{S}^1 \rightarrow [-1, 1] \quad z \mapsto \max_{\|k\|_2 \leq q} \text{Re}(z^\ast z_k),
\]

we get \( G_{\max x}(x) = h_{\max}(Z_X(x)) \). According to Hypothesis 1, \( Z_X(x) \) follows a uniform distribution on \( \mathbb{S}^1 \). Therefore,

\[
\mathbb{E}[G_{\max x}(x)^p] = \frac{1}{2\pi} \int_{\mathbb{S}^1} h_{\max}(z)^p \, d\vartheta(z),
\]

which proves that \( \mathbb{E}[G_{\max x}(x)^p] \) does not depend on \( x \). Let us split the unit circle \( \mathbb{S}^1 \) into the arcs \( \mathcal{A}_0, \ldots, \mathcal{A}_{N_q - 1} \) such as defined in (47):

\[
\mathbb{E}[G_{\max x}(x)^p] = \frac{1}{2\pi} \sum_{i=0}^{N_q-1} \int_{\mathcal{A}_i} h_{\max}(z)^p \, d\vartheta(z). \tag{111}
\]

Let \( i \in \{0 \ldots N_q - 1\} \). We can show that

\[
\forall z \in \mathcal{A}_i, h_{\max}(z) = \max \{\text{Re}(z^\ast z_i^q), \text{Re}(z^\ast z_{i+1}^q)\}
\]

Therefore, \( h_{\max} \) is symmetric with respect to the center value of \( \mathcal{A}_i^q \), denoted by \( z_i^q \). We denote by \( \mathcal{A}_i^q := [z_i^q, z_{i+1}^q) \), the first half of arc \( \mathcal{A}_i^q \).

Then,

\[
\forall z \in \mathcal{A}_i^q, h_{\max}(z) = \text{Re}(z^\ast z_i^q).
\]

As a consequence, using symmetry, we get

\[
\int_{\mathcal{A}_i^q} h_{\max}(z)^p \, d\vartheta(z) = 2 \int_{\mathcal{A}_i^q} h_{\max}(z)^p \, d\vartheta(z) = 2 \int_{\mathcal{A}_i^q} \text{Re}(z^\ast z_i^q)^p \, d\vartheta(z).
\]

By using the change of variable formula [59, p. 81] with \( z \mapsto e^{i\theta} \), we get

\[
\int_{\mathcal{A}_i^q} h_{\max}(z)^p \, d\vartheta(z) = 2 \int_{\theta_i^q} \cos^p (\theta - \theta_i^q) \, d\theta,
\]

where \( \theta_i^q := (\theta_i^q + \theta_{i+1}^q)/2 \) denotes the argument of \( z_i^q \).

Then, the change of variable \( \omega \mapsto \theta - \theta_i^q \) yields

\[
\int_{\mathcal{A}_i^q} h_{\max}(z)^p \, d\vartheta(z) = 2 \int_{0}^{\omega_i^q/2} \cos^p \omega \, d\omega. \tag{112}
\]

Now, we insert (112) into (111), and compute \( \mathbb{E}[G_{\max x}(x)^p] \) for \( p = 1 \) and \( p = 2 \). We get

\[
\mathbb{E}[G_{\max x}(x)] = \frac{1}{\pi} \sum_{i=0}^{N_q-1} \sin \frac{\omega_i^q}{2},
\]

\[
\mathbb{E}[G_{\max x}(x)^2] = \frac{1}{2} + \frac{1}{4\pi} \sum_{i=0}^{N_q-1} \sin \omega_i^q.
\]

We recall that \( Q_X := 1 - G_{\max x} \). By linearity of \( \mathbb{E} \), we get

\[
\mathbb{E}\left[Q_X(z)^2\right] := \frac{3}{2} + \frac{1}{4\pi} \sum_{i=0}^{N_q-1} \left(\sin \omega_i^q - 8\sin \frac{\omega_i^q}{2}\right).
\]

Using the notation introduced in (48) concludes the proof.

**Appendix K**

**Proof of Proposition 5**

**Proof**: We suppose that Hypothesis 2 is satisfied and we consider \( x \in \mathbb{R}^2 \). For a given \( n \in \mathbb{N}^* \), we introduce the random variable

\[
\tilde{S}_X^{(n)} := \sqrt{\sum_{|k|_\infty \leq n} M_X(x_k)^2}.
\]
According to Hypothesis 2, $Z_X(x)$ is jointly independent of $M_X(x_k)$ for $k \in \{-n \ldots n\}^2$. Therefore, by composition, $Z_X(x)$ is also independent of $\tilde{S}_X$. Moreover, according to (52) and (60), $\tilde{S}_X$ converges almost surely towards $S_X$, which proves independence between $Z_X(x)$ and $\tilde{S}_X$.

Now, we prove conditional independence between $Z_X(x)$ and $M_X(x)$ given $S_X$. According to Hypothesis 2,

$$ (M_X(x), \tilde{S}_X^{(n)}) \perp Z_X(x), $$

where $\perp$ stands for independence. This is because $\tilde{S}_X^{(n)}$ only depends on a finite number of $M_X(x_k)$. Therefore,

$$ Z_X(x) \perp M_X(x) \mid \tilde{S}_X^{(n)}. $$

Finally, since $\tilde{S}_X^{(n)}$ converges almost surely towards $\tilde{S}_X$, it comes that $Z_X(x)$ and $M_X(x)$ are conditionally independent given $S_X$.

**APPENDIX M**

**PROOF OF THEOREM 3**

**Proof:** Using the triangular inequality, we compute

$$ \left\| R_{m,q}(T_u X) - R_{m,q} X \right\|_2 \leq \left\| C_{2m}(T_u X) \right\|_2 \bar{P}_{\tau_u X} + \left\| C_{2m}(T_u X) - C_{2m} X \right\|_2 \bar{P}_X, $$

where $\bar{P}_X$ and $\bar{P}_{\tau_u X}$ are defined in (55). Since, by hypothesis, $\kappa \leq \pi/m$, expression (21) in Proposition 3 states that

$$ \left\| C_{2m}(T_u X) \right\|_2 = \left\| C_{2m} X \right\|_2. $$

Moreover, according to (63), we can apply (20) in Theorem 1 on the third term of the above expression. We get

$$ \left\| R_{m,q}(T_u X) - R_{m,q} X \right\|_2 \leq \left[ \bar{P}_{\tau_u X} + \bar{P}_X + \alpha(\kappa u) \right] \left\| C_{2m} X \right\|_2. $$

Then, by linearity of $E$, we get

$$ E[\bar{R}_{X,u}] \leq E[\bar{P}_{\tau_u X}] + E[\bar{P}_X] + \alpha(\kappa u). \quad (114) $$

For any stochastic process $X'$ satisfying Hypotheses 1 and 2, expression (61) in Theorem 2 and Jensen’s inequality yield:

$$ E[\bar{P}_{X'}] \leq (1 + \beta_q(m\kappa)) \gamma_q(m\xi). \quad (115) $$

Since Hypotheses 1 & 2 are satisfied for $Z_X$ and $M_X$, Lemma 3 implies that they are also true for $Z_{\tau_u X}$ and $M_{\tau_u X}$. Therefore, (115) is valid for $X' \leftarrow X$ and $X' \leftarrow \tau_u X$, and plugging this expression into (114) concludes the proof.

**APPENDIX N**

**PROOF OF PROPOSITION 6**

**Proof:** This proposition is a simple reformulation of the well-known result that two successive convolutions can be written as another convolution with a wider kernel. We introduce the upsampling operator: $(X \uparrow m)[n] := X[n/m]$ if $n/m \in \mathbb{Z}^2$, and 0 otherwise. We also consider the “identity” filter $I \in l^2(\mathbb{Z}^2)$ such that $I[0] = 1$ and $I[n] = 0$ otherwise.

First, for any $U, V \in l^2(\mathbb{Z}^2)$ and any $s, t \in \mathbb{N}^*$, we have $((U \downarrow s) * V) \downarrow t = (U * (V \uparrow s)) \downarrow (st)$. Then, a simple reasoning by induction yields the result, with

$$ V^{(0)}_l := I; \quad V^{(j+1)}_{4l+k} := V^{(j)}_l * (G_k \uparrow 2^j) $$

for any $l \in \{0 \ldots j - 1\}$ and any $k \in \{0 \ldots 3\}$.

**APPENDIX O**

**PROOF OF PROPOSITION 7**

**Proof:** For any $k \in \{0 \ldots 3\}$, Proposition 6 guarantees the existence of $V^{(j)}_{k,l} \in l^2(\mathbb{Z}^2)$ such that

$$ X^{(j)}_{k,l} := \left( X * V^{(j)}_{k,l} \right) \downarrow 2^j. \quad (116) $$

Then, the result is obtained by plugging (116) into (82) for $k \in \{0 \ldots 3\}$, and by denoting

$$ \begin{pmatrix} W^{(j)}_{l} \\ W^{(j)}_{4^l+k} \end{pmatrix} := \begin{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} & \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \end{pmatrix} \begin{pmatrix} V^{(j)}_{0,l} \\ V^{(j)}_{1,l} \end{pmatrix}.$$
C. Vasconcelos, H. Larochelle, V. Dumoulin, N. L. Roux, and N. Kingsbury, “Complex wavelets for shift invariant analysis and filter-