Appendix for "Transparent Vision: A Theory of Hierarchical Invariant Representations"

A1. Historical Perspectives of Invariance

We provide some historical perspectives on the invariance in the development of image representations. The quest for invariance dates back to the gestation of computer vision itself, spanning both hand-crafted and learning approaches [3]:

- In the hand-crafted approach, symmetry priors (*e.g.*, invariance and equivariance) *w.r.t.* geometric transformations (*e.g.*, translation, rotation, and scaling) have been recognized as main ideas in designing representations. Such ideas cover almost all classical and state-of-the-art methods, from global features (*e.g.*, moment invariants [25]), to local sparse features (*e.g.*, SIFT [20]), and to local dense features (*e.g.*, DAISY [33]). However, these hand-crafted representations are all fixed in design, relying on *(under)-complete* dictionaries, and therefore fail to provide sufficient discriminability at larger scales, *e.g.*, ImageNet classification task [28].
- In the learning approach, CNNs achieve *over-complete* representations of strong discriminative power for larger-scale vision tasks, through a cascade of learnable nonlinear transformations. As a textbook view of deep learning, representations should be learned not designed [18]. Therefore, typical CNN representations are equipped with very few symmetry priors, typically just translation equivariance [15], but has recently been proven to no longer hold in deeper layers of the CNNs with downsampling structures [42]. In general, these learning representations lack robustness and interpretability guarantees, *e.g.*, the presence and understanding of adversarial perturbations [6], and therefore cannot be well extended to trustworthy tasks [32].

Historically, to a certain extent, efforts at invariance and discriminability have developed independently in hand-crafted and learning approaches. The compatibility between invariance and discriminability has emerged as a tricky problem when moving towards trustworthy AI.

A2. Related Works

In this section, we supplement the developmental efforts of scattering and equivariant networks.

A2.1. Scattering Networks

Theoretical works further explored various geometric invariants [29], more general mathematical formulations [36], and the potential for improving the efficiency, interpretability, and robustness of state-of-the-art CNN techniques [24]. Regarding applications, they provided competitive results in a variety of tasks on audio [1], image [5, 24, 29] and graph [9] data, some of which are even interdisciplinary [10, 41].

A2.2. Equivariant Networks

Theoretical works further explored the equivariance for rotation [12, 35, 38], flipping [12], scaling [30, 37], and their combination [31] from various mathematical theories, including steerable filters [12], harmonic analysis [38], scale space [37], Lie groups [13], and B-spline interpolation [4]. Regarding applications, they played a key role in low-level vision tasks [39], especially scientific discoveries with symmetry priors [2, 34].

A3. Foundations of Invariant Theory

Our work develops from the theory of moment invariants. Therefore, we begin with a brief review on the foundations of moment invariants, covering some concepts, notations, and definitions.

A3.1. Global and Local Representations

In general, classical moments and moment invariants are *global* representations of images, where the theory is built on the following definition [25]:

$$\langle f, V_{nm} \rangle = \iint_D V_{nm}^*(x,y) f(x,y) dxdy, \tag{A1}$$

where f is the image function, V_{nm} is the basis function with order parameter $(n,m) \in \mathbb{Z}^2$ on domain D, and $*$ is the complex conjugate. Note that the domains of f and V_{nm} in (A1) have the same/similar location and scale, implying the global nature of the representation information.

With the sparse prior and geometric prior for natural images, two typical constraints, *i.e.*, orthogonality and rotation invariance, often imposed on the explicit definition of V_{nm},

leading to the following polar form:

$$\langle f, V_{nm} \rangle = \iint_D R_n^*(r) A_m^*(\theta) f(r,\theta) r dr d\theta, \tag{A2}$$

where $V_{nm}(\underbrace{r\cos\theta}_{x}, \underbrace{r\sin\theta}_{y}) \equiv V_{nm}(r,\theta)$ is separated as the product of the angular basis function $A_m(\theta) = \exp(\boldsymbol{\jmath} m\theta)$ ($\boldsymbol{\jmath} = \sqrt{-1}$) and the radial basis function R_n, subject to the weighted orthogonality condition $\int_0^1 R_n(r) R_{n'}^*(r) r dr = \frac{1}{2\pi}\delta_{nn'}$. Note that the basis function $V_{nm} = R_n A_m$ in (A2) is orthogonal on D, and the magnitude of $\langle f, V_{nm} \rangle$ is invariant to the rotation on the image f (see [25] for a survey).

In our recent work, moments and moment invariants are extended to *local* representations of images, where the theory is built on the following definition [26]:

$$\langle f, V_{nm}^{uvw} \rangle = \iint_D R_n^*(r') A_m^*(\theta') f(x,y) dx dy, \tag{A3}$$

where the new basis function V_{nm}^{uvw} introduces position parameters (u,v) and scale parameter w. It can be interpreted as a translated and scaled version of the global V_{nm} with the following coordinate relationship:

$$\begin{cases} r' = \frac{1}{w}\sqrt{(x-u)^2 + (y-v)^2} \\ \theta' = \arctan(\frac{y-v}{x-u}) \end{cases}, \tag{A4}$$

where the domain is a disk centered at (u,v) and with radius w: $D = \{(x,y) : (x-u)^2 + (y-v)^2 \leq w^2\}$. Note that (A3) allows the domain of V_{nm}^{uvw} to be built in different positions and scales *w.r.t.* the domain of f, implying the local nature of the representation information. Also, the classical definition (A2) is in fact a special case of the new definition (A3) with $(u,v) = (0,0)$ and $w = 1$ (see [26] for details).

A3.2. Invariance, Equivariance, and Covariance

The terms of invariance, equivariance, and covariance appear in the fields of computer vision, graphics, geometry, and physics. We use the following identities to generally denote such terms [19, 22]:

- invariance — $\mathcal{R}(\mathcal{D}(f)) \equiv \mathcal{R}(f)$,
- equivariance — $\mathcal{R}(\mathcal{D}(f)) \equiv \mathcal{D}(\mathcal{R}(f))$,
- covariance — $\mathcal{R}(\mathcal{D}(f)) \equiv \mathcal{D}'(\mathcal{R}(f))$,

where $\mathcal{R}$ is a representation, $\mathcal{D}$ is a degradation, and $\mathcal{D}'$ is a composite function of $\mathcal{D}$. Note that invariance and equivariance are special cases of covariance with $\mathcal{D}' = \mathrm{id}$ and $\mathcal{D}' = \mathcal{D}$.

Starting from the local representation (A3), one can verify that $\langle f, V_{nm}^{uvw} \rangle$ exhibits the following properties *w.r.t.* translation, rotation, flipping, and scaling on images (see [26] for details).

The image *translation* leads to

$$\begin{aligned} &\langle f(x+\Delta x, y+\Delta y), V_{nm}^{uvw}(x,y) \rangle \\ &= \left\langle f(x,y), V_{nm}^{(u+\Delta x)(v+\Delta y)w}(x,y) \right\rangle, \end{aligned} \tag{A5}$$

where $(\Delta x, \Delta y)$ is the translation offset of the image f. Note that the same $(\Delta x, \Delta y)$ appears in position parameters (u,v), implying the equivariance *w.r.t.* the image translation.

Since the translation equivariance holds, the following analysis (A6) $\sim$ (A8) will consider only center-aligned geometric transformations, *i.e.*, we can restrict $(u,v) = (0,0)$ without loss of generality.

The image *rotation* leads to

$$\begin{aligned} &\langle f(r,\theta+\phi), V_{nm}^{uvw}(r',\theta') \rangle \\ &= \langle f(r,\theta), V_{nm}^{uvw}(r',\theta') \rangle \, A_m^*(-\phi), \end{aligned} \tag{A6}$$

with $(u,v) = (0,0)$, where ϕ is the rotation angle *w.r.t.* the center of the image f. Note that the same ϕ appears in the phase of the representation, implying the covariance *w.r.t.* the center-aligned rotation. It is straightforward that the covariance (A3) will specialize to the invariance when taking the magnitude as $|\langle f(r,\theta+\phi), V_{nm}^{uvw}(r',\theta') \rangle| = |\langle f(r,\theta), V_{nm}^{uvw}(r',\theta') \rangle|$.

The image *flipping* leads to

$$\begin{aligned} &\langle f(r,-\theta), V_{nm}^{uvw}(r',\theta') \rangle \\ &= (\langle f(r,\theta), V_{nm}^{uvw}(r',\theta') \rangle)^*, \end{aligned} \tag{A7}$$

with $(u,v) = (0,0)$, where $f(r,-\theta)$ is a vertically flipped version of the image f *w.r.t.* the center. Note that center-aligned vertical flipping again only affects the phase of the representation, implying the covariance similar to (A6). As for other flipping orientations, the same conclusion can be derived from the composite of rotation and vertical flipping. It is straightforward that the joint invariance of center-aligned rotation and flipping holds when taking the magnitude of the representation.

The image *scaling* leads to

$$\begin{aligned} &\langle f(sx, sy), V_{nm}^{uvw}(x,y) \rangle \\ &= \left\langle f(x,y), V_{nm}^{uv(ws)}(x,y) \right\rangle, \end{aligned} \tag{A8}$$

with $(u,v) = (0,0)$, where s is the scaling factor *w.r.t.* the center of the image f. Note that the same s appears in the scale parameter w, implying the covariance *w.r.t.* center-aligned scaling.

For the representation properties when $(u,v) \neq (0,0)$, they can be derived from the composite of translation with center-aligned rotation, flipping, and scaling, respectively. Hence, the magnitude of the representation has *joint equivariance for any translation, rotation, and flipping* on (u,v) domain, as well as *covariance for any scaling* on w domain.

A4. Proofs

A4.1. Equivariance Properties

Proof. First, let us examine the behavior of a representation unit $\mathbb{U}$ on $\mathfrak{G}_1$:

$$\begin{aligned}\mathbb{U}(\mathfrak{g}_1 M) &= \mathbb{P}\circ\mathbb{S}\circ\mathbb{C}(\mathfrak{g}_1 M)\\ &= \mathbb{P}\circ\mathbb{S}\circ\mathfrak{g}_1'\mathbb{C}(M)\\ &= \mathbb{P}\circ\mathfrak{g}_1\mathbb{S}\circ\mathbb{C}(M)\\ &= \mathfrak{g}_1\mathbb{P}\circ\mathbb{S}\circ\mathbb{C}(M)\\ &= \mathfrak{g}_1\mathbb{U}(M),\end{aligned} \tag{A9}$$

where the first pass comes from the covariance of $\mathbb{C}$ for rotation and flipping, *i.e.*, (A6) and (A7), and $\mathfrak{g}_1'$ is a predictable operation acting in the phase domain of $\mathbb{C}(M)$; the second pass comes from the specialization of $\mathbb{S}$ to the covariant $\mathfrak{g}_1'$ – the magnitude operation removes the extra phase variations, leading to a pure equivariance $\mathfrak{g}_1$; the third pass comes from the identity function of $\mathbb{P}$, which becomes approximately equal when the downsampled $\mathbb{P}$ is used.

Here, $\mathbb{U}(\mathfrak{g}_1 M) = \mathfrak{g}_1\mathbb{U}(M)$ means that the representation unit $\mathbb{U}$ can be considered as an *equivariant layer* for any $\mathfrak{g}_1 \in \mathfrak{G}_1$ and $M \in X$ – in other words, the single $\mathbb{U}$ and $\mathfrak{g}_1$ operations on $M \in X$ are *exchangeable*. Furthermore, with a notation $M_{[l]} \triangleq \mathbb{U}_{[l]}\circ\cdots\circ\mathbb{U}_{[1]}(M) = \mathbb{U}_{[l]}M_{[l-1]}$, we have $M_{[l]} \in X$ for any $l \in \{1, 2, \cdots, L\}$. Therefore, $\mathfrak{g}_1$ and any composition of $\mathbb{U}$ are exchangeable, implying the correctness of Property 1. □

A4.2. Covariance Properties

Proof. First, let us examine the behavior of a representation unit $\mathbb{U}^w$ on $\mathfrak{G}_2$:

$$\begin{aligned}\mathbb{U}^w(\mathfrak{g}_2 M) &= \mathbb{P}\circ\mathbb{S}\circ\mathbb{C}^w(\mathfrak{g}_2 M)\\ &= \mathbb{P}\circ\mathbb{S}\circ\mathfrak{g}_2'\mathbb{C}^w(M)\\ &= \mathbb{P}\circ\mathbb{S}\circ\mathfrak{g}_2\mathbb{C}^{ws}(M)\\ &= \mathbb{P}\circ\mathfrak{g}_2\mathbb{S}\circ\mathbb{C}^{ws}(M)\\ &= \mathfrak{g}_2\mathbb{P}\circ\mathbb{S}\circ\mathbb{C}^{ws}(M)\\ &= \mathfrak{g}_2\mathbb{U}^{ws}(M)\\ &= \mathfrak{g}_2'\mathbb{U}^w(M),\end{aligned} \tag{A10}$$

where the first pass comes from the covariance of $\mathbb{C}$ for scaling, *i.e.*, (A8), and $\mathfrak{g}_2'$ is a predictable operation acting in both the Ω domain (*i.e.*, the same scaling $\mathfrak{g}_2$) and the w domain (*i.e.*, the factor s) of $\mathbb{C}^w(M)$; the second and third passes come from the element-wise act of $\mathbb{S}$ and the identity function of $\mathbb{P}$, respectively.

Here, $\mathbb{U}^w(\mathfrak{g}_2 M) = \mathfrak{g}_2'\mathbb{U}^w(M)$ means that the representation unit $\mathbb{U}^w$ can be considered as an *covariant layer* for any $\mathfrak{g}_2 \in \mathfrak{G}_2$ and $M \in X$ – in other words, the single $\mathbb{U}^w$ and $\mathfrak{g}_2$ operations on $M \in X$ are *exchangeable but with the parameter changing* of ws. Furthermore, we have $M_{[l]} \in X$ for any $l \in \{1, 2, \cdots, L\}$. Therefore, $\mathfrak{g}_2$ and any composition of $\mathbb{U}^w$ are exchangeable while changing the scale parameter to ws, implying the correctness of Property 2. □

A4.3. Invariance Properties

Proof. We can rewrite $\mathbb{I}(\mathfrak{g}_0' M)_{[L]}$ as:

$$\begin{aligned}\mathbb{I}(\mathfrak{g}_0 M)_{[L]} &= \mathbb{I}\circ\mathbb{U}_{[L]}\circ\cdots\circ\mathbb{U}_{[2]}\circ\mathbb{U}_{[1]}(\mathfrak{g}_0 M)\\ &= \mathbb{I}(\mathfrak{g}_0'\mathbb{U}_{[L]}\circ\cdots\circ\mathbb{U}_{[2]}\circ\mathbb{U}_{[1]}(M))\\ &= \mathbb{I}\circ\mathbb{U}_{[L]}\circ\cdots\circ\mathbb{U}_{[2]}\circ\mathbb{U}_{[1]}(M)\\ &= \mathbb{I}M_{[L]},\end{aligned} \tag{A11}$$

where the first pass comes from Properties 1 and 2, note that $\mathfrak{g}_0 \in \mathfrak{G}_0 \subseteq \mathfrak{G}_1 \times \mathfrak{G}_2$, $\mathfrak{g}_0'$ is related to $\mathfrak{g}_1$ and $\mathfrak{g}_2'$; the second pass comes from our assumption $\mathbb{I}(\mathfrak{g}_0' M) = \mathbb{I}(M)$ for any $\mathfrak{g}_0 \in \mathfrak{G}_0$ and $M \in X$, with $M_{[l]} \in X$ for any $l \in \{1, 2, \cdots, L\}$. Therefore, $\mathbb{I}(\mathfrak{g}_0' M)_{[L]} \equiv \mathbb{I}M_{[L]}$, implying the correctness of Property 3. □

A5. Theoretical Comparisons

It is necessary to highlight the theoretical relationships with typical related works:

- **Traditional Invariants:** Our work generalizes this theory by unifying the global and local invariant representations into a new framework of HIR. More specifically, we formalize layers $\mathbb{C}$, $\mathbb{S}$, and $\mathbb{P}$ based on the theory of local invariants [26] (Definitions $1 \sim 3$), arguing the equivariance/covariance can be preserved across layers under a certain cascade (Properties $1 \sim 2$). We also formalize layer $\mathbb{I}$ based on the theory of global invariants [25] (Definition 4), arguing the successes of global invariance for image domains can be directly generalized to equivariant/covariant deep feature domains (Property 3). Under our hierarchical invariance, classical global [25] and local [26] invariants can be considered as special cases, *i.e.*, $\mathbb{I}f$ and $\mathbb{I}\circ\mathbb{S}\circ\mathbb{C}f$ (Definition 5).
- **Traditional CNNs:** Our work has a similar hierarchical architecture but with better properties in geometric symmetry, allowing for robust and interpretable image representations. More specifically, we introduce the discriminative design of CNNs in our invariants, *i.e.*, overcomplete representation with deep cascading [42]. On the other hand, we criticize typical CNN modules (Formulation 1), allowing fully transparent geometric symmetries across layers of our representation (Properties $1 \sim 3$). As a result, the proposed representation serves as an effective alternative to the highly black-box CNNs in trustworthy tasks.
- **Scattering Networks:** Our work is more compact in achieving rotation invariance. As a main competitor, scattering networks are also based on deep cascading of

explicit transforms (wavelets) [5], with similar concepts to our work. However, constructing rotation invariants from scattering networks is complicated, which requires parallel convolution and cross-channel pooling of multiple oriented wavelets; increasing the orientation sampling will result in an exponential growth of the complexity. Whereas our approach benefits from classical invariant theory, rotation invariance is continuous and one-shot (Property 1), providing better efficiency while easily enlarging the network size to improve the representation capacity.

- **Equivariant Networks:** Our work is non-learning while being more compact in achieving continuous and joint invariance. As a secondary competitor, equivariant networks are also guaranteed by group theory [11], with similar concepts to our work. However, the convolutional layers in equivariant networks are learned, leading to varying degrees of data dependence. In particular, it has a similar parallel structure to scattering networks, leading to exponential complexity and optimization challenges. Although equivariant networks are a very generic design, our approach provides better efficiency for continuous and joint invariance (Properties $1 \sim 3$), while easily enlarging the network size to improve the representation capacity.

A6. Practical Details

In this section, we discuss more practical details on numerical implementations, network parameters, layer settings.

A6.1. Fast and Accurate Implementation

We will complement the numerical implementation of HIR, especially the fast and accurate computations of Definition 1 from our previous work [26]. Note that the discussion here is very general, with no restrictions on the specific definitions of the basis functions.

Definition (Fast Implementation)**.** *Let us introduce the **convolution theorem** as a fast implementation of Definition 1, such that the spatial domain convolution of (A9) can be converted to the following frequency domain product form [26]:*

$$\mathbb{C}M = \mathcal{F}^{-1}(\mathcal{F}(M(i,j;k)) \odot \mathcal{F}((H^{w}_{nm}(i,j))^{T})), \quad \text{(A12)}$$

where $\mathcal{F}$ is the Fourier transform and $\odot$ is the point-wise multiplication.

Property (Complexity Analysis)**.** *In Definition 1, the (1) dominates the computational complexity due to the dense convolution. For the input feature map $M(i,j;k)$ with $\Omega = \{1,2,\cdots,N_i\} \times \{1,2,\cdots,N_j\}$ and $\mathrm{H} = \mathfrak{C}^K$, we assume that a set of $\mathbb{C}M$ needs to be computed, where the scale parameter $w \in S_w$ with a fixed order (n,m) and a fixed channel k, and denote the number of feature map samples as $N_{ij} = N_i N_j$ and the number of scale samples as $N_w = |S_w|$. With the above definition and the Fast Fourier Transform (FFT), we can compute the set of $\mathbb{C}M$ in $\mathcal{O}(N_w N_{ij} \log N_{ij})$ multiplications, as opposed to the complexity of $\mathcal{O}(N_w N_{ij} {w_{\max}}^2)$ by the direct Definition 1, where $w_{\max}$ is the maximum scale in S_w. Note that the big difference between square and logarithmic growths in the complexity (removing the same terms), where the above definition will exhibit batter efficiency when $w_{\max}$ is sufficiently large such that ${w_{\max}}^2 > \log N_{ij}$.*

Definition (Accurate Implementation)**.** *Let us introduce the **higher-order numerical integration** as an accurate implementation of Definition 1, such that the two-dimensional continuous integral of (3) can be converted to the following summation form [26]:*

$$h^{uvw}_{nm}(i,j) \simeq \sum_{(a,b)\in S_{ab}} c_{ab} (V^{uvw}_{nm}(x_a, y_b))^{*} \frac{\Delta i \Delta j}{w^2}, \quad \text{(A13)}$$

where the set of numerical integration samples S_{ab} encodes the points $(x_a, y_b) \in D_{ij}$ and the corresponding weights c_{ab}, which are specified by a certain numerical integration strategy, such as Gaussian quadrature.

Property (Accurate Analysis)**.** *In Definition 1, (3) dominates the computational accuracy due to the continuous integration of complicated functions. We assume that h^{uvw}_{nm} with a fixed order (n,m) and position (u,v) needs to be computed, and denote the number of numerical integration samples as $N_{ab} = |S_{ab}|$. The implementation based on the above definition exhibits an approximation error of $\mathcal{O}((\frac{\Delta i \Delta j}{w^2})^{N_{ab}+1})$. Note that when there is more than one sample within each pixel region, i.e., $N_{ab} > 1$, the above definition will exhibit batter accuracy than the error of $\mathcal{O}((\frac{\Delta i \Delta j}{w^2})^2)$ by the direct Definition 1 (zero-order approximation).*

A6.2. Parameters of Single-scale Networks

Here, the order parameter (n,m) of the previous unit (blue) is always smaller than that of the subsequent ones (under a specific norm), so that the path exhibits an increasing trend in the order. With this design, the main information can be passed through the early nodes, and hence the subsequent nodes capture rich features. Also, the identity function (black) is introduced as a skip-connection trick, allowing the information to be passed to deeper nodes. In this paper, all units from the same level l are specified separately from the set $\{(n,m) : n+m = l, (n,m) \in \mathbb{N}^2\}$, *i.e.*, their orders are equal under the ℓ_1 norm.

A6.3. Radial Basis Functions

In our previous work [27], two generic classes of radial basis functions have been introduced, based on a family of

harmonic functions:

$$R_n(\alpha, r) = \sqrt{\frac{\alpha r^{\alpha-2}}{2\pi}} \exp(\boldsymbol{j} 2n\pi r^{\alpha}), \tag{A14}$$

and a family of polynomial functions:

$$R_n^{\alpha}(p, q, r) = \sqrt{\frac{\alpha r^{\alpha q-2}(1-r^{\alpha})^{p-q}(p+2n)\Gamma(q+n)n!}{2\pi\Gamma(p+n)\Gamma(p-q+n+1)}} \times \sum_{k=0}^{n} \frac{(-1)^k \Gamma(p+n+k) r^{\alpha k}}{k!(n-k)!\Gamma(q+k)}, \tag{A15}$$

respectively, where the fractional parameter $\alpha \in \mathbb{R}^+$, the polynomial parameters $p, q \in \mathbb{R}$ must fulfill $p - q > -1$ and $q > 0$. Both classes of functions can be used to define R_n in the (A2), satisfying the orthogonality condition.

For the sake of simplicity, a family of cosine functions are chosen in all experiments and applications, as a special case of the (A14):

$$R_n(r) = \begin{cases} \frac{1}{\sqrt{\pi}} & n = 0 \\ \sqrt{\frac{2}{\pi}} \cos(n\pi r^2) & n > 0 \end{cases}, \tag{A16}$$

i.e., forming a hierarchical invariant version of the Polar Cosine Transform (PCT) [40]. Note that we try to show the superiority of the hierarchical invariant framework itself, even if relying on naive (A16).

A6.4. Invariant Layer

In the monograph [14] and our previous work [25], a number of strategies for directly constructing global invariants in image domains have been presented. They can be naturally used to define $\mathcal{I}$ in (6), with the equivariant or covariant behavior of deep feature maps (Properties $1 \sim 3$). In all experiments and applications of this paper, a class of global invariants is concisely designed based on frequency pooling.

Regarding (6), we first let the Fourier basis be $V_{nm}(x_i, y_j)$. Note that the Fourier Transform (FT) is highly understood in the signal processing community and can be considered a good foundation for interpretability. Then, based on the order/frequency sampling of the FT $(n, m) \in [-K, K]^2$, we define $\mathcal{I}$ as a frequency-band integral in the polar system:

$$\mathcal{I}(\{\langle M, V_{nm}\rangle\}) \triangleq \{I_i = \sum_{(n,m)\in\mathcal{B}_i} \{|\langle M, V_{nm}\rangle|\} : i = 1, 2, \cdots, \#_B\}, \tag{A17}$$

where $\mathcal{B}_i = \{(n, m) : \sqrt{2}K(i-1)/\#_B| \leq ||(n, m)||_2 \leq \sqrt{2}Ki/\#_B\}$ is the i-th frequency band under the ℓ_2 norm, with the number of bands $\#_B$.

Here, we can state that the above feature vector $\{I_i : i = 1, 2, ..., \#_B\}$ directly satisfies the invariance for $\mathfrak{G}_1$, in light of Property 1 and the translation, rotation and flipping properties of FT. As for scaling, $\mathcal{I}$ is compatible with both single-scale and multi-scale networks: 1) regarding the single-scale case, a certain degree of robustness is provided for $\mathfrak{G}_2$ (at least up to the bandwidth), in light of Property 2 and the scaling property of FT; 2) regarding the multi-scale case, the scaling covariance has been eliminated before feeding into $\mathcal{I}$, and thus will satisfy the joint invariance for $\mathfrak{G}_0 = \mathfrak{G}_1 \times \mathfrak{G}_2$.

Note that the well-known average pooling is in fact a special case of (A17), with $K = 0$ and $\#_B = 1$. Our frequency-band integral $\mathcal{I}$ can be regarded as a generic design of global pooling, with comprehensive consideration on interpretability, invariance, and discriminability.

A7. Implementation Details of Experiments/Applications

All experiments/applications are executed in Matlab R2023a under Microsoft Windows environment, based on 2.90-GHz CPU, RTX-3060 GPU, and 16-GB RAM.

Experiment: Our HIR is implemented here as a single-scale network, where scale parameter $w = 10$ and composition length $L = 6$; its invariant layer (A17) is specialized to the average pooling, with $K = 0$ and $\#_B = 1$, for a fair comparison with the deep representations by average pooling. Note that the adaptability strategies of Section 3.2 are not employed here, for a direct assessment of its discriminative power. All features are fed into a PCA classifier, trained on features of the training set. Unless otherwise stated, the training and testing sets are formed without any crossover by random sampling at 80% and 20% ratios on the original dataset, respectively.

Application: Our HIR is implemented here as a single-scale network, where scale parameter $w = 10$ and composition length $L = 7$; its invariant layer (A17) is specialized with $K = N_{ij}/2$ and $\#_B = 30$, for improving the discriminability of digital artifacts. Note that the feature/architecture selection strategy of Section 3.2 is employed for data adaptability and discriminability, where the top-ranked 500- and 1000-dimensional features are selected for AIGC and adversarial perturbation, respectively. All features are fed into both NN and SVM classifiers, for evaluating the sensitivity *w.r.t.* the classifier. Unless otherwise stated, the training and testing sets are formed without any crossover by random sampling at 50% and 50% ratios on the original dataset, respectively.

Table A1. Classification scores (%) and runtime (second) for different representations on a small-scale texture benchmark.

Method	Time	**Original**			**Orien. & Flip.**		
	GPU†	Pre.	Rec.	F1	Pre.	Rec.	F1
Classical:							
Cosine	5	70.74	67.50	66.85	69.65	66.25	65.30
Wavelet	6	69.43	64.38	64.68	62.34	58.13	57.82
Kraw.	5	70.67	67.50	66.30	64.41	60.00	59.55
Learning:							
SimpleNet	52†	70.33	67.50	67.09	54.63	43.13	41.31
SimpleNet+	52†	46.93	49.38	46.06	47.18	48.13	44.93
AlexNet	42†	98.82	98.75	98.75	91.69	91.25	91.28
AlexNet+	41†	87.61	84.38	84.05	88.37	85.63	85.76
VGGNet	266†	99.41	99.38	99.37	92.18	91.25	91.37
VGGNet+	609†	91.34	90.00	89.81	92.15	91.25	91.08
Invariant:							
ScatterNet	42	98.89	98.75	98.75	84.98	83.13	83.08
HIR	27	96.98	96.88	96.87	96.32	96.25	96.23

A8. Supplementary Experiments/Applications

A8.1. Texture Experiments

As shown in Fig. 4, the experiment is executed on dataset KTH-TIPS[1], a typical benchmark for texture image classification. This dataset has 10 classes, each containing 81 instances, the total size is $10 \times 81 = 810$, and hence is considered as a small-scale vision problem.

As shown in Table A1, we list performance scores of the competing representations on this benchmark, as well as the elapsed time, *i.e.*, CPU featuring time or GPU training time. Besides this direct protocol on the original dataset, we also consider testing image variants with random orientation (*w.r.t.* $\{0, 90, 180, 270\}$ degree) or flipping (*w.r.t.* x or y axis).

- The classical (over-)complete representations fail to achieve a satisfactory level of discriminability, even in the direct protocol of such small-scale benchmark.
- The learning CNN family achieves significantly higher scores due to its over-complete and data-adaptive properties, especially the AlexNet and VGGNet with large-scale pre-training and transfer learning. Whereas, the SimpleNet performs relatively poorly, indicating the sensitivity of learning to network size and training strategy. Under the variant protocol, they exhibit a significant performance degradation, suggesting the learned features lack invariance *w.r.t.* natural geometric variations of texture. After introducing the augmented training, the CNN scores become more stable, but at the cost of discriminability. A potential reason for this phenomenon is the small amount of training data. Moreover, the computational cost is considerable for this small-scale problem, and a certain training instability is observed.

[1] https://www.csc.kth.se/cvap/databases/kth-tips/index.html

Table A2. Classification scores (%) and runtime (second, for train./test. = 8/2) for different representations on a large-scale parasite benchmark.

Method	Time	**Train./Test. = 8/2**			**Train./Test. = 1/9**		
	GPU†	Pre.	Rec.	F1	Pre.	Rec.	F1
Classical:							
Cosine	37	36.19	32.60	29.85	49.40	41.97	43.80
Wavelet	39	41.68	45.20	41.79	53.69	47.97	49.27
Kraw.	42	66.56	69.49	67.21	71.60	57.88	61.10
Learning:							
SimpleNet	2244†	90.15	89.25	89.65	84.51	76.14	78.84
AlexNet	1796†	98.87	98.40	98.63	95.92	94.69	95.27
VGGNet	9184†	99.24	98.97	99.11	97.95	97.37	97.65
Invariant:							
ScatterNet	1277	68.41	69.71	67.55	72.52	63.30	65.70
HIR	823	88.73	92.18	90.10	91.26	88.76	89.85

- The scattering networks provide a high level of discriminability and robustness without feature training and data augmentation, indicating the success of extending classical wavelets to deep representations.
- Our work further extends such success: the HIR achieves a similar level of discriminability as the learning CNN family, while exhibiting superior robustness in the variant protocol than all competing representations. In particular, such representation success build on our compact and efficient framework, with lower runtimes than scattering networks and learning CNN family.

A8.2. Parasite Experiments

As shown in Fig. 4, the experiment is executed on micrographic dataset[2], a typical benchmark for parasite image classification. This dataset has 6 parasite classes and 2 host classes, with real-world diversity regarding imaging, background, morphology, and geometry, the total size is 34298, and hence is considered as a large-scale vision problem.

As shown in Table A2, we list performance scores and elapsed times of the competing representations on this benchmark. Note that we also consider a protocol with different training-testing ratios to analyze the data dependence and sample efficiency.

- In this large-scale problem, the scores of the classical representations drop further, implying a limited level of discriminability. On the other hand, their performance is relatively stable when training samples are reduced, and even better in the $1/9$ case, indicating a good efficiency.
- In the learning CNN family, the direct-learning SimpleNet exhibits a clear data dependence. Specifically, it achieves $\sim$90% scores in the $8/2$ case (similar to HIR), while the scores drop significantly in the $1/9$ case (below than HIR). In contrast, the AlexNet and VGGNet achieve

[2] https://data.mendeley.com/datasets/38jtn4nzs6/3

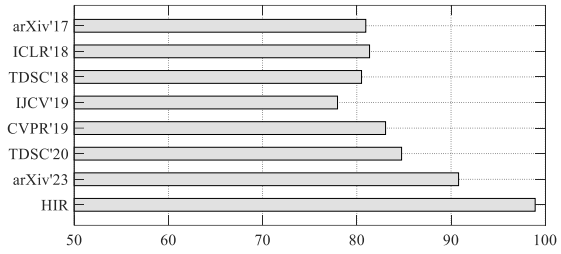

Figure A1. A comparison of adversarial perturbation forensic scores (F1, %) *w.r.t.* current forensic solutions on the UP benchmark.

good discriminability and stability in the $1/9$ case, revealing that the transfer strategy effectively inherits the pre-training prior on ImageNet. On the other hand, the cost of pre-training and transfer learning is still considerable, without guaranteed robustness or adaptability for a given data domain.

- Despite outperforming the original wavelets, scattering networks fail to provide a competitive discriminability in the era of deep learning. Here, the common failure of such hand-crafted representations on larger-scale discriminability can be regarded as important evidence for our motivation.
- The HIR achieves a SimpleNet-level discriminability, outperforming our competitor scattering networks significantly. Also, the HIR is not sensitive to the reduction of training samples, outperforming the learning CNN family in data dependence and sample efficiency. Note that the discriminability of the fixed features from HIR is still lower than the transfer learning with large-scale pre-training. Therefore, in the next applications, the HIR features will be empowered with data adaptability strategies in Section 3.2.

A8.3. Adversarial Perturbation Forensic Applications

As shown in Fig. 5, the dataset ImageNet[3] is perturbed through 6 adversarial methods[4], *i.e.*, BIM [17], CW [7], Damage [8], FGSM [16], PGD [21], and UP [23], respectively, resulting in 6 benchmarks, each containing 5000 clean images and 5000 perturbed versions. This task exhibits real-world discriminative challenges, in light of the rich variability of the perturbations themselves and the underlying ImageNet.

In Fig. A1, we first provide a comparison with the current solutions of perturbation forensics on the basic and realistic UP benchmark. Despite the fixed perturbation pattern, there are still competing methods failing to achieve good scores. Such methods are with under-complete representations, and thereby unable to comprehensively capture perturbation patterns. In contrast, over-complete arXiv'23 and our HIR all achieve $> 90\%$ scores, further revealing the fundamental role of representation in forensic tasks. Thus, we will next further compare relevant representation strategies.

[3]https://www.image-net.org/

[4]https://github.com/Harry24k/adversarial-attacks-pytorch

Table A3. Adversarial perturbation forensic scores (F1, %) for different representations *w.r.t.* various types of perturbations.

Method	BIM	CW	DAmage	FGSM	PGD	UP	**AVG**	**MIN**
Classical:								
Cosine NN	34.63	33.19	90.78	39.80	34.69	2.22	39.22	2.22
Cosine SVM	79.57	83.34	97.26	78.24	79.22	96.68	85.72	78.24
Wavelet NN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Wavelet SVM	72.83	82.09	97.77	78.21	71.80	95.87	83.10	71.80
Kraw. NN	66.43	66.49	90.86	66.43	66.44	0.00	59.44	0.00
Kraw. SVM	0.00	55.87	0.00	56.44	0.00	70.37	30.45	0.00
Learning:								
SimpleNet	4.24	3.24	92.13	49.89	33.13	99.86	47.08	3.24
AlexNet	90.20	72.72	96.63	94.61	90.91	98.45	90.59	72.72
VGGNet	96.04	62.50	99.08	98.12	96.99	99.15	91.98	62.50
GoogLeNet	90.29	80.04	97.09	95.29	89.94	98.75	91.90	80.04
ResNet	90.22	75.59	97.35	94.66	90.17	98.40	91.07	75.59
DenseNet	98.93	90.19	99.34	99.23	98.85	99.76	97.72	90.19
InceptionNet	98.70	85.14	97.38	97.32	98.66	99.41	96.10	85.14
MobileNet	92.51	82.67	97.37	96.81	92.10	98.19	93.27	82.67
Invariant:								
ScatterNet NN	81.30	70.23	95.27	91.17	82.65	94.64	85.88	70.23
ScatterNet SVM	84.40	69.49	96.77	90.57	83.86	95.12	86.70	69.49
HIR NN	89.66	84.92	98.89	93.26	90.08	97.78	92.43	84.92
HIR SVM	92.30	89.10	99.30	95.96	91.60	98.93	94.53	89.10

In Table A3, we train and test all representations on the 6 benchmarks, presenting the corresponding F1 scores, as well as the average and worst score statistics. This protocol exhibits richer intra-class variability over the fixed perturbation.

- The frequency difference between natural and perturbed data is a fruitful forensic clue. Therefore, the classical (time)-frequency representations achieve higher scores than generally expected on this large-scale problem. However, such features exhibit significant sensitivity to classifiers. A potential reason is the restricted separability, where one must resort to complex classification strategies in the feature space.
- In the learning CNN family, all large-scale networks exhibit $> 90\%$ average scores, especially DenseNet and InceptionNet. The phenomenon suggests that the transfer learning is good at capturing discriminative features with sufficient training data and aligned testing protocol. As for the attacks, the CW is more challenging and dominates the worst scores, mainly due to its variable and weak patterns.
- The scattering networks achieve similar scores and much better classifier stability than the original wavelets, suggesting an improvement in the separability. However, its average scores did not reach 90%, failing to provide a similar level of discriminability as learning CNNs.
- Our HIR is very robust to classifier changes, also achieving a MobileNet-level of discriminability, slightly lower

Table A4. Adversarial perturbation forensic scores (%) for different representations on a real-world (hybrid) benchmark.

Method	Train./Test. = 5/5			Train./Test. = 1/9		
	Pre.	Rec.	F1	Pre.	Rec.	F1
Classical:						
Cosine NN	0.00	0.00	0.00	0.00	0.00	0.00
Cosine SVM	79.08	73.33	76.10	81.13	68.79	74.45
Wavelet NN	0.00	0.00	0.00	0.00	0.00	0.00
Wavelet SVM	77.53	66.95	71.85	76.05	61.13	67.78
Kraw. NN	50.53	15.22	23.40	50.00	15.10	23.20
Kraw. SVM	50.03	65.34	56.67	49.75	48.77	49.26
Learning:						
SimpleNet	47.31	48.11	47.71	50.59	63.63	56.36
AlexNet	81.46	87.35	84.30	72.24	61.36	66.35
VGGNet	81.41	90.04	85.51	78.83	75.35	77.05
GoogLeNet	82.74	85.46	84.08	63.35	57.74	60.42
ResNet	80.93	84.70	82.77	68.48	66.64	67.55
DenseNet	87.92	93.25	90.51	82.07	83.96	83.00
InceptionNet	84.60	90.92	87.65	69.58	70.77	70.17
MobileNet	83.07	88.07	85.50	68.73	69.50	69.11
Invariant:						
Scatter. NN	69.85	68.94	69.39	74.93	77.31	76.10
Scatter. SVM	75.70	72.07	73.84	76.42	78.63	77.51
HIR NN	81.27	80.68	80.98	79.09	82.17	80.60
HIR SVM	86.20	86.06	86.13	83.42	83.29	83.35

than DenseNet and InceptionNet, and significantly better than the direct competitor scattering networks. Therefore, our strategy has a better combined performance in robustness, interpretability, and discriminability. Its efficiency benefit will be highlighted in the next experimental protocol.

In Table A4, we train and test all representations on a hybrid of the 6 perturbation benchmarks, presenting scores at two training-testing ratios. This protocol is more challenging due to very complex intra-class variability, while being more practical for real-world forensic scenarios.

- In line with previous observations, the classical representations still exhibit score fluctuations on the two classifiers. We also note a performance degradation compared to the case of Table A3, due to the discriminative challenges by this hybrid protocol. On the other hand, their performance is stable *w.r.t.* the reduction of training samples, further validating the inherent advantages in sample efficiency.
- Moving into this hybrid benchmark, the learning CNN family yields consistent and large performance degradation, especially for the $1/9$ case with fewer samples. This phenomenon is direct evidence for the data dependence in learning representations (even with transfer strategy). In fact, real-world forensics often face the situation where the perturbation types are diverse and some of them lack samples. Therefore, such data-dependent forensics typically exhibit time-consuming (re-)training, while failing to guarantee their validity for under-sampled perturbation patterns.
- The scattering networks basically continue the discriminability level and classifier stability from Table A3. Note that its scores in the $1/9$ case are higher than most classical and learning representations, reflecting the superior performance in both discriminability and efficiency.
- In this challenging protocol, the hand-crafted HIR still achieves a learning-level discriminability and consistently outperforms scattering networks. More importantly, our HIR is significantly less dependent on training samples than learning CNNs, meaning it can better cope with under-sampled perturbation patterns in practice. For the AIGC forensic task, the comprehensive advantages of HIR over learning CNNs will be further highlighted, in robustness, interpretability, discriminability, and efficiency.

References

[1] Joakim Andén and Stéphane Mallat. Deep scattering spectrum. *IEEE Trans. Signal Process.*, 62(16):4114–4128, 2014. 1

[2] Kenneth Atz, Francesca Grisoni, and Gisbert Schneider. Geometric deep learning on molecular representations. *Nature Mach. Intell.*, 3(12):1023–1032, 2021. 1

[3] V Balntas, K Lenc, A Vedaldi, T Tuytelaars, J Matas, and K Mikolajczyk. H-Patches: A benchmark and evaluation of handcrafted and learned local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(11):2825–2841, 2019. 1

[4] Erik J Bekkers. B-spline CNNs on Lie groups. In *Proc. Int. Conf. Learn. Representations*, 2019. 1

[5] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35 (8):1872–1886, 2013. 1, 4

[6] Cameron Buckner. Understanding adversarial examples requires a theory of artefacts for deep learning. *Nature Mach. Intell.*, 2(12):731–736, 2020. 1

[7] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Proc. IEEE Symp. Secur. Privacy*, pages 39–57, 2017. 7

[8] Sizhe Chen, Zhengbao He, Chengjin Sun, Jie Yang, and Xiaolin Huang. Universal adversarial attack on attention and the resulting dataset DamageNet. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(4):2188–2197, 2020. 7

[9] Xu Chen, Xiuyuan Cheng, and Stéphane Mallat. Unsupervised deep Haar scattering on graphs. *Proc. Adv. Neural Inf. Process. Syst.*, 27, 2014. 1

[10] Sihao Cheng, Yuan-Sen Ting, Brice Ménard, and Joan Bruna. A new approach to observational cosmology using the scattering transform. *Mon. Not. R. Astron. Soc.*, 499(4): 5902–5914, 2020. 1

[11] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proc. Int. Conf. Mach. Learn.*, pages 2990–2999, 2016. 4

[12] Taco S Cohen and Max Welling. Steerable CNNs. In *Proc. Int. Conf. Learn. Representations*, 2016. 1

[13] Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to Lie groups on arbitrary continuous data. In *Proc. Int. Conf. Mach. Learn.*, pages 3165–3176, 2020. 1

[14] Jan Flusser, Barbara Zitova, and Tomas Suk. *Moments and moment invariants in pattern recognition*. John Wiley & Sons, 2009. 5

[15] Kunihiko Fukushima and Sei Miyake. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognit.*, 15(6):455–469, 1982. 1

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 7

[17] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*, pages 99–112. Chapman and Hall/CRC, 2018. 7

[18] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. 1

[19] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 991–999, 2015. 2

[20] David G Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60:91–110, 2004. 1

[21] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proc. Int. Conf. Learn. Representations*, 2018. 7

[22] Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 5048–5057, 2017. 2

[23] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1765–1773, 2017. 7

[24] Edouard Oyallon, Sergey Zagoruyko, Gabriel Huang, Nikos Komodakis, Simon Lacoste-Julien, Matthew Blaschko, and Eugene Belilovsky. Scattering networks for hybrid representation learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41 (9):2208–2221, 2018. 1

[25] Shuren Qi, Yushu Zhang, Chao Wang, Jiantao Zhou, and Xiaochun Cao. A survey of orthogonal moments for image representation: theory, implementation, and evaluation. *ACM Comput. Surv.*, 55(1):1–35, 2021. 1, 2, 3, 5

[26] Shuren Qi, Yushu Zhang, Chao Wang, Jiantao Zhou, and Xiaochun Cao. A principled design of image representation: Towards forensic tasks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(5):5337–5354, 2022. 2, 3, 4

[27] Shuren Qi, Yushu Zhang, Chao Wang, Tao Xiang, Xiaochun Cao, and Yong Xiang. Representing noisy image without denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 4

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115: 211–252, 2015. 1

[29] Laurent Sifre and Stéphane Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1233–1240, 2013. 1

[30] Ivan Sosnovik, Michał Szmaja, and Arnold Smeulders. Scale-equivariant steerable networks. In *Proc. Int. Conf. Learn. Representations*, 2019. 1

[31] Zikai Sun and Thierry Blu. Empowering networks with scale and rotation equivariance using a similarity convolution. In *Proc. Int. Conf. Learn. Representations*, 2023. 1

[32] Mariarosaria Taddeo, Tom McCutcheon, and Luciano Floridi. Trusting artificial intelligence in cybersecurity is a double-edged sword. *Nature Mach. Intell.*, 1(12):557–560, 2019. 1

[33] Engin Tola, Vincent Lepetit, and Pascal Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(5):815–830, 2009. 1

[34] Raphael JL Townshend, Stephan Eismann, Andrew M Watkins, Ramya Rangan, Masha Karelina, Rhiju Das, and Ron O Dror. Geometric deep learning of RNA structure. *Science*, 373(6558):1047–1051, 2021. 1

[35] Maurice Weiler, Fred A Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant CNNs. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 849–858, 2018. 1

[36] Thomas Wiatowski and Helmut Bölcskei. A mathematical theory of deep convolutional neural networks for feature extraction. *IEEE Trans. Inf. Theory*, 64(3):1845–1866, 2017. 1

[37] Daniel Worrall and Max Welling. Deep scale-spaces: Equivariance over scale. *Proc. Adv. Neural Inf. Process. Syst.*, 32, 2019. 1

[38] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 5028–5037, 2017. 1

[39] Qi Xie, Qian Zhao, Zongben Xu, and Deyu Meng. Fourier series expansion based filter parametrization for equivariant convolutions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45 (4):4537–4551, 2022. 1

[40] Pew-Thian Yap, Xudong Jiang, and Alex Chichung Kot. Two-dimensional polar harmonic transforms for invariant image representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(7):1259–1270, 2009. 5

[41] Sunkyu Yu. Evolving scattering networks for engineering disorder. *Nature Comput. Sci.*, 3(2):128–138, 2023. 1

[42] Richard Zhang. Making convolutional networks shift-invariant again. In *Proc. Int. Conf. Mach. Learn.*, pages 7324–7334, 2019. 1, 3