Structured Invariance for Vision Backbones: A Geometric Route to MetaFormer

Shuren Qi, Yushu Zhang, Yuming Fang, Xiaochun Cao, and Fenglei Fan

Abstract—Invariance has long been a foundational prior in vision, but its role in vision backbones under scaling laws has become increasingly ambiguous. In practice, it is often learned implicitly through data, augmentation, and scale, while making it explicit is commonly viewed as restricting representational flexibility. This paper studies the role of invariance in modern vision backbones through a geometric route to MetaFormer, centered on two questions: where invariance should reside, and whether explicit invariance can yield gains under scaling. To answer these questions, we extend the theory of moments and moment invariants to modern vision backbones by formulating a unified learnable framework for global, local, and hierarchical invariance, in which the role of learnable operators becomes analytically explicit. Within this framework, we establish a classification theorem showing that, under suitable assumptions, every admissible continuous local linear operator with scalar-like channels reduces to shared pointwise 1×1 convolution. This theorem reveals a principled role for 1×1 convolution in geometry-compatible learning. Because the admissible operator is spatially blind, geometric structure must be maintained through the interaction of two roles throughout the hierarchy: geometry-aware spatial operators and generic pointwise channel operators. This compositional pattern provides a geometric interpretation of the MetaFormer decomposition into spatial and channel mixing. Guided by the structured view, we instantiate GeoFormer, where geometric priors reside primarily in spatial mixing while channel mixing remains generic and pointwise. Experiments show that this structured design improves the invariance-discriminability trade-off and exhibits favorable scaling behavior.

Index Terms—Representation, invariance, equivariance, scaling, MetaFormer.

1 Introduction

Geometric variability is one of the most persistent sources of complexity in vision. The same semantic content may appear with translations, rotations, reflections, or mild viewpoint changes, while the desired prediction (e.g., class identity) often remains unchanged. *Invariance*—the property that a representation is stable under a family of transformations—has therefore served as a central inductive bias throughout the development of computer vision [1]. Yet in today's large-scale vision backbones, including Convolutional Neural Networks (CNNs) and vision Transformers, the status of invariance has become ambiguous. The strongest models often acquire invariance implicitly through large datasets, aggressive augmentation, and over-parameterization [2]. Conversely, making invariance explicit by hard-wiring geometric constraints has been repeatedly criticized as being overly restrictive, potentially harming discriminability and scalability [3].

1.1 Invariance as Inductive Bias

From a statistical learning perspective, invariance can be viewed as compressing *equivalence classes* induced by transformations [4]. In vision, semantic categories are naturally stable under geometric transformations, so invariance reduces dependence on exhaustive data coverage and mitigates the *curse of dimensionality*.

- *S. Qi and F. Fan are with the Frontier of Artificial Networks (FAN) Group, Department of Data Science, City University of Hong Kong, Hong Kong, China (e-mail: shurenqi, fenglfan@cityu.edu.hk).*
- *Y. Zhang and Y. Fang are with the School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang, China (e-mail: zhangyushu@jxufe.edu.cn, fa0001ng@e.ntu.edu.sg).*
- *X. Cao is with the School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen, China (e-mail: caoxiaochun@mail.sysu.edu.cn).*

However, stronger invariance is not always better. Over-invariance can erase fine-grained spatial information that is crucial for discrimination. The goal, therefore, is to develop a *structured* form of invariance that is (i) *selective* with respect to specified transformation families, (ii) *controllable* for varying task requirements, and (iii) *compatible* with scalable end-to-end representation learning.

1.2 State of the Art and Motivation

Existing approaches to invariance in vision can be broadly organized into three routes, each making different trade-offs among geometric robustness, discriminability, and scalability.

Hand-crafted invariants, such as SIFT, HOG, and moment invariants, provide clear geometric guarantees and interpretability [1], [5], [6]. However, they rely on fixed, manual designs that do not adapt to data distributions. This route operates largely outside the hierarchical feature formation of modern backbones, behaving more like front-end feature engineering than an integrated component of end-to-end representation learning.

Implicit learning lets invariance emerge through data and scale. Over-parameterized CNN/Transformer backbones trained with large datasets and strong augmentations can acquire a degree of invariance [2], [7]. This route scales exceptionally well in engineering terms, and it usually achieves strong discriminative performance. However, the learned invariance is hard to treat as a reusable structural asset. In a scaling context, it can also be data- and compute-hungry, since the model may repeatedly relearn symmetries.

Equivariant networks aim for a principled middle ground: instead of demanding invariance everywhere, they enforce

equivariance so that features transform predictably under the group action [3], [8]. This makes invariance derivable and auditable. Yet equivariance often comes with a specific structural commitment: to preserve group structure, channels commonly carry structured group representations, which in turn constrains channel mixing to be representation-compatible. Geometry is no longer only a concern in the spatial domain; it is *moved into* the channel dimension and directly restricts cross-channel interaction. In a regime where generic backbones increasingly seek stronger channel mixing, wider/deeper models, and high-throughput scaling, this geometry–channel binding can introduce friction—both in expressive flexibility and in engineering scalability. A relevant practical bottleneck is that many constructions rely on discrete group sampling, which trades geometric fidelity against compute, memory, and optimization complexity.

Motivation. These observations suggest two questions that are increasingly pressing in the era of scaling laws: 1) Where should invariance reside in vision backbones—what form should it take, and which architectural components should carry it? 2) When scaling, does explicit invariance yield predictable and reusable gains? These questions matter because they determine whether geometric structure is repeatedly relearned from data or encoded once as a reusable architectural prior, and thus whether invariance remains a scaling-compatible source of efficiency that helps vision backbones balance robustness, discriminability, and scalability.

1.3 Main Ideas and Contributions

This paper develops a systematic answer to the above questions. Our central hypothesis is that a vision backbone should remain generic where semantic learning benefits most from data and scale, yet encode geometric priors in a *structured* way—in the right places and in the right forms—so that basic symmetries need not be repeatedly relearned. In this view, invariance is not opposed to scaling; it can serve as a reusable prior that reduces representational redundancy, preserves flexibility, and supports scaling-friendly design.

To make this hypothesis precise, we develop the argument in four steps:

1) We lift the global, local, and hierarchical invariants [9], [10] into a *unified feature-learning framework*, in which geometric structure is carried by the invariant representations themselves rather than prescribed through group-structured channel types [3], [8]. Here, we introduce learnable operators into the hierarchy, turning their compatibility with geometry into an *analytical* question.

2) Within this framework formulation, we ask what kind of learnable operator can remain compatible with sufficiently rich geometric constraints while staying as generic as possible. Under suitable assumptions, we establish a *classification theorem* showing that every admissible continuous local linear operator with scalar-like channels reduces to a constant pointwise channel operator, which on a discrete lattice is exactly shared 1×1 convolution.

3) We derive the structural consequence of this theorem. Since the admissible learnable operator is spatially blind, a geometry-engineered front-end followed by a purely learnable back-end cannot preserve spatial discrimination. This leads to a *compositional principle*: geometry-aware spatial operators and admissible pointwise channel operators must be interleaved throughout the hierarchy. In *MetaFormer* terms, this is precisely the decomposition into spatial mixing and channel mixing [11].

4) Guided by this principle, we instantiate our theory and analysis in GeoFormer, a *Geometric MetaFormer* that combines geometrically grounded spatial mixing with generic pointwise channel mixing, and test whether this form of structured invariance delivers practical gains in robustness and scaling.

Correspondingly, our contributions are fourfold:

- **Framework formulation.** We extend the foundational theory of moments and moment invariants to the setting of modern vision backbones, placing global, local, and hierarchical invariance [9], [10] within a unified learnable representation framework that renders the role of learnable operators analytically explicit. To the best of our knowledge, this is the first systematic bridge from classical invariant theory to scalable backbone design.
- **Operator characterization.** We establish a rigorous classification theorem, revealing a principled role for 1×1 convolution beyond its familiar engineering roles in channel adjustment and efficient computation. Within our theory, it is nearly unique in retaining generic learnability while remaining compatible with geometric structure, suggesting benefits for redundancy reduction and scaling-friendly design.
- **Compositional principle.** We show that the classification theorem yields a compositional principle, which provides an early geometric interpretation of the MetaFormer decomposition [11], a pattern that is pervasive in vision Transformers but has so far been justified mainly empirically.
- **Architectural instantiation.** We instantiate the theory in GeoFormer, where geometric priors reside primarily in spatial mixing, while channel mixing remains generic and pointwise. Through carefully controlled and aligned experiments, we show that this structured design yields the theory-predicted benefits: a better invariance–discriminability trade-off and more favorable scaling behavior of invariance priors.

2 Related Work

We situate our work relative to four closely related lines, focusing on how geometric priors are introduced, learned, or structurally encoded.

2.1 Hand-crafted Invariants

Classical computer vision treated invariance as an explicit design objective. Representative examples include moment invariants and local descriptors such as SIFT and HOG, which were designed to remain stable under translation, rotation, scaling, and related geometric transformations [5], [6], [12], [13]. Among them, moment invariants are

especially relevant here because they provide an explicit representation-level route to geometric invariance.

The main limitation of this line is that the resulting features are usually fixed by design and often global, limiting spatial selectivity, adaptability, and integration with modern end-to-end hierarchical learning. Subsequent work therefore extended classical moment invariants toward local and hierarchical invariant representations [9], [10]. Our work continues in this direction, but shifts the focus from descriptor construction to the architectural question of how such geometric compatibility can be preserved in scalable learnable backbones.

2.2 CNNs and Implicit Learning

CNNs replaced hand-crafted descriptors with learned hierarchical representations and established convolution as the dominant spatial inductive bias in vision. Relative to fully connected MLP-style models, convolution introduces locality, weight sharing, and translation equivariance, yielding substantially better statistical efficiency and practical performance on images [7], [14], [15], [16]. This success established CNNs as the standard example of how built-in structure benefits visual learning.

At the same time, the geometric bias of standard CNNs is concentrated mainly on translation, while robustness to broader transformations is usually acquired implicitly through data augmentation, large-scale supervision, and model capacity [17], [18]. CNNs thus occupy an important middle ground: they show the value of explicit spatial bias, but do not by themselves resolve how richer invariance can be incorporated in a scalable and controllable way.

2.3 Equivariant CNNs and Geometric Deep Learning

Equivariant networks were developed to combine the geometric explicitness of invariant methods with the hierarchical learning capacity of deep models. Rather than enforcing invariance directly, they preserve transformation structure across layers so that invariant representations can be derived in a principled way. Representative directions differ mainly in how equivariance is instantiated, including group convolutions, harmonic bases, orientation channels, steerable representations, and gauge-equivariant formulations [3], [8], [19], [20], [21], [22]. Complementary work extends the underlying symmetry from planar roto-translations to scale, spherical/3D rotations, Lie groups, and more general geometric domains [23], [24], [25], [26].

These models offer strong geometric faithfulness and can improve robustness or sample efficiency when the assumed symmetry matches the data. Their main relevance to our work lies in a contrasting structural choice: geometry is typically encoded through representation-carrying channels, so channel mixing must remain compatible with the associated group action. This geometry–channel coupling is mathematically principled, but it can restrict generic cross-channel interaction and increase implementation complexity through specialized kernels, group discretization, and less hardware-friendly computation. Our work instead seeks to keep channel mixing generic and place geometric prior primarily in spatial interaction.

2.4 Transformers and Inductive Bias Reconsidered

Recent vision backbones reopened the question of how much explicit inductive bias is necessary. After CNNs had established the effectiveness of convolutional spatial bias, ViT [2], DeiT [27], and MLP-Mixer [28] showed that architectures with much weaker built-in spatial priors could still become highly competitive at sufficient scale. A common lesson is that weaker architectural bias tends to place greater demands on data scale and training recipe, including augmentation and optimization [29], [30], [31].

Subsequent work moved toward more balanced designs by reintroducing lightweight spatial priors into otherwise generic backbones. Swin Transformer [32] restores locality and hierarchy through window-based attention, S2-MLP [33] introduces local spatial interaction through shift-based mixing, and ConvNeXt [34] reincorporates convolutional priors such as locality, hierarchy, and weight sharing. MetaFormer then abstracts this trend into a general design: a vision backbone can be viewed as the decomposition of a spatial token mixer and a channel mixer, with the token mixer carrying much of the effective inductive bias [11]. Our work is closely related to this view but differs in emphasis. MetaFormer is primarily an architectural abstraction distilled from empirical design convergence, whereas we seek to explain why such a decomposition is natural when one wants explicit invariance without sacrificing scalable generic channel learning. In the sequel, we use this terminology at the level of architectural roles: geometry-aware spatial operators correspond to token mixers, whereas learnable pointwise channel operators correspond to channel mixers.

3 FOUNDATIONS

This section reviews the progression from global to local and hierarchical invariants in classical moment theory [1], [9], [10].

3.1 Notations

We use a minimal set of symbols for scalar fields, multi-channel feature maps, geometric groups, and proof-level test spaces.

- $\Omega \subset \mathbb{R}^2$: Spatial domain.
- $f : \Omega \to \mathbb{R}$: Generic scalar field.
- $F : \Omega \to \mathbb{R}^N$: Generic N-channel feature map.
- Y: Final output space.
- $D_N(\Omega) := C_c^\infty(\Omega, \mathbb{R}^N)$: Smooth N-channel test-function space with compact support in Ω; in particular, $D(\Omega) := D_1(\Omega)$.
- $\mathfrak{G}_\bullet$: Generic geometric group in the framework, with instances $\mathfrak{G}_{\mathrm{geo}}$: group of translations, rotations, and reflections; $\mathfrak{G}_{\mathrm{scale}}$: scaling group; $\mathfrak{G}_{\mathrm{inv}} \leq \mathfrak{G}_{\mathrm{geo}} \ltimes \mathfrak{G}_{\mathrm{scale}}$: target invariance group, where $\ltimes$ denotes the semidirect product; and $\mathfrak{G}_{\mathrm{probe}}$: probe group used in the operator classification theorem.
- $\mathfrak{g}_\bullet \in \mathfrak{G}_\bullet$: Generic group element. Group actions are by pullback whenever defined: $[\mathfrak{g}_\bullet \cdot f](x) = f(\mathfrak{g}_\bullet^{-1} x)$ and $[\mathfrak{g}_\bullet \cdot F](x) = F(\mathfrak{g}_\bullet^{-1} x)$.
- $\mathcal{I}, \mathcal{M}, \mathcal{T}, \mathcal{R}$: Global invariant readout, local geometric operator, learnable feature-map operator, and generic representation map.

- $\equiv$: Equality up to boundary effects, discretization, or other benign implementation details.

In Sec. 3, moments and invariant operators are defined on generic scalar fields f. From Sec. 4 onward, we additionally use F for generic multi-channel feature maps whose channels are scalar-like fields.

For a representation $\mathcal{R}$ and a geometric group $\mathfrak{G}$ with generic element $\mathfrak{g} \in \mathfrak{G}$, geometric stability is characterized as follows.

- **Invariance**: $\mathcal{R}(\mathfrak{g} \cdot f) \equiv \mathcal{R}(f)$.
- **Equivariance**: if $\mathcal{R}(f)$ is again a spatial field on Ω, then $\mathcal{R}(\mathfrak{g} \cdot f) \equiv \mathfrak{g} \cdot \mathcal{R}(f)$ whenever the same action is defined on the codomain.
- **Covariance**: $\mathcal{R}(\mathfrak{g} \cdot f) \equiv \tau_{\mathfrak{g}}(\mathcal{R}(f))$, where $\tau_{\mathfrak{g}}$ denotes the transformation induced by $\mathfrak{g}$ on the codomain of $\mathcal{R}$.

3.2 Global Invariants

Classical moment theory starts from global moments. Let V_{nm} be an orthogonal basis on Ω. The global moments of a scalar field f are

$$\mu_{nm}(f) = \langle f, V_{nm} \rangle = \iint_{\Omega} V_{nm}^{*}(\xi) f(\xi)\, d\xi, \tag{1}$$

where $\xi \in \Omega$ is the spatial variable. For polar moments, $V_{nm}(r, \theta) = R_n(r) \exp(\mathrm{j} m\theta)$, with $\mathrm{j} = \sqrt{-1}$ and radial functions satisfying $\int_0^1 R_n(r) R_{n'}^{*}(r)\, r\, dr = \frac{1}{2\pi} \delta_{nn'}$ [12], [13], [35].

For a rotation $\mathfrak{g}_{\mathrm{geo},\alpha}$ by angle α, the polar moments satisfy $\mu_{nm}(\mathfrak{g}_{\mathrm{geo},\alpha} \cdot f) = \exp(-\mathrm{j} m\alpha) \mu_{nm}(f)$, hence $|\mu_{nm}(\mathfrak{g}_{\mathrm{geo},\alpha} \cdot f)| = |\mu_{nm}(f)|$. Therefore, $|\mu_{nm}(f)|$ is a rotation invariant. More generally, global invariants are obtained from the moment set $\{\mu_{nm}(f)\}_{n,m}$ by classical elimination of the group action. We write this abstractly as $\mathcal{I}(f) = \Psi(\{\mu_{nm}(f)\}_{n,m})$, where Ψ denotes an invariant elimination map; magnitudes are the simplest example for rotations.

3.3 Local Invariants

Local invariants extend global invariants to spatially varying fields [9]. Let $c \in \Omega$ denote the local center and $w > 0$ be the local scale. Define the local support $B(c, w) = \{\xi \in \Omega : \|\xi - c\| \le w\}$. For $\lambda = (n, m, w)$, the corresponding local moment at center c is

$$\nu_{\lambda}[f](c) = \langle f, V_{\lambda}^{c} \rangle = \frac{1}{w^2} \iint_{B(c,w)} \big(V_{\lambda}^{c}(\xi)\big)^{*} f(\xi)\, d\xi, \tag{2}$$

where $V_{\lambda}^{c}(\xi) = R_n(r_c(\xi)) \exp(\mathrm{j} m \theta_c(\xi))$, with $r_c(\xi) = \sqrt{(\xi_1 - c_1)^2 + (\xi_2 - c_2)^2}/w$ and $\theta_c(\xi) = \mathrm{Arg}((\xi_1 - c_1) + \mathrm{j}(\xi_2 - c_2))$ for $\xi \neq c$; the value at $\xi = c$ is immaterial for the integral and may be assigned arbitrarily.

This notation separates the operator index $\lambda = (n, m, w)$ from the output coordinate c. The local center is therefore not treated as a layer parameter but the spatial coordinate of the resulting feature field.

Under geometric transformations, the local polar moments satisfy (see Sec. A.2 for proofs):

- **Translation**: for $\mathfrak{g}_{\mathrm{geo},\Delta}$, $\nu_{\lambda}[\mathfrak{g}_{\mathrm{geo},\Delta} \cdot f](c) = \nu_{\lambda}[f](c - \Delta)$.
- **Rotation / Flipping**: for $\mathfrak{g}_{\mathrm{geo},\rho}$, $\nu_{\lambda}[\mathfrak{g}_{\mathrm{geo},\rho} \cdot f](c) = \chi_m(\mathfrak{g}_{\mathrm{geo},\rho})\, \nu_{\lambda}[f]\big((\mathfrak{g}_{\mathrm{geo},\rho})^{-1} c\big)$, where $\chi_m(\mathfrak{g}_{\mathrm{geo},\rho})$ is the induced unimodular phase/sign factor.
- **Scaling**: for $\mathfrak{g}_{\mathrm{scale},s}$, $\nu_{\lambda}[\mathfrak{g}_{\mathrm{scale},s} \cdot f](c) = \nu_{\lambda_s}[f](s^{-1} c)$ with $\lambda_s = (n, m, w/s)$.

Here $\mathfrak{g}_{\mathrm{geo},\Delta}$ denotes translation by $\Delta \in \mathbb{R}^2$, $\mathfrak{g}_{\mathrm{geo},\rho}$ denotes a planar orthogonal transformation $\rho \in O(2)$ (rotation or reflection), and $\mathfrak{g}_{\mathrm{scale},s}$ denotes isotropic scaling by factor $s > 0$.

A local geometric operator is then obtained by applying the same elimination idea locally:

$$\mathcal{M}^{\lambda}(f)(c) = |\nu_{\lambda}[f](c)| = |\langle f, V_{\lambda}^{c} \rangle|. \tag{3}$$

Its geometric behavior is (see Sec. A.2 for proofs):

- $\mathfrak{G}_{\mathrm{geo}}$**-equivariance**: $\mathcal{M}^{\lambda}(\mathfrak{g}_{\mathrm{geo}} \cdot f) \equiv \mathfrak{g}_{\mathrm{geo}} \cdot \mathcal{M}^{\lambda}(f)$.
- $\mathfrak{G}_{\mathrm{scale}}$**-covariance**: $\mathcal{M}^{\lambda}(\mathfrak{g}_{\mathrm{scale},s} \cdot f) \equiv \mathfrak{g}_{\mathrm{scale},s} \cdot \mathcal{M}^{\lambda_s}(f)$.

3.4 Hierarchical Invariants

Hierarchical invariants are obtained by cascading local geometric operators and applying a global invariant readout at the end [10], [36], [37]. Let $p = (\lambda_{[1]}, \lambda_{[2]}, \ldots, \lambda_{[L]})$ be a path of local indices, where each $\lambda_{[l]} = (n_{[l]}, m_{[l]}, w_{[l]})$. Starting from $f_{[0]} = f$, define recursively $f_{[l]} = \mathcal{M}^{\lambda_{[l]}}(f_{[l-1]})$ for $l = 1, \ldots, L$. The hierarchical representation associated with the path p is

$$\mathcal{R}_p(f) = \mathcal{I}(f_{[L]}) = \mathcal{I} \circ \mathcal{M}^{\lambda_{[L]}} \circ \cdots \circ \mathcal{M}^{\lambda_{[1]}}(f). \tag{4}$$

Here $\mathcal{I}$ denotes the global invariant readout on the final scalar field: one first computes the global moments $\{\mu_{nm}(f_{[L]})\}_{n,m}$ and then derives invariants through a classical elimination procedure, exactly as in the global-invariant construction above. Rotation invariants such as $|\mu_{nm}(f_{[L]})|$ are particular cases within this general construction.

The resulting hierarchical representation satisfies (see Sec. A.3 for proofs):

- $\mathfrak{G}_{\mathrm{geo}}$**-equivariance of the cascade**: $\mathcal{M}^{\lambda_{[L]}} \circ \cdots \circ \mathcal{M}^{\lambda_{[1]}}(\mathfrak{g}_{\mathrm{geo}} \cdot f) \equiv \mathfrak{g}_{\mathrm{geo}} \cdot \big(\mathcal{M}^{\lambda_{[L]}} \circ \cdots \circ \mathcal{M}^{\lambda_{[1]}}(f)\big)$.
- $\mathfrak{G}_{\mathrm{scale}}$**-covariance of the cascade**: $\mathcal{M}^{\lambda_{[L]}} \circ \cdots \circ \mathcal{M}^{\lambda_{[1]}}(\mathfrak{g}_{\mathrm{scale},s} \cdot f) \equiv \mathfrak{g}_{\mathrm{scale},s} \cdot \big(\mathcal{M}^{\lambda_{[L],s}} \circ \cdots \circ \mathcal{M}^{\lambda_{[1],s}}(f)\big)$, where $\lambda_{[l],s} = (n_{[l]}, m_{[l]}, w_{[l]}/s)$.
- $\mathfrak{G}_{\mathrm{inv}}$**-invariance of the readout**: $\mathcal{R}_p(\mathfrak{g}_{\mathrm{inv}} \cdot f) \equiv \mathcal{R}_p(f)$.

This progression, together with its continuation toward learnable operators and architectural instantiation, is summarized in Fig. 1.

4 A Geometric Route to MetaFormer

This section develops a geometric route from invariants to MetaFormer; Fig. 1 summarizes the route. We proceed in three steps. First, we lift the invariant constructions of Sec. 3 from scalar fields to multi-channel feature maps. Second, once learnable operators are introduced, we ask what form they may take if they remain generic while still respecting geometric prior. Third, we show that, under a sufficiently rich equivariance requirement, local linear learnable mixing collapses to shared 1×1 convolution. Its geometric compatibility is preserved, but its lack of spatial interaction makes a separate geometry-aware token mixer necessary. MetaFormer emerges from this structural tension.

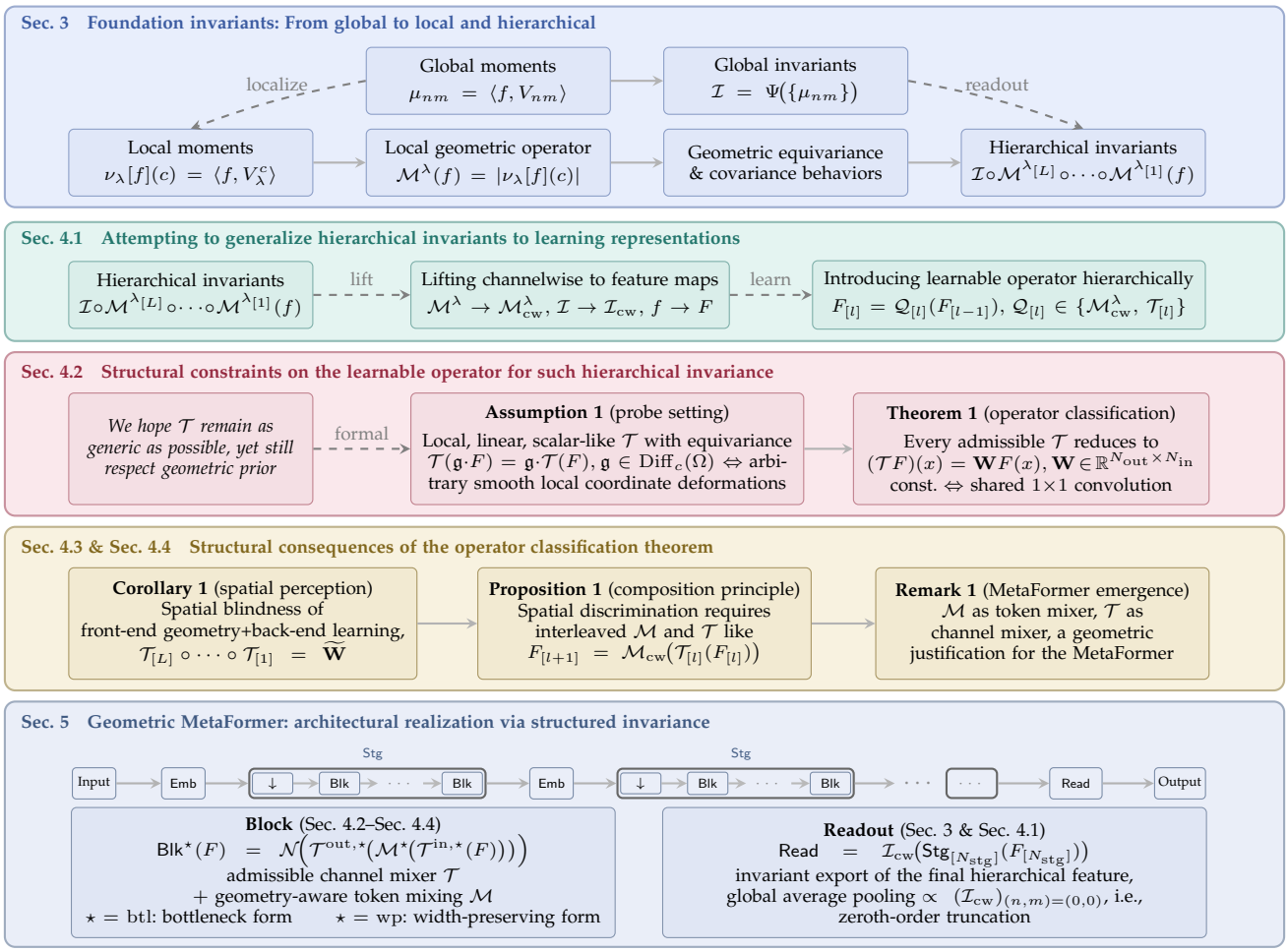


Fig. 1. **A Geometric Route From Invariants to MetaFormer.** (*Sec. 3*) Global moments μ_{nm} induce the invariant readout $\mathcal{I}$, local moments ν_λ induce the local geometric operator $\mathcal{M}^\lambda$, and cascading these local operators yields hierarchical invariants. (*Sec. 4.1*) This route is then lifted from scalar fields to multi-channel feature maps and extended by inserting learnable operators $\mathcal{T}$. (*Sec. 4.2*) Under the probe setting of Assumption 1, Theorem 1 shows that every admissible continuous local linear learnable operator reduces to shared 1×1 convolution. (*Sec. 4.3 & Sec. 4.4*) Corollary 1 shows that a purely learnable back-end is spatially blind, while Proposition 1 requires geometry-aware spatial operators and learnable pointwise channel operators to be interleaved throughout the hierarchy. Remark 1 interprets this paired structure in MetaFormer terms, with $\mathcal{M}$ as token mixer and $\mathcal{T}$ as channel mixer. (*Sec. 5*) The resulting Geometric MetaFormer is realized through the coordinated design of embedding, stage, block, and readout.

4.1 From Invariants to Learnable Representations

The preceding section develops a progression from global moment invariants to local and hierarchical invariant representations. These constructions provide explicit geometric stability, but once specified they are essentially fixed. To connect this route with modern architecture, we now pass from scalar-field constructions to general multi-channel feature maps.

We denote such feature maps by F, whose channels are scalar-like spatial fields. The scalar geometric operators introduced in Sec. 3 lift channelwise to multi-channel feature maps. For a scalar operator $\mathcal{M}^\lambda$, its channelwise lifting is

$$\big(\mathcal{M}^\lambda_{\mathrm{cw}}(F)\big)^{(k)} = \mathcal{M}^\lambda\big(F^{(k)}\big), \qquad k = 1, \ldots, N. \tag{5}$$

Likewise, the scalar invariant readout $\mathcal{I}$ lifts channelwise as

$$\big(\mathcal{I}_{\mathrm{cw}}(F)\big)^{(k)} = \mathcal{I}\big(F^{(k)}\big), \qquad k = 1, \ldots, N. \tag{6}$$

This preserves the geometric action at the scalar-channel level while opening a separate question: what forms of learnable operator remain geometrically compatible at the feature-map level?

Let $F_{[0]}$ denote the initial feature map derived from the input image f. Rather than restricting learning to a terminal stage, we allow learnable operators to appear throughout the hierarchy. Accordingly, we consider recursive feature updates of the form

$$F_{[l]} = \mathcal{Q}_{[l]}(F_{[l-1]}), \qquad l = 1, \ldots, L, \tag{7}$$

where each intermediate operator $\mathcal{Q}_{[l]}$ is either a local geometric operator or a learnable operator:

$$\mathcal{Q}_{[l]} \in \Big\{\mathcal{M}^{\lambda_{[l]}}_{\mathrm{cw}}, \mathcal{T}_{[l]}\Big\}. \tag{8}$$

The final representation is

$$\mathcal{R}(f) = \mathcal{I}_{\mathrm{cw}}(F_{[L]}). \tag{9}$$

The original hierarchical invariant representation is recovered as the scalar special case $N = 1$ in which no learnable operators are inserted, every $\mathcal{Q}_{[l]}$ reduces to $\mathcal{M}^{\lambda_{[l]}}$, and the final readout reduces to $\mathcal{I}$.

The central question is therefore to characterize the geometrically compatible form of the learnable feature-map operators $\mathcal{T}_{[l]}$.

4.2 The Equivariance Constraint: Why 1×1 Convolution is Necessary

Once learnable operators enter the invariant hierarchy, a tension appears immediately:

> *The learnable operator should remain as generic as possible, yet it should still respect geometric prior.*

This is the central structural tension highlighted in Fig. 1.

If the learnable component is left unconstrained, then geometric robustness is relegated to an implicit byproduct of data and training. If, conversely, one hard-wires a particular invariant construction into the learnable operator, then the operator ceases to be generic.

What is needed, therefore, is a constraint that enforces geometric compatibility without prescribing the operator in advance.

A natural candidate is equivariance. However, while useful for modeling task-level symmetries, equivariance with respect to only a task-specific small transformation group does not serve our present purpose: it ties the analysis to a particular geometric prior from the outset, and is often too weak to sharply restrict the form of a generic local linear operator. Our aim here is different. We do not seek to model a chosen task symmetry, but to determine how far geometric compatibility alone can constrain a generic local linear operator.

For that purpose, the transformation class must be sufficiently rich. In the local smooth setting, this naturally leads to compactly supported *diffeomorphisms*, which we invoke not as a literal symmetry model of data, but as a *structural probe* on geometrically compatible local learnable operator. In this way, equivariance becomes a genuinely selective test of compatibility with *arbitrary smooth local coordinate deformations*.

Assumption 1 (Probe setting)**.** *Let $\Omega \subset \mathbb{R}^2$ be a connected open set, and let*

$$\mathfrak{G}_{\mathrm{probe}} := \mathrm{Diff}_c(\Omega) \tag{10}$$

be the group of diffeomorphisms with compact support in Ω. Consider a linear operator

$$\mathcal{T} : D_{N_{\mathrm{in}}}(\Omega) \to D_{N_{\mathrm{out}}}(\Omega), \tag{11}$$

acting on multi-channel feature maps F. We assume:

(i) ***Scalar-like channels:*** *geometry acts by pullback on the spatial variable and trivially on channel indices; equivalently,*

$$[\mathfrak{g}_{\mathrm{probe}} \cdot F](x) = F(\mathfrak{g}_{\mathrm{probe}}^{-1}x), \forall \mathfrak{g}_{\mathrm{probe}} \in \mathfrak{G}_{\mathrm{probe}}; \tag{12}$$

(ii) ***Continuity and locality:*** *$\mathcal{T}$ is continuous, linear, and local;*

(iii) ***Diffeomorphism equivariance:*** *for every $\mathfrak{g}_{\mathrm{probe}} \in \mathfrak{G}_{\mathrm{probe}}$,*

$$\mathcal{T}(\mathfrak{g}_{\mathrm{probe}} \cdot F) = \mathfrak{g}_{\mathrm{probe}} \cdot \mathcal{T}(F), \forall F \in D_{N_{\mathrm{in}}}(\Omega). \tag{13}$$

Assumption 1 isolates a local linear operator under a strong notion of geometric compatibility. The domain and test-function spaces provide a standard continuous setting for locality and smooth deformations; the group $\mathfrak{G}_{\mathrm{probe}} = \mathrm{Diff}_c(\Omega)$ serves as a structural probe of arbitrary smooth local coordinate deformations; and the scalar-channel hypothesis ensures that geometric structure is imposed on the operator rather than encoded a priori in the channel space.

By a standard linear Peetre-type reduction [38], [39], [40], Assumption 1(ii) implies that $\mathcal{T}$ is locally a finite-order differential operator (see Sec. B.1 for proofs). The problem is thus reduced to classifying which local differential terms can survive full diffeomorphism equivariance.

Theorem 1 (Classification of local linear diffeomorphism-equivariant operators)**.** *Under Assumption 1, there exists a constant matrix*

$$\mathbf{W} \in \mathbb{R}^{N_{\mathrm{out}} \times N_{\mathrm{in}}} \tag{14}$$

such that

$$(\mathcal{T}F)(x) = \mathbf{W}F(x), \qquad \forall F \in D_{N_{\mathrm{in}}}(\Omega),\ \forall x \in \Omega. \tag{15}$$

Equivalently, every continuous local linear operator with scalar-like channels that is equivariant under all compactly supported diffeomorphisms reduces to constant pointwise channel mixing.

Proof sketch. By continuity and locality, $\mathcal{T}$ is locally a finite-order differential operator. Fix $x_0 \in \Omega$. Using a compactly supported diffeomorphism that agrees with a pure dilation near x_0, equivariance (13) implies that $(\mathcal{T}F)(x_0)$ must be invariant under local rescaling of coordinates. Derivatives of order $m \geq 1$ scale nontrivially under such a dilation, whereas the zeroth-order term is scale-neutral. Hence all positive-order coefficients vanish, and $\mathcal{T}$ reduces locally to pointwise channel mixing: $(\mathcal{T}F)(x) = \mathbf{W}(x)F(x)$. A second equivariance argument, using compactly supported transport between arbitrary points in the connected domain Ω, forces $\mathbf{W}(x)$ to be independent of x, yielding (15). Upon discretization on a regular lattice, constant pointwise channel mixing is realized exactly as shared 1×1 convolution (see Sec. B for the full proof). □

Regarding Theorem 1, we should also review its scope, minimal probe conditions, and vision-facing interpretation.

- **Scope.** The use of the full group $\mathrm{Diff}_c(\Omega)$ is a convenient *probe choice*, rather than the minimal assumption required by the proof, and Theorem 1 should be interpreted accordingly. The result is a structural classification within the present setting—namely, local linear operators, scalar-like channels, and equivariance under a sufficiently rich local probe—rather than a universal classification of all geometry-compatible vision operators. In particular, if channels carry nontrivial geometric representations, if nonlocal operators are permitted, or if equivariance is imposed only with respect to a smaller transformation class, then additional admissible operators may arise (see also Sec. B.5).
- **Minimal probe conditions.** A closer inspection of the proof shows that the argument relies on only two ingredients: *localized isotropic rescaling*, which eliminates all positive-order terms, and *transport* across the connected domain, which forces the remaining zeroth-order coefficient to be spatially constant. In

the special case $\Omega = \mathbb{R}^d$, these may be read heuristically as isotropic scaling and translation, respectively (see also Sec. B.5).

- **Vision-facing interpretation.** We retain $\mathrm{Diff}_c(\Omega)$ in the theorem statement because it provides a standard, self-contained, and technically convenient maximal probe under which these two ingredients are immediate. Its role here is diagnostic rather than descriptive: we do not claim that natural images are literally symmetric under arbitrary compactly supported diffeomorphisms. From a vision-facing perspective, the same structural message may be understood through a smaller localized-similarity probe, i.e., a class of compactly supported transformations that, on a neighborhood of each point, coincide with a similarity map generated by translation, rotation/reflection, and isotropic dilation, provided that the two ingredients above remain available (see also Sec. B.5).

4.3 The Spatial Perception Dilemma: Why 1×1 Convolution Alone Is Insufficient

As indicated in Fig. 1, Theorem 1 should be understood in the cascade context of Sec. 4, where the key issue is the *composition pattern* between geometric and learnable operators. A particularly important case is the classical design in which a front-end first computes geometry-engineered features through several non-learnable operators, and a learnable back-end is applied only afterwards. We now ask whether such a back-end can still contribute fresh spatial discrimination.

To formulate the required capability, we use the following notion.

Definition 1 (Nontrivial spatial interaction)**.** *Let $\mathcal{S} : D_{N_{\mathrm{in}}}(\Omega) \to D_{N_{\mathrm{out}}}(\Omega)$ be an operator. We say that $\mathcal{S}$ realizes* nontrivial spatial interaction *if there exist $x \in \Omega$ and $H, K \in D_{N_{\mathrm{in}}}(\Omega)$ such that*

$$H(x) = K(x) \qquad \textit{but} \qquad (\mathcal{S}H)(x) \neq (\mathcal{S}K)(x). \tag{16}$$

Corollary 1 (Spatial blindness of a purely learnable back-end)**.** *Under the hypotheses of Theorem 1, any composition of admissible local linear learnable operators remains pointwise in space and therefore does not realize nontrivial spatial interaction.*

Proof. By Theorem 1, each admissible local linear learnable operator has the form $(\mathcal{T}F)(x) = \mathbf{W}F(x)$ for a constant matrix $\mathbf{W}$. Hence any composition of such operators is still pointwise, i.e.,

$$(\mathcal{T}_{[L]} \circ \cdots \circ \mathcal{T}_{[1]}F)(x) = \widetilde{\mathbf{W}}F(x) \tag{17}$$

for some constant matrix $\widetilde{\mathbf{W}}$. The conclusion then follows immediately from Definition 1. □

Corollary 1 identifies the limitation of the front-end/back-end split. Once geometry-aware spatial aggregation has been completed in the front-end, a purely learnable back-end can only remix the channels of the resulting features. It may improve channel-space separability, but it cannot introduce new cross-location interaction or build fresh spatially discriminative patterns. This is the spatial perception dilemma behind Theorem 1. In the pattern of *front-end geometry + back-end learning*, spatial perception is exhausted by the hand-crafted front-end, while the trainable back-end is forced to remain spatially blind [37]. Under our theory, this makes the architecture structurally inefficient.

4.4 The Emergence of the MetaFormer Design

Corollary 1 indicates that the front-end/back-end split is not the appropriate composition pattern for representation learning. Equivalently, in the MetaFormer terminology recalled above, token mixing cannot be confined to a one-shot geometric front-end and followed by a purely channel-mixing back-end. The resulting architectural principle may be summarized as follows.

Proposition 1 (Composition principle of geometric and learnable operators)**.** *Under the hypotheses of Theorem 1, if an invariant representation is to renew spatial discrimination throughout the hierarchy, then the architecture must interleave admissible learnable pointwise channel operators with geometry-aware spatial operators, i.e., it should be organized as a cascade of geometric–learnable paired units of the form*

$$F_{[l+1]} = \mathcal{M}_{\mathrm{cw}}^{\lambda_{[l]}}(\mathcal{T}_{[l]}(F_{[l]})). \tag{18}$$

In particular, a one-shot geometric front-end followed by a purely learnable back-end cannot achieve such renewal of spatial discrimination.

Justification. By Corollary 1, once geometry-aware spatial aggregation is completed in a fixed front-end, any subsequent back-end formed only by admissible local linear learnable operators remains pointwise in space and cannot realize any further nontrivial spatial interaction. The paired composition (18) resolves this limitation by repeatedly coupling the two roles throughout the cascade: $\mathcal{T}_{[l]}$ performs adaptive pointwise channel recombination on the current representation, thereby changing the semantic organization of the channels, and $\mathcal{M}_{\mathrm{cw}}^{\lambda_{[l]}}$ then performs a new round of local geometry-aware spatial aggregation on this renewed channel basis. Consequently, spatially discriminative patterns are not frozen once and for all in an initial front-end, but are progressively rebuilt and enriched layer by layer.

Moreover, this repeated coupling remains compatible with the invariant route established earlier. The geometric operators carry the explicit geometric prior and preserve the equivariant/covariant structure developed in Sec. 3, while the admissible learnable operators remain pointwise and therefore do not destroy that structure. Thus the cascade can increase discriminative power throughout the hierarchy without sacrificing the final invariant representation. □

Viewed through an architectural lens, Proposition 1 can be interpreted as a geometric route to the MetaFormer [11].

Remark 1 (Connection to the MetaFormer design)**.** *MetaFormer architectures are built around the separation between a token mixer and a channel mixer; in our framework, $\mathcal{M}_{\mathrm{cw}}^{\lambda}$ plays the role of the token mixer, while $\mathcal{T}$ serves as the channel mixer. In this sense, our framework re-derives a MetaFormer-type architecture from the requirements of invariance. At the same time, our construction clarifies the scope of this connection. In the present framework, the token mixer is geometrically grounded and endowed with explicit stability properties. In a*

general MetaFormer, however, the token mixer need not explicitly encode geometric invariance; more broadly, it may be regarded as a generic encoder of geometric or spatial patterns. Our result therefore does not reduce MetaFormer to geometry, but shows that geometric invariant representation theory naturally leads to it as one principled special case.

This interpretation is also consistent with the parameter allocation commonly observed in MetaFormer architectures. In many practical instances, the channel mixer carries the dominant share of the learnable parameters, whereas the token mixer remains comparatively light—often accounting for more than 90% of the block parameters in pooling- or depthwise-based variants, and still for roughly two-thirds in attention-based ones. From the present geometric viewpoint, such an allocation is natural. The channel mixer forms the principal learnable component and serves as a general equivariance-preserving carrier of semantic adaptation, thereby supporting both representational universality and favorable scaling behavior, since symmetry need not be repeatedly relearned by heavy spatial modules. The comparatively light token mixer then introduces a controlled geometric inductive bias through modest spatial aggregation.

This perspective provides one possible explanation for the empirical success of the MetaFormer design: it is less constrained than conventional convolutional or strictly group-equivariant architectures, while remaining more geometrically structured than a plain ViT with weaker inductive bias.

5 The Geometric MetaFormer with Structured Invariance

We now turn from the operator-level analysis to its architectural realization. The operator characterization developed in Sec. 4 naturally leads to a separation between geometry-aware spatial mixing and generic channel mixing, echoing the MetaFormer decomposition. Building on this connection, we realize structured invariance within the MetaFormer framework by using geometric operators as spatial mixers and pointwise channel mixing as the generic learnable pathway, yielding the GeoFormer backbone[1].

5.1 Overview

The backbone GeoFormer is organized into four components: *block*, *stage*, *embedding*, and *readout*. This terminology follows the design perspective introduced in Sec. 4.4: a block realizes the repeated coupling of geometry-aware token mixing and learnable channel mixing, stages organize such blocks hierarchically, embeddings prepare stage inputs, and the readout exports the final invariant representation.

The principle of GeoFormer is *structured invariance*. Its practical meaning is that the main design trade-offs are handled *structurally*, rather than being forced into a single uniformly constrained operator. Concretely:

- **Block.** A block is the basic theory-driven learning unit that resolves the tension between generic learning and geometric compatibility. As shown in Sec. 4, admissible learnable local linear mixing reduces to pointwise channel mixing under the equivariance constraint, so geometry-aware token mixing must supply the nontrivial spatial interaction. Its role is therefore to couple generic channel adaptation with explicit geometric spatial aggregation, thereby renewing spatial discrimination throughout the hierarchy.
- **Stage.** A stage is a hierarchical module defined by a stage-initial transition operator followed by a sequence of blocks. Its purpose is to organize the backbone into a scalable multi-resolution hierarchy. The associated trade-off is among computational efficiency, effective receptive field, multi-scale abstraction, and geometric fidelity: stage transition improves efficiency and receptive-field growth, but may introduce approximation errors such as aliasing or loss of fine geometric detail.
- **Embedding.** An embedding module is a stage-entry projection/reparameterization. Its purpose is to prepare the incoming representation through channelization and low-level coding. The associated trade-off is between low-level image-structure discriminability and geometric fidelity: more flexible local coding is often beneficial for edges, textures, and early visual structure, but such flexibility is typically not itself geometry preserving.
- **Readout.** The readout is the terminal invariant export, defined as the invariant global operator acting on the final hierarchical feature. Its purpose is to convert the backbone output into a task-level invariant representation. The associated trade-off is between invariant retention and export simplicity: a richer invariant readout preserves more global geometric structure, whereas simpler truncations yield a cheaper and more compact task interface.

Let $F_{[i]}$ denote the feature entering stage i, where $i = 1, \dots, N_{\mathrm{stg}}$. The backbone recursion is organized at the stage level as

$$\begin{aligned} F_{[1]} &= \mathsf{Emb}_{[1]}(f), \\ F_{[i+1]} &= \mathsf{Emb}_{[i+1]}\Big(\mathsf{Stg}_{[i]}(F_{[i]})\Big), \\ \mathcal{R} &= \mathsf{Read}\Big(\mathsf{Stg}_{[N_{\mathrm{stg}}]}(F_{[N_{\mathrm{stg}}]})\Big). \end{aligned} \tag{19}$$

where the second line is understood for $i = 1, \dots, N_{\mathrm{stg}} - 1$.

A feature first enters stage i, undergoes the stage-initial transition, and is then processed by the internal block sequence. The resulting representation is subsequently reparameterized by the next embedding module before entering stage $i + 1$. The internal blockwise realization is specified below.

5.2 Block

The block is the principal learning module of the GeoFormer. More specifically, it is the architectural unit most directly dictated by the theory of Sec. 4, i.e., Theorem 1, Corollary 1, and Proposition 1. The block is precisely the architectural realization of this paired unit.

Accordingly, the minimal theory-driven form of a block is

$$\widetilde{\mathsf{Blk}}_{[i,j]}(F) = \mathcal{M}_{[i,j]}\big(\mathcal{T}_{[i,j]}(F)\big), \tag{20}$$

1. Code available at https://github.com/ShurenQi/GeoFormer.

where $\mathcal{T}_{[i,j]}$ is an admissible pointwise learnable channel operator, and $\mathcal{M}_{[i,j]}$ is a geometry-aware token mixer built from the channelwise liftings $\mathcal{M}^{\lambda}_{\mathrm{cw}}$.

In practice, we use the enriched form

$$\begin{aligned}&\mathsf{Blk}^{\star}_{[i,j]}(F)\\&=\mathcal{N}_{[i,j]}\Big(\mathcal{T}^{\mathrm{out},\star}_{[i,j]}\big(\mathcal{M}^{\star}_{[i,j]}(\mathcal{T}^{\mathrm{in},\star}_{[i,j]}(F))\big)\Big),\star\in\{\mathrm{btl},\mathrm{wp}\},\end{aligned}\tag{21}$$

where $\mathcal{T}^{\mathrm{in},\star}_{[i,j]}$ and $\mathcal{T}^{\mathrm{out},\star}_{[i,j]}$ are admissible pointwise learnable maps, $\mathcal{M}^{\star}_{[i,j]}$ is the geometry-aware token mixer, and $\mathcal{N}_{[i,j]}$ absorbs training-oriented wrappers such as residual connections, normalization, squeeze-and-excitation, gating, or dropout, provided that they introduce no new learnable spatial mixing.

Let $\Lambda_{[i,j]}$ be the finite set of geometric branch indices used by block (i,j). Two practical realizations are especially natural.

A first realization is the *bottleneck form*. Its motivation is straightforward. If the geometric token mixer applies multiple geometric branches $\lambda\in\Lambda_{[i,j]}$ to every input channel, then the output width grows proportionally to the number of branches. Applying the full branch family directly at the ambient channel width is therefore often prohibitively expensive. A natural remedy is to first compress the channels by a geometrically compatible pointwise map, perform the geometry-aware mixing in a lower-dimensional bottleneck space, and then re-expand or fuse the result afterward. Formally,

$$\mathcal{M}^{\mathrm{btl}}_{[i,j]}(F)=\operatorname*{Concat}_{\lambda\in\Lambda_{[i,j]}}\mathcal{M}^{\lambda}_{\mathrm{cw}}(F).\tag{22}$$

Substituting this into (21) gives

$$\begin{aligned}&\mathsf{Blk}^{\mathrm{btl}}_{[i,j]}(F)\\&=\mathcal{N}_{[i,j]}\Big(\mathcal{T}^{\mathrm{out,btl}}_{[i,j]}\big(\operatorname*{Concat}_{\lambda\in\Lambda_{[i,j]}}\mathcal{M}^{\lambda}_{\mathrm{cw}}\big(\mathcal{T}^{\mathrm{in,btl}}_{[i,j]}(F)\big)\big)\Big).\end{aligned}\tag{23}$$

Its advantage is that every bottleneck channel is exposed to the full family of geometric branches. As a result, geometric prior is injected densely and uniformly, and the resulting structure is the closest architectural realization of the operator-level decomposition into geometry-aware spatial mixing plus pointwise channel mixing. Its limitation is equally clear: because compression is performed before the geometric operators, some information may be discarded prior to geometry-aware extraction.

A second realization is the *width-preserving form*. Its motivation is complementary to that of the bottleneck design. If the token mixer is required to preserve channel width, then it cannot, in general, apply the full geometric branch family to every channel without causing channel expansion. The usual solution is therefore to assign different geometric branches to different channel subsets, so that width is preserved while geometry-aware mixing remains tractable. Let $\Pi^{\lambda}_{[i,j]}$ denote the channel-selection operator associated with branch λ, with the partition property that, for every intermediate feature F,

$$F=\operatorname*{Concat}_{\lambda\in\Lambda_{[i,j]}}\Pi^{\lambda}_{[i,j]}(F).$$

We then define

$$\mathcal{M}^{\mathrm{wp}}_{[i,j]}(F)=\operatorname*{Concat}_{\lambda\in\Lambda_{[i,j]}}\mathcal{M}^{\lambda}_{\mathrm{cw}}\big(\Pi^{\lambda}_{[i,j]}(F)\big),\tag{24}$$

and therefore

$$\begin{aligned}&\mathsf{Blk}^{\mathrm{wp}}_{[i,j]}(F)\\&=\mathcal{N}_{[i,j]}\Big(\mathcal{T}^{\mathrm{out,wp}}_{[i,j]}\big(\operatorname*{Concat}_{\lambda\in\Lambda_{[i,j]}}\mathcal{M}^{\lambda}_{\mathrm{cw}}\big(\Pi^{\lambda}_{[i,j]}(\mathcal{T}^{\mathrm{in,wp}}_{[i,j]}(F))\big)\big)\Big).\end{aligned}\tag{25}$$

Its advantage is that no pre-compression is required: the original channel information enters the geometric module directly, and the resulting design is typically more favorable for throughput, memory usage, and large-scale model scaling. Its cost is that each geometric branch acts only on a subset of channels. Accordingly, the injection of geometric prior becomes sparser than in the bottleneck form and is therefore less exhaustive.

Note that, in both cases, the same theoretical principle is preserved: the block remains a theory-driven coupling of admissible learning and explicit geometry-aware spatial interaction.

5.3 Stage

While the block is the principal learning module, the stage is the carrier of hierarchical organization. Its role is not to introduce a new learning principle beyond the block, but to arrange computation across scales in a structurally controlled and computationally scalable backbone. Accordingly, a stage is realized by a stage-initial transition operator followed by a sequence of blocks.

For an input stage feature $F_{[i]}$, let

$$\begin{aligned}&F_{[i,0]}=\mathcal{D}_{[i]}\big(F_{[i]}\big),\\&F_{[i,j]}=\mathsf{Blk}_{[i,j]}\big(F_{[i,j-1]}\big),\qquad j=1,\dots,N^{\mathrm{blk}}_{[i]}.\end{aligned}\tag{26}$$

where $\mathcal{D}_{[i]}$ is the stage-transition operator.

The stage output is then defined by

$$\mathsf{Stg}_{[i]}\big(F_{[i]}\big)=F_{[i,N^{\mathrm{blk}}_{[i]}]}.\tag{27}$$

When the feature entering stage i has spatial resolution $H_i\times W_i$ and the stage-initial transition uses stride s_i, the lattice processed by the blocks in that stage has resolution

$$\widetilde{H}_i=H_i/s_i,\qquad\widetilde{W}_i=W_i/s_i.\tag{28}$$

The role of the stage is therefore a controlled trade-off among four factors: computational efficiency, effective receptive field, multi-scale abstraction, and geometric fidelity. The stage-initial downsampling establishes a coarser working scale, thereby improving efficiency, enlarging the effective receptive field, and enabling progressively higher-level semantic abstraction. At the same time, such hierarchical transition is not strictly geometry preserving: it may introduce aliasing, anisotropy, and the loss of fine local structure. In the present framework, this is an intentional and structured concession. Exact geometry-aware interaction is concentrated in the blocks, whereas the stage bears the approximation cost required to establish scalable hierarchy before blockwise processing resumes at the new scale.

This also clarifies the design space of $\mathcal{D}_{[i]}$. Standard implementations such as patch merging may be used, but more geometry-friendly choices are often preferable when possible, including *average pooling*, *isotropic Gaussian kernels*, or *anti-aliased downsampling kernels* [18]. Such choices do

not remove the approximation nature of stage transition, but they reduce unnecessary geometric distortion before the block sequence at that stage begins.

5.4 Embedding

The embedding module is a stage-entry projection/reparameterization. Its role is to prepare the representation entering a stage, primarily by adjusting channelization and low-level coding, rather than by serving as the main carrier of geometric interaction.

A practical default choice is a conventional convolutional embedding:

$$\mathsf{Emb}_{[i]}(F) = \mathrm{Conv}_{[i]}(F), \tag{29}$$

where $\mathrm{Conv}_{[i]}$ denotes a convolution with kernel size, stride, and output channel number indexed by i. For the input image, $\mathsf{Emb}_{[1]}$ maps raw pixels into the latent feature space of the first stage; for later stages, $\mathsf{Emb}_{[i]}$ reparameterizes the output of the previous stage before it enters stage i.

Its role is best understood as a controlled trade-off between low-level image-structure discriminability and geometric fidelity. On the one hand, early visual processing benefits from flexible local coding, edge and texture sensitivity, and sufficient discriminative capacity for low-level structures. On the other hand, such flexibility is typically realized by ordinary local filters, which do not in themselves guarantee the geometric behavior established for the theory-driven block operators. In the present framework, this trade-off is handled structurally rather than uniformly: the embedding module is allowed to prioritize low-level discriminative coding, while the subsequent blocks repeatedly re-impose explicit geometry-aware interaction throughout the backbone. This is also broadly consistent with *biological vision*, where early processing is associated more with local feature extraction than with fully formed invariant object representation, and stronger transformation tolerance emerges progressively at later stages [41].

This structural view is important in practice. Even when the embedding module is implemented by an ordinary convolution, the dominant share of parameters and learning capacity in the backbone still lies in the geometry-compatible blocks. Accordingly, compared with a typical CNN backbone—where a large number of unconstrained learnable spatial kernels must gradually acquire geometric robustness from data alone—the part that still needs to be compensated by data augmentation here is only a lightweight boundary layer with a relatively small parameter budget. For this reason, ordinary convolutional embedding combined with standard data augmentation can still be expected to reach geometric robustness more efficiently than in a typical CNN backbone, because most of the burden of invariance is no longer carried by unconstrained spatial kernels throughout the network.

From the viewpoint of invariant representation theory, more controlled choices such as *shared* 1×1 *projections*, *hand-crafted invariant descriptors*, or *group-equivariant convolutions* remain possible and are closer to the idealized operator-level picture. However, in realistic large-scale vision backbones, ordinary convolutional embedding is often a reasonable compromise because the main body of the network remains structured by repeated geometry-aware blocks.

5.5 Readout

The readout is the terminal interface through which task-level invariance is exported. At the architectural level, we define it directly by the invariant operator introduced in the theoretical development:

$$\mathcal{R} = \mathsf{Read}\Big(\mathsf{Stg}_{[N_{\mathrm{stg}}]}\big(F_{[N_{\mathrm{stg}}]}\big)\Big) = \mathcal{I}_{\mathrm{cw}}\Big(\mathsf{Stg}_{[N_{\mathrm{stg}}]}\big(F_{[N_{\mathrm{stg}}]}\big)\Big). \tag{30}$$

Under the decomposition of the invariant readout into frequency-indexed components, conventional Global Average Pooling (GAP) is recovered as the lowest-order special case, namely $(n,m)=(0,0)$:

$$\mathrm{GAP}(F) \propto (\mathcal{I}_{\mathrm{cw}})_{(n,m)=(0,0)}(F). \tag{31}$$

Under the corresponding normalization convention, this proportionality becomes equality. Therefore, GAP should be interpreted not as an external exception, but as the zeroth-order truncation of the invariant readout. The practical trade-off is then transparent: a richer invariant readout preserves more global geometric structure at the output interface, whereas GAP retains only the coarsest invariant statistic but is simpler and cheaper.

6 Experiments

The experiments are organized around two questions that follow directly from the preceding theory and architecture.

Q1. In a controlled classification setting, does structured invariance improve the trade-off between clean discriminability and geometric robustness?

Q2. Does the same advantage remain meaningful under model scaling and on large-scale robustness benchmarks?

To answer these questions, we proceed from controlled medium-scale evaluation to large-scale analysis. CIFAR-100 is used to study the clean/robust trade-off under rotation, while ImageNet-1K is used to evaluate scalability, robustness, and architectural positioning in the modern backbone landscape [42], [43].

6.1 CIFAR: The Invariance–Discriminability Trade-off

This subsection primarily addresses **Q1**.

6.1.1 Protocol and comparison groups

CIFAR-100 is used to evaluate whether explicit geometric structure improves the balance between clean discriminability and arbitrary-angle rotation robustness. Its scale is sufficient for nontrivial model comparison while remaining controlled enough to make the clean/robust trade-off visible under different augmentation regimes.

Experimental protocol. To make this trade-off explicit, we evaluate each model under four training protocols:

- N90° protocol: training uses rotations by multiples of 90° (i.e., $0^\circ, 90^\circ, 180^\circ, 270^\circ$);
- ±180° protocol: training uses arbitrary-angle rotations (i.e., $-180^\circ \sim 180^\circ$);
- 0° protocol: training uses no rotation augmentation;
- ±15° protocol: training uses only small-angle rotations (i.e., $-15^\circ \sim 15^\circ$).

TABLE 1
CIFAR-100 Accuracy (%) Under Four Training Protocols.

	N90°		±180°		0°		±15°	
Model	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust
Generic backbones								
VGGNet	71.4	32.3	72.2	53.4	70.8	24.3	71.3	31.3
ResNet	**75.5**	40.7	**76.7**	62.5	**75.5**	29.9	**75.4**	35.4
ViT	51.4	30.7	53.0	39.0	54.7	23.9	53.1	26.0
ConvNeXt	71.4	36.9	73.3	63.0	71.2	24.5	71.1	31.4
PoolFormer	67.8	39.2	68.5	54.9	69.6	29.9	70.0	33.9
Geometric baselines								
WaveNet	45.6	18.1	45.6	25.2	50.2	16.7	47.9	19.5
GroupNet	64.7	46.3	63.9	<u>63.3</u>	58.2	43.2	61.8	<u>51.4</u>
Invariant baselines								
InvPoint	36.1	34.1	37.5	35.9	39.8	35.4	38.4	34.9
InvConv3	67.5	41.5	68.0	51.7	67.6	27.5	67.2	29.5
InvPool	55.4	48.5	54.1	51.3	55.4	<u>48.1</u>	56.1	50.1
InvGauss	54.3	<u>51.1</u>	53.2	51.5	54.9	**50.3**	54.6	51.0
InvShift	65.5	41.5	67.6	52.6	66.5	26.1	66.8	28.1
Proposed model								
GeoFormer	<u>72.8</u>	**56.9**	<u>74.0</u>	**68.8**	<u>75.0</u>	46.6	<u>74.4</u>	**54.8**

TABLE 2
Statistical Indicators of CIFAR-100 Accuracy (%) Under Four Training Protocols.

Model	$\mathbb{E}$ Clean	$\mathbb{E}$ Robust	$\mathbb{E}$ C.−R.	$\mathbb{E}$ C.&R.
Generic backbones				
VGGNet	71.4	35.3	36.1	53.4
ResNet	**75.8**	42.1	33.6	<u>58.9</u>
ViT	53.0	29.9	23.1	41.5
ConvNeXt	71.8	38.9	32.8	55.3
PoolFormer	69.0	39.5	29.5	54.2
Geometric baselines				
WaveNet	47.3	19.9	27.5	33.6
GroupNet	62.2	<u>51.1</u>	11.1	56.6
Invariant baselines				
InvPoint	38.0	35.1	**2.9**	36.5
InvConv3	67.6	37.6	30.0	52.6
InvPool	55.3	49.5	5.7	52.4
InvGauss	54.2	51.0	<u>3.3</u>	52.6
InvShift	66.6	37.1	29.5	51.8
Proposed model				
GeoFormer	<u>74.1</u>	**56.8**	17.3	**65.4**

In all four cases, robustness is evaluated on the same arbitrary-angle rotated test set, and clean accuracy is reported on the standard test set. This design keeps the test-time condition fixed while varying the extent of geometric exposure during training, making it possible to distinguish structural robustness from augmentation-dependent robustness.

Comparison groups. The compared models fall into four groups.

- **Generic backbones:** VGGNet, ResNet, ViT, ConvNeXt, and PoolFormer [2], [7], [11]. These models are not built around explicit rotation-aware structure and therefore mainly test how far invariance can be acquired implicitly from data and augmentation.
- **Geometric baselines:** the wavelet scattering network WaveNet and the group equivariant network GroupNet [3], [37]. WaveNet represents the route of geometry-engineered front-end structure with limited back-end learning, whereas GroupNet represents the group-equivariant route in which geometry is encoded more deeply through constrained convolution, channels, and representation structure.
- **Invariant baselines:** a family of models that share the same front-end geometry and differ only in the back-end learning. Specifically, all these baselines first compute the hand-crafted hierarchical invariants established in Sec. 3.4, and then apply different learning operators and/or spatial mixers on top of invariants, as an engineering trial for Sec. 4.1 and a validation of the theoretical conclusions in Sec. 4.2 and Sec. 4.3. For brevity, we denote them by InvPoint, InvConv3, InvPool, InvGauss, and InvShift. Here, InvPoint uses only a pointwise convolution operator and corresponds to geometry-compatible but spatially blind pointwise learning; InvConv3 replaces it with a learnable local 3×3 convolution operator with unconstrained local interaction; InvPool, InvGauss, and InvShift retain the same pointwise operator as InvPoint while further introducing, respectively, pooling, Gaussian, and shift-based spatial mixers.
- **Proposed model:** GeoFormer is instantiated here as a bottleneck Geometric MetaFormer with 1×1 embedding, which better matches the experimental purpose. We also use geometry-friendly average-pooling stage transitions and a global geometric readout for the same reason (see Sec. C for the full architectural and training details).

6.1.2 Main findings

The CIFAR-100 experiments reveal the main structural conclusion of the paper. Generic backbones achieve strong clean recognition but depend heavily on augmentation to acquire rotation robustness; rigid geometric or invariant pipelines improve robustness but often sacrifice discriminability; and the pattern of *front-end geometry + back-end learning* remains structurally limited because spatially meaningful interaction is not renewed effectively across depth. Against this background, GeoFormer occupies the most favorable clean–robust regime by combining geometry-aware token mixing with flexible pointwise channel recombination. The remainder of this subsection unpacks this conclusion through the different comparison groups.

Table 1 gives accuracy on the standard test set and on the arbitrarily rotated test set under all four training protocols. Table 2 highlights the statistical differences. Fig. 2 visualizes the positions of each method on the invariance–discriminability trade-off landscape. In Tables 1 and 2, the best and second-best values are highlighted in bold and underlined, respectively.

6.1.3 Generic backbones: strong discrimination, weak structural robustness

A first clear pattern is that generic backbones do not resolve rotation robustness structurally, even when their clean accuracy is strong. ResNet and ConvNeXt are the clearest examples. On the clean test set, they remain among the strongest methods across all training protocols. However,

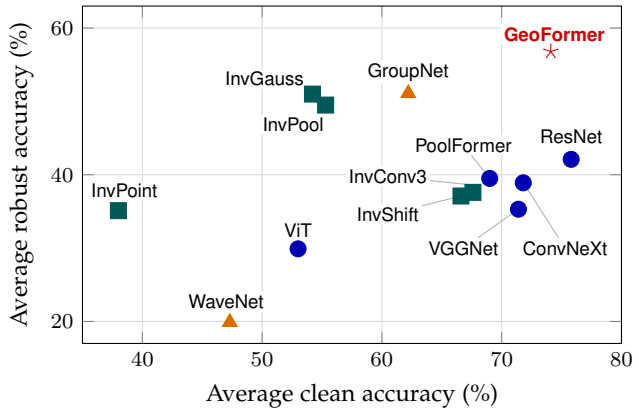


Fig. 2. Visualization of the invariance–discriminability trade-off on CIFAR-100, using average clean accuracy and average robust accuracy across the four training protocols. ● generic backbones, ▲ geometric baselines, ■ invariant-front-end baselines, and ✶ GeoFormer. GeoFormer lies near the favorable frontier, combining high clean recognition with substantially improved rotational robustness.

their rotated accuracy depends strongly on the augmentation used during training. Under the ±180° protocol, they reach 62.5 and 63.0 robust accuracy, whereas under the 0° protocol they drop to 29.9 and 24.5. This large sensitivity to train-time exposure is exactly what one expects when invariance is learned only implicitly from data.

ViT and PoolFormer fit the same broader picture from different starting points. ViT remains weak on this data scale, and PoolFormer is more stable than several CNN-style baselines but still falls well short of solving arbitrary-angle robustness. Taken together, the generic models show that rotation robustness can be helped by augmentation, but is not reliably internalized as a structural property.

6.1.4 *Invariant baselines: direct tests of the theory*

The invariant-front-end baselines provide the most direct empirical probe of the theory because they hold the feature preprocessing fixed. In all these models, the front-end is the hand-crafted hierarchical invariant construction introduced in Sec. 3.4; only the learnable operator and the form of spatial mixing after that front-end are varied. This makes the comparison particularly diagnostic for the structural claims of the paper.

InvPoint is the clearest realization of the theorem-level conclusion that geometry-compatible local linear learning collapses to pointwise channel mixing. Its behavior is precisely what the theory predicts: it preserves a full degree of geometric stability, but fails badly in clean discrimination because it cannot create fresh spatial interaction. This is the practical manifestation of spatial blindness.

InvConv3, which replaces the pointwise back-end with 3×3 convolution operator, changes the failure mode. Clean accuracy recovers substantially, which shows that unconstrained local spatial interaction is highly effective for discrimination. However, the robust performance becomes much more fragile, and the average clean/robust gap increases sharply to 30.0. This is exactly the trade-off one would expect when spatial expressivity is restored in a generic way, without maintaining geometric control.

InvPool, InvGauss, and InvShift refine this diagnosis further. Their results confirm that some form of spatial interaction is indeed necessary. At the same time, they also reveal the limitation of simple repair. InvPool and InvGauss substantially improve robust accuracy relative to InvPoint, and their small average gaps show that pooling or Gaussian mixers can preserve considerable stability. However, their clean averages remain far below those of the strongest discriminative models. In other words, such heuristic spatial interactions help, but they do not restore the full balance. InvShift reveals a complementary point. Because shift-style mixing is less geometry-controlled than pooling or isotropic Gaussian smoothing, it recovers more clean accuracy than InvPool or InvGauss, but its robust performance falls back toward the pattern of InvConv3.

Taken together, these baselines show a continuous structural spectrum: too little spatial mixing gives robustness without recognition, fully generic spatial mixing gives recognition without sufficient control, and lightweight repair occupies only an intermediate regime.

6.1.5 *Geometric baselines: explicit geometry helps, but not all explicit geometry is equally favorable*

The geometric baselines are also instructive, especially because they represent different ways of making geometry explicit.

WaveNet can be viewed as a strongly geometry-engineered route in which invariance is built largely through a front-end. Its results show good stability relative to its very limited flexibility, but its clean and robust accuracies remain low overall. This is consistent with the long-standing strength and limitation of hand-crafted or semi-hand-crafted invariant pipelines: they can encode stable geometric priors, but often do so at the cost of discriminative adaptability.

GroupNet represents a different route, namely explicit geometry through group-equivariant structure, where convolution, channels, and representation organization are constrained jointly by the underlying group representation. Its average robust accuracy of 51.1 is much stronger than that of most generic backbones, confirming that explicit geometry can indeed improve robustness. At the same time, its average clean accuracy of 62.2 remains clearly below GeoFormer. This comparison is especially informative because it separates two claims that are often conflated. The first is that geometry matters; GroupNet already supports this. Our finding is that *where* geometry is imposed matters just as much as whether it is imposed at all.

From this perspective, WaveNet and GroupNet serve as two complementary references. The former is closer to the route of geometry-dominant front-end design; the latter is closer to the route of geometry-constrained internal representation learning. Both confirm the value of explicit geometry, but neither achieves the same balance as the proposed factorization.

6.1.6 *GeoFormer: the favorable regime*

GeoFormer consistently occupies the favorable upper-right regime. It is not simply the most invariant model, nor merely another strong clean classifier. Its distinctive advantage is that it preserves much of the clean discriminability

TABLE 3
ImageNet-1K Parameter Counts, MACs, and Top-1 Accuracy.

Model	Params (M)	MACs (G)	Acc. (%)
Representative external models			
RSB-ResNet-152 [7], [44]	60	11.6	81.8
RegNetY-8G [44], [45]	39	8.0	82.1
ConvNeXt-S [34]	50	8.7	83.1
VAN-B4 [46]	60	12.2	84.2
RepLKNet-31B [47]	79	15.3	83.5
ConvFormer-M36 [48]	57	12.8	84.5
DeiT-B [27]	86	17.5	81.8
PVT-Large [49]	61	9.8	81.7
T2T-ViT-24 [50]	64	13.8	82.3
Swin-S [32]	50	8.7	83.0
CSWin-B [51]	78	15.0	84.2
MViTv2-B [52]	52	10.2	84.4
MLP-Mixer-B/16 [28]	59	12.7	76.4
Swin-Mixer-B/D24 [32]	61	10.4	81.3
gMLP-B [53]	73	15.8	81.6
CoAtNet-1 [54]	42	8.4	83.3
UniFormer-B [55]	50	8.3	83.9
iFormer-L [56]	87	14.0	84.8
MaxViT-S [57]	69	11.7	84.5
CAFormer-M36 [48]	56	13.2	85.2
Internal models			
IdentityFormer-M48	73	11.5	80.4
RandFormer-M48	73	11.9	81.4
PoolFormer-M48	73	11.6	82.5
GeoFormer-M48	73	12.0	82.7

of the best generic backbones while substantially improving robust performance.

This is most visible in the averages of Table 2. GeoFormer reaches 74.1 clean accuracy and 56.8 robust accuracy, yielding the strongest descriptive mean among all methods. Relative to ResNet and ConvNeXt, it gives much higher robust performance while remaining close in clean accuracy. Relative to the more rigid invariant baselines, it preserves much stronger clean recognition while still maintaining high stability.

The regime-wise behavior reinforces the same interpretation. Under the ±180° protocol, GeoFormer achieves the best robust accuracy at 68.8 while remaining competitive on the clean test set, indicating that explicit geometric structure can fully exploit favorable training exposure. Under the N90° protocol and the ±15° protocol, it again gives the strongest robust results, which shows that its robustness is not tied to exact augmentation matching. The 0° protocol is particularly informative: therein, some rigid invariant baselines, especially InvGauss, can exceed GeoFormer in robust accuracy, but only at the cost of a severe collapse in clean recognition. GeoFormer is therefore not maximizing invariance at any price; it strikes the best balance.

This behavior is precisely what one should expect from the proposed architecture. By placing geometry in token-side spatial interaction while keeping channel recombination generic, GeoFormer avoids the two extremes exposed by the controlled baselines: the spatial weakness of purely pointwise learning and the rigidity of fixed or overly constrained spatial structure.

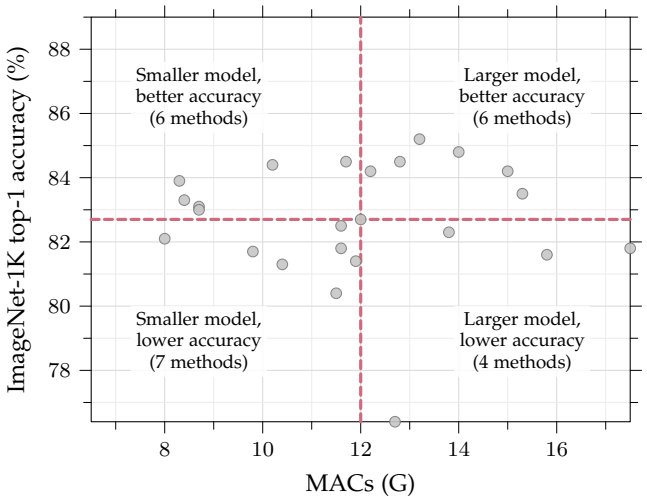


Fig. 3. Positioning of GeoFormer-M48 on ImageNet-1K using MACs and top-1 accuracy. GeoFormer-M48 is placed at the physical center of the plot, and the horizontal and vertical reference lines partition the plane into four interpretable quadrants.

6.2 ImageNet: Scaling Behavior of Structured Invariance

This subsection primarily addresses **Q2**.

6.2.1 Protocol and comparison groups

ImageNet-1K is used to test whether the geometric prior introduced by GeoFormer remains useful at large scale, where strong generic backbones already learn powerful representations from data. This setting therefore evaluates whether the proposed token-side geometric structure remains beneficial under increasing data scale, model scale, and realistic robustness benchmarks.

Experimental protocol. We evaluate each model on the standard ImageNet-1K validation set and on four widely used robustness benchmarks: ImageNet-A [58], ImageNet-R [58], ImageNet-S [59], and ImageNet-C [60]. These benchmarks are important here because they are *not* laboratory-style tests built from explicit simulated rotations. Instead, they represent realistic failure modes of large-scale visual recognition: naturally occurring hard examples, cross-rendering and depiction shift, sparse contour-dominant sketches, and common corruptions. Accordingly, they provide a more meaningful test of whether the proposed geometric prior improves real-world robustness and transfer beyond clean top-1 accuracy.

To study scaling directly, we evaluate five backbone sizes, namely S12, S24, S36, M36, and M48. This progression makes it possible to ask whether the effect of structured invariance diminishes, persists, or becomes more useful as the backbone grows.

Comparison groups. The compared models fall into three groups.

- **Internal baselines:** IdentityFormer, RandFormer, and PoolFormer. These models share the same overall MetaFormer-style scaffold as much as possible with GeoFormer and differ primarily in the token mixer. They therefore serve as the main diagnostic controls for isolating the structural role of token-side spatial

TABLE 4
Scores on Clean ImageNet-1K and Related Robustness Benchmarks under Five Model Scales.

Bench.: Metric	Model	S12	S24	S36	M36	M48
Clean: Acc. (%)	IdentityFormer	74.6	78.2	79.3	80.0	80.4
	RandFormer	76.6	78.8	79.5	81.2	81.4
	PoolFormer	77.2	80.3	81.4	82.1	82.5
	GeoFormer	**78.4**	**80.6**	**81.6**	**82.2**	**82.7**
A: Acc. (%)	IdentityFormer	5.4	9.7	13.4	17.4	19.1
	RandFormer	7.1	13.2	16.9	20.4	21.9
	PoolFormer	6.5	14.0	18.1	22.2	23.6
	GeoFormer	**9.0**	**15.5**	**20.6**	**26.0**	**27.9**
R: Acc. (%)	IdentityFormer	32.8	35.8	37.2	38.2	38.9
	RandFormer	34.6	37.8	39.9	40.6	40.9
	PoolFormer	38.2	41.8	42.4	43.9	44.1
	GeoFormer	**39.4**	**42.2**	**43.3**	**45.3**	**45.7**
S: Acc. (%)	IdentityFormer	21.0	23.6	24.8	25.6	26.4
	RandFormer	22.5	25.5	27.5	27.6	28.3
	PoolFormer	25.7	29.2	30.7	31.5	31.7
	GeoFormer	**27.0**	**29.3**	**30.8**	**31.6**	**32.3**
C: mCE↓	IdentityFormer	77.9	70.7	69.0	67.4	67.1
	RandFormer	73.1	66.5	64.6	62.6	61.6
	PoolFormer	69.1	**61.4**	**59.0**	**56.3**	**54.8**
	GeoFormer	**68.5**	61.7	60.2	57.5	56.0

mixing under near-matched architectural and computational conditions. In particular, IdentityFormer tests the degenerate case with no token-side spatial interaction, RandFormer tests whether arbitrary fixed spatial mixing is already sufficient, and PoolFormer serves as a generic internal baseline with a lightweight hand-designed mixer.

- **External positioning models:** a representative set of modern ImageNet backbones, including CNN-, Transformer-, MLP-, and MetaFormer-style architectures such as RSB-ResNet, ConvNeXt, Swin, MViTv2, MaxViT, and CAFormer. These models are not used for controlled internal diagnosis; instead, they provide broader architectural context for evaluating where GeoFormer-M48 stands in the contemporary ImageNet backbone landscape.
- **Proposed model:** In contrast to the CIFAR instantiation, we deliberately align the surrounding scaffold with the standard MetaFormer/PoolFormer template: width-preserving Geometric MetaFormer, patch-embedding-style input and inter-stage modules are used without geometry-friendly modifications, and the readout reduces to global average pooling, which corresponds to the zeroth-order truncation of the geometric readout. This alignment is intentional. It ensures that the internal comparison among IdentityFormer, RandFormer, PoolFormer, and GeoFormer is as clean as possible, with the token mixer as the principal architectural variable (see Sec. D for the full architectural and training details).

6.2.2 Main findings

The ImageNet study leads to four main conclusions. First, GeoFormer reaches a competitive large-scale baseline in the modern backbone landscape despite using a fixed geometric token mixer. Second, the effect of structured invariance remains visible under scaling when geometry is introduced as a lightweight token-side prior rather than as a pervasive constraint on channel learning. Third, under nearly matched compute, this prior improves large-scale robustness and generalization, especially on ImageNet-A, ImageNet-R, and ImageNet-S. Fourth, the ablations indicate that the benefit is structural rather than tied to one narrowly tuned mixer realization. The remainder of this subsection provides the evidence for each of these claims.

TABLE 5
Statistical Indicators of Scores on Clean ImageNet-1K and Related Robustness Benchmarks under Five Model Scales.

Model	$\mathbb{E}$ Clean	$\mathbb{E}$ A	$\mathbb{E}$ R	$\mathbb{E}$ S	$\mathbb{E}$ C ↓
IdentityFormer	78.5	13.0	36.6	24.3	70.4
RandFormer	79.5	15.9	38.8	26.3	65.7
PoolFormer	80.7	16.9	42.1	29.8	**60.1**
GeoFormer	**81.1**	**19.8**	**43.2**	**30.2**	60.8

Table 3 and Fig. 3 report the external positioning of GeoFormer-M48. Table 4 reports the full internal results across scales and benchmarks. Table 5 summarizes the scale-averaged indicators. Fig. 4 visualizes the clean–robustness scaling trajectories. Fig. 5 highlights the benchmark-wise advantage of GeoFormer over the strongest internal generic baseline PoolFormer. Table 6 reports representative mixer ablations at small and large scale.

6.2.3 External positioning: a competitive baseline in the modern backbone landscape

The first large-scale conclusion is that GeoFormer reaches a *competitive baseline position* among modern ImageNet backbones. Table 3 shows that GeoFormer-M48 attains 82.7 top-1 accuracy with 73M parameters and 12.0G MACs. Fig. 3 then places this model at the physical center of the MACs–accuracy plane and partitions the surrounding method space into four interpretable quadrants.

This visualization is informative for two reasons. First, GeoFormer-M48 is not an isolated or fragile outlier. It lies in a genuinely competitive region of the contemporary ImageNet design space, near a substantial set of strong models with comparable or larger computational budgets. In this sense, the proposed method should be understood not as a niche geometry-driven curiosity, but as a serious large-scale baseline.

Second, this position is especially significant because the token mixer of GeoFormer is still *fixed*. The spatial inductive bias is hand-designed and geometry-aware, but it is not itself a learned large-scale mixer. Consequently, the present result should be interpreted as a strong lower-bound-style reference point for future token-mixer design: if a learned spatial mixer is to justify its added flexibility at ImageNet scale and beyond, then its baseline target should be at least to surpass this level of performance. In other words, GeoFormer establishes that even a fixed geometry-aware mixer, when placed in the right architectural location, already reaches a regime that any stronger learned mixer must take seriously.

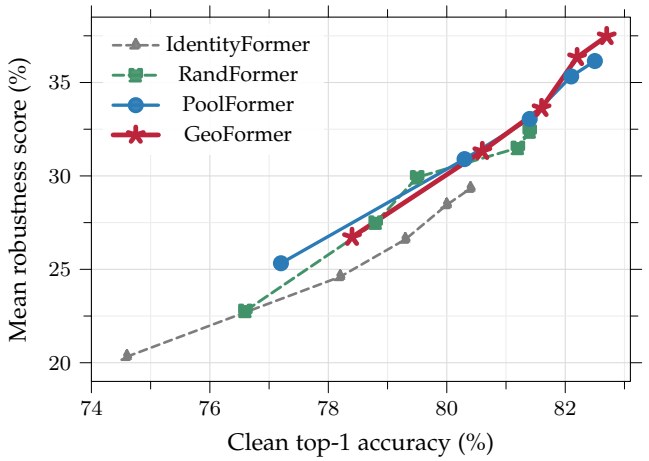


Fig. 4. Scaling trajectories on the clean–robustness frontier. Each polyline traces one model family from S12 to M48, using Clean accuracy on the x-axis and the mean of ImageNet-A, -R, -S, and -C ($100 - \mathrm{mCE}$) on the y-axis.

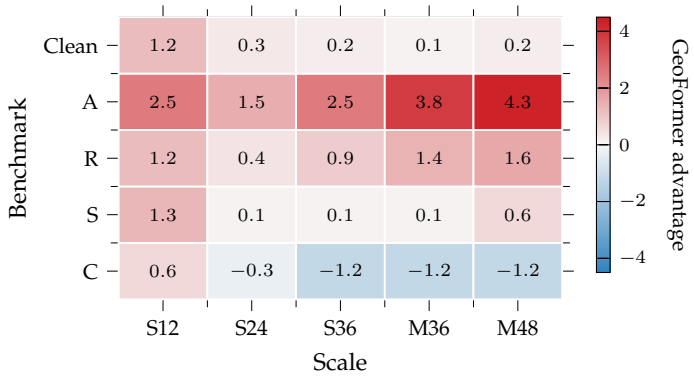


Fig. 5. Benchmark-wise improvement of GeoFormer over PoolFormer across five ImageNet scales.

To the best of our knowledge, this is among the first instances in our line of work where an invariant-based method reaches a competitive baseline position on ImageNet-scale classification. This matters because it changes the role of geometric prior in the discussion. Geometry is no longer merely a robustness-oriented add-on or a small-scale regularizer; it becomes a viable ingredient of modern scalable backbone design.

6.2.4 Scaling analysis: the geometric prior remains useful because it is placed in the right form and at the right location

The internal comparison reveals why this is possible. Table 4 shows that all four families improve with scale: clean accuracy rises from S12 to M48, ImageNet-A, ImageNet-R, and ImageNet-S accuracy also improve, and ImageNet-C mCE decreases. Thus larger models and larger data help across the board. However, scale does *not* erase the effect of token-mixer design.

The ordering among the four families remains highly structured. IdentityFormer is consistently weakest, showing that retaining only channel mixing and stage hierarchy is insufficient even at ImageNet scale. RandFormer improves over IdentityFormer, which confirms that some spatial interaction is better than none, but it still remains substantially below the stronger structured mixers. PoolFormer is the strongest generic internal baseline and therefore the most meaningful control. Yet GeoFormer remains ahead on clean accuracy at every scale, reaching 78.4, 80.6, 81.6, 82.2, and 82.7 from S12 to M48, compared with 77.2, 80.3, 81.4, 82.1, and 82.5 for PoolFormer.

This pattern is central to the paper. It shows that the success of GeoFormer neither come from adding geometry in an arbitrary way, nor from hard-wiring invariance everywhere. Instead, it comes from inserting geometry in the specific structural role predicted by Sec. 4: the channel mixer remains generic and scalable, while the token mixer carries a lightweight but explicit geometry-aware spatial prior. This is exactly the architectural factorization derived from Theorem 1, Corollary 1, and Proposition 1. The data then show that this factorization scales well in practice.

Fig. 4 makes the same point from a trajectory perspective. Each polyline traces one model family from S12 to M48 on the clean–robustness plane, where robustness is summarized by the mean of accuracy on ImageNet-A, ImageNet-R, ImageNet-S, and $100 - \mathrm{mCE}$ on ImageNet-C. The important observation is not merely that GeoFormer is strong at one endpoint, but that it stays on the favorable frontier throughout scaling. This is precisely the evidence needed for a scaling claim: the advantage is not a one-scale accident, but a persistent data-driven trend across model sizes.

Table 5 reinforces the same conclusion statistically. Averaged across the five scales, GeoFormer achieves the best mean clean accuracy and the best mean scores on ImageNet-A, ImageNet-R, and ImageNet-S. This is the empirical signature of a geometric prior that remains compatible with modern scaling, rather than being dominated by it.

6.2.5 Beyond clean accuracy: geometry improves real-world generalization and robustness under nearly matched compute

The most important large-scale diagnostic result lies beyond clean top-1 accuracy. Relative to the three highly aligned internal models, the introduction of geometric prior substantially improves performance on the benchmarks that are most closely related to geometric and structural variation in real-world visual data.

This comparison is especially strict because the internal models are nearly compute-matched. At M48, all four models have almost the same parameter count, while the MACs are 11.5 for IdentityFormer, 11.9 for RandFormer, 11.6 for PoolFormer, and 12.0 for GeoFormer. In other words, the advantage of GeoFormer is not purchased by a qualitatively different backbone scale or a major increase in computation.

Against this tightly aligned background, the gains are highly systematic. On ImageNet-A, GeoFormer outperforms PoolFormer at every scale, and the margin grows with scale from +2.5 points at S12 to +4.3 points at M48. On ImageNet-R, GeoFormer is again consistently better at every scale, with gains ranging from +0.4 to +1.6 points. On ImageNet-S, the advantage is smaller but still positive throughout. These improvements are summarized compactly in Fig. 5, which shows that the largest and most

TABLE 6
Ablations of the Geometric Token Mixer with Multi-scale and Higher-order Variants on Small and Large Models.

Variant	Clean	A	R	S	C ↓
Small scale: GeoFormer-S12					
Basic	78.4	9.0	39.4	27.0	68.5
Multi-scale variant	79.1	9.8	39.1	28.5	68.0
Higher-order variant	79.4	11.9	40.6	28.1	65.9
Large scale: GeoFormer-M48					
Basic	82.7	27.9	45.7	32.3	56.0
Multi-scale variant	82.9	29.6	44.5	32.2	56.3
Higher-order variant	82.8	28.0	44.9	33.1	56.7

systematic gains concentrate exactly on the benchmarks most related to structured visual variation.

This is a strong result for the intended claim of the paper. We are not testing explicit synthetic rotations or a narrow laboratory perturbation model here. Instead, the advantage appears on natural hard examples, depiction shift, sketch-like abstraction, and common corruptions. A plausible explanation is that the proposed token-side geometric prior does not merely encode one hand-crafted symmetry in isolation; rather, it biases spatial aggregation toward more stable geometric organization, thereby reducing sensitivity to incidental local layout and representation-level shape variation while leaving the channel pathway generic and scalable. In this way, explicit geometry does not compete with large-scale data and model capacity, but makes their learned semantics more reusable under visual shift.

ImageNet-C provides an instructive nuance. GeoFormer is clearly stronger than IdentityFormer and RandFormer throughout scaling, but PoolFormer remains slightly better in mCE at several medium-to-large scales. We regard this not as a contradiction, but as a useful boundary of the claim. The proposed geometric prior is especially effective for geometry-related and representation-related shifts, whereas corruption robustness depends more strongly on broader low-level stability properties. Even so, when Clean, A, R, S, and C are viewed jointly through Fig. 4 and Table 5, GeoFormer still occupies the most favorable overall regime.

6.2.6 Ablation: the benefit is structural rather than tied to a single design trick

Table 6 examines whether the ImageNet behavior of GeoFormer depends on one narrowly-tuned token-mixer design. Two representative variants are tested at both small and large scales: a multi-scale variant and a higher-order variant.

The main observation is that both variants remain strong and preserve the overall pattern of the base model. At small scale, both improve clean accuracy over the basic GeoFormer-S12, and the higher-order variant improves all three geometry-related accuracy benchmarks while also reducing ImageNet-C mCE. At large scale, the picture becomes crystal clear: the multi-scale variant gives the best ImageNet-A score, the higher-order variant gives the best ImageNet-S score, and the basic model remains strongest on ImageNet-R and ImageNet-C among the three variants.

This is exactly the kind of ablation outcome we would hope to see. It shows that the success of GeoFormer is not tied to one isolated engineering trick or one uniquely optimal branch configuration. Instead, the main advantage comes from the structural principle itself: explicit geometry is introduced in token-side spatial interaction, while the rest of the MetaFormer scaffold remains generic and scalable. The detailed mixer design still matters, but it acts as a secondary trade-off knob rather than as the sole source of success.

7 Conclusion

This paper has revisited invariance as a structural question for modern vision backbones. Rather than viewing invariance as either a fixed hand-crafted constraint or a property to be learned implicitly from scale alone, we have argued that its value in the scaling era depends crucially on how it is placed within the architecture. From this perspective, explicit geometry remains useful not as a replacement for generic learning, but as a complementary prior that reduces redundant relearning of symmetry while preserving scalability.

A central conclusion of this work is that sufficiently rich geometric compatibility imposes strong restrictions on admissible local linear learnable mixing. In the scalar-like setting studied here, such mixing reduces to pointwise channel interaction, which is geometrically compatible yet spatially blind. This, in turn, suggests a principled separation of roles: geometry should govern spatial interaction, while flexible learning should reside in channel mixing. Under this view, the MetaFormer decomposition can be understood not only as an effective design pattern, but also as a geometric resolution of this tension.

Our experiments support this interpretation. Structured invariance consistently improves the trade-off between clean discriminability and geometric robustness, and its benefits remain visible under model scaling and on larger-scale robustness benchmarks. These results suggest that geometric bias does not become obsolete in the presence of scale; rather, it remains effective when introduced in a form compatible with modern backbone design.

Several directions remain for future work, including extensions beyond local linear operators, stronger geometry-aware token mixers, and broader evaluation on dense prediction and more geometrically variable domains. More broadly, we hope this work helps position invariance not as a legacy notion of classical vision, but as a continuing critical perspective for scalable representation learning.

References

[1] S. Qi, Y. Zhang, C. Wang, J. Zhou, and X. Cao, "A survey of orthogonal moments for image representation: Theory, implementation, and evaluation," *ACM Computing Surveys*, vol. 55, no. 1, pp. 1–35, 2021.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[3] T. S. Cohen and M. Welling, "Group equivariant convolutional networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2016, pp. 2990–2999.

[4] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

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

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 886–893.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[8] T. S. Cohen, M. Weiler, B. Kicanaoglu, and M. Welling, "Gauge equivariant convolutional networks and the icosahedral CNN," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2019, pp. 1321–1330.

[9] S. Qi, Y. Zhang, C. Wang, J. Zhou, and X. Cao, "A principled design of image representation: Towards forensic tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5337–5354, 2022.

[10] S. Qi, Y. Zhang, C. Wang, Z. Xia, X. Cao, and F. Fan, "Transparent vision: A theory of hierarchical invariant representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[11] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "Metaformer is actually what you need for vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 819–10 829.

[12] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[13] J. Flusser, T. Suk, and B. Zitová, *Moments and Moment Invariants in Pattern Recognition*. Chichester, UK: Wiley, 2009.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2012, pp. 1097–1105.

[16] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[17] A. Azulay and Y. Weiss, "Why do deep convolutional networks generalize so poorly to small image transformations?" *Journal of Machine Learning Research*, vol. 20, no. 184, pp. 1–25, 2019.

[18] R. Zhang, "Making convolutional networks shift-invariant again," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2019, pp. 7324–7334.

[19] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow, "Harmonic networks: Deep translation and rotation equivariance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5028–5037.

[20] Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, "Oriented response networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 519–528.

[21] T. S. Cohen and M. Welling, "Steerable CNNs," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[22] M. Weiler and G. Cesa, "General $E(2)$-equivariant steerable CNNs," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[23] D. Marcos, M. Volpi, N. Komodakis, and D. Tuia, "Scale-equivariant steerable networks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[24] C. Esteves, C. Allen-Blanchette, X. Zhou, and K. Daniilidis, "Learning SO(3) equivariant representations with spherical CNNs," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 52–68.

[25] M. Finzi, S. Stanton, P. Izmailov, and A. G. Wilson, "A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 3318–3328.

[26] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," *arXiv preprint arXiv:2104.13478*, 2021.

[27] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers and distillation through attention," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 10 347–10 357.

[28] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "MLP-Mixer: An all-MLP architecture for vision," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[29] A. P. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your ViT? data, augmentation, and regularization in vision transformers," *Transactions on Machine Learning Research*, 2022.

[30] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12 106–12 116.

[31] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1–41, 2022.

[32] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 012–10 022.

[33] T. Yu, X. Li, Y. Cai, M. Sun, and P. Li, "S2-MLP: Spatial-shift MLP architecture for vision," *arXiv preprint arXiv:2106.07477*, 2022.

[34] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 976–11 986.

[35] M. R. Teague, "Image analysis via the general theory of moments," *Journal of the Optical Society of America*, vol. 70, no. 8, pp. 920–930, 1980.

[36] S. Mallat, "Group invariant scattering," *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, 2012.

[37] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.

[38] J. Peetre, "Une caracterisation abstraite des operateurs differentiels," *Mathematica Scandinavica*, vol. 7, pp. 211–218, 1960.

[39] J. Slovák, "Peetre theorem for nonlinear operators," *Annals of Global Analysis and Geometry*, vol. 11, no. 3, pp. 273–283, 1993.

[40] I. Kolář, P. W. Michor, and J. Slovák, *Natural Operations in Differential Geometry*. Berlin, Germany: Springer, 1993.

[41] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[42] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.

[43] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[44] R. Wightman, H. Touvron, and H. Jégou, "Resnet strikes back: An improved training procedure in timm," *arXiv preprint arXiv:2110.00476*, 2021.

[45] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 428–10 436.

[46] M.-H. Guo, C. Lu, Q. Hou, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, pp. 733–752.

[47] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in CNNs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 963–11 975.

[48] W. Yu, C. Si, P. Zhou, M. Luo, Y. Zhou, J. Feng, S. Yan, and X. Wang, "Metaformer baselines for vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 2, pp. 896–912, 2023.

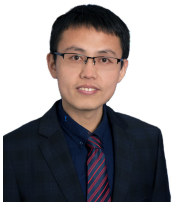
[49] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 568–578.

[50] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. H. Tay, J. Feng, and S. Yan, "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 558–567.

[51] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "CSWin transformer: A general vision transformer back-

bone with cross-shaped windows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12 124–12 134.
[52] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "MViTv2: Improved multiscale vision transformers for classification and detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4804–4814.
[53] H. Liu, Z. Dai, D. R. So, and Q. V. Le, "Pay attention to MLPs," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
[54] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying convolution and attention for all data sizes," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
[55] K. Li, Y. Wang, P. G. He, Y. Wang, L. Wang, Y. Qiao, and D. Lin, "Uniformer: Unifying convolution and self-attention for visual recognition," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.
[56] C. Si, W. Yu, P. Zhou, Y. Zhou, X. Wang, and S. Yan, "Inception transformer," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 495–23 509, 2022.
[57] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "MaxViT: Multi-axis vision transformer," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, pp. 459–479.
[58] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, and E. Dorundo, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 8340–8349.
[59] H. Wang, S. Ge, E. P. Xing, and Z. C. Lipton, "Learning robust global representations by penalizing local predictive power," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
[60] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Shuren Qi received the Ph.D. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2024. He is currently a Postdoctoral Fellow with the Department of Data Science, City University of Hong Kong, Hong Kong, China. He was previously a Postdoctoral Fellow with The Chinese University of Hong Kong, Hong Kong, China. His research interests include invariants, representations, and geometric deep learning.



Yushu Zhang (Senior Member, IEEE) received the Ph.D. degree from Chongqing University, Chongqing, China, in 2014. He is currently a Professor with the School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang, China. His research interests include privacy, security, and trustworthy AI. He serves on the Editorial Boards of *IEEE Transactions on Dependable and Secure Computing*, *Signal Processing*, and *Information Sciences*.

Yuming Fang (Fellow, IEEE) received the Ph.D. degree from Nanyang Technological University, Singapore, in 2013. He is currently a Professor and the Vice President of Jiangxi University of Finance and Economics, Nanchang, China. His research interests include attention modeling, quality assessment, computer vision, and 3D image/video processing.

Xiaochun Cao (Senior Member, IEEE) received the Ph.D. degree from the University of Central Florida, Orlando, FL, USA, in 2006. He is currently a Professor and the Dean of the School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University, Shenzhen, China. His research interests include computer vision, multimedia analysis, and artificial intelligence. He serves on the Editorial Boards of *IEEE Transactions on Pattern Analysis and Machine Intelligence* and *IEEE Transactions on Image Processing*.

Fenglei Fan received the Ph.D. degree from Rensselaer Polytechnic Institute, Troy, NY, USA, in 2021. He is currently an Assistant Professor with the Department of Data Science, City University of Hong Kong, Hong Kong, China, where he directs the Frontier of Artificial Networks Group. He was previously a postdoctoral researcher with Cornell University, NY, USA, and a Research Assistant Professor with The Chinese University of Hong Kong, Hong Kong, China. His research interests include NeuroAI, model compression, and medical imaging.

APPENDIX A
PROOFS FOR THE FOUNDATIONAL INVARIANTS

This appendix provides supporting proofs for the core claims in Sec. 3. Throughout, we work in the ideal continuous setting and assume that transformed local supports remain inside Ω, so that the equalities below hold exactly. The symbol $\equiv$ used in the main text then absorbs boundary truncation, discretization, and other benign implementation effects.

A.1 Auxiliary Setup

For $\lambda = (n, m, w)$ and $c \in \Omega$, recall the normalized local moment

$$\nu_\lambda[f](c) = \frac{1}{w^2} \iint_{B(c,w)} \big(V_\lambda^c(\xi)\big)^* f(\xi)\, d\xi, \tag{A.1}$$

where $B(c, w) = \{\xi \in \Omega : \|\xi - c\| \le w\}$. For local polar moments, the basis is

$$V_\lambda^c(\xi) = R_n\big(r_c(\xi)\big) \exp\big(\mathrm{j} m \theta_c(\xi)\big),$$
$$r_c(\xi) = \frac{\|\xi - c\|}{w},$$
$$\theta_c(\xi) = \mathrm{Arg}\big((\xi_1 - c_1) + \mathrm{j}(\xi_2 - c_2)\big), \quad \xi \neq c.$$

Here $\theta_c(\xi) \in (-\pi, \pi]$ denotes the principal polar angle. The value of $\theta_c(\xi)$ at $\xi = c$ is immaterial for the integral and may be assigned arbitrarily. The local geometric operator is

$$\mathcal{M}^\lambda(f)(c) = |\nu_\lambda[f](c)|.$$

A.2 Proof of Local Invariants

Lemma A.1 (Translation property)**.** *For a translation $\mathfrak{g}_{\mathrm{geo},\Delta}$ with offset $\Delta \in \mathbb{R}^2$,*

$$\nu_\lambda[\mathfrak{g}_{\mathrm{geo},\Delta} \cdot f](c) = \nu_\lambda[f](c - \Delta).$$

Proof. By the pullback action,

$$[\mathfrak{g}_{\mathrm{geo},\Delta} \cdot f](\xi) = f(\xi - \Delta).$$

Substituting into Eq. (A.1),

$$\nu_\lambda[\mathfrak{g}_{\mathrm{geo},\Delta} \cdot f](c) = \frac{1}{w^2} \iint_{B(c,w)} \big(V_\lambda^c(\xi)\big)^* f(\xi - \Delta)\, d\xi.$$

Let $\eta = \xi - \Delta$. Then $d\eta = d\xi$, and $B(c, w)$ is mapped to $B(c - \Delta, w)$. Since

$$(\eta + \Delta) - c = \eta - (c - \Delta),$$

the centered basis satisfies

$$V_\lambda^c(\eta + \Delta) = V_\lambda^{c-\Delta}(\eta).$$

Therefore,

$$\begin{aligned}\nu_\lambda[\mathfrak{g}_{\mathrm{geo},\Delta} \cdot f](c) &= \frac{1}{w^2} \iint_{B(c-\Delta,w)} \big(V_\lambda^{c-\Delta}(\eta)\big)^* f(\eta)\, d\eta \\ &= \nu_\lambda[f](c - \Delta).\end{aligned}$$

□

Lemma A.2 (Rotation and flipping property)**.** *Let $\mathfrak{g}_{\mathrm{geo},\rho}$ be a center-aligned rotation or reflection. Then*

$$\nu_\lambda[\mathfrak{g}_{\mathrm{geo},\rho} \cdot f](c) = \chi_m(\mathfrak{g}_{\mathrm{geo},\rho})\, \nu_\lambda[f]\big((\mathfrak{g}_{\mathrm{geo},\rho})^{-1} c\big),$$

where $\chi_m(\mathfrak{g}_{\mathrm{geo},\rho})$ is the induced unimodular phase/sign factor satisfying $|\chi_m(\mathfrak{g}_{\mathrm{geo},\rho})| = 1$.

Proof. Let Q_ρ be the orthogonal matrix associated with $\mathfrak{g}_{\mathrm{geo},\rho}$. By the pullback action,

$$[\mathfrak{g}_{\mathrm{geo},\rho} \cdot f](\xi) = f(Q_\rho^{-1}\xi).$$

Using the change of variables $\eta = Q_\rho^{-1}\xi$, we have $d\eta = d\xi$ since $|\det Q_\rho| = 1$, and the support $B(c, w)$ is mapped to $B(Q_\rho^{-1}c, w)$. The radial coordinate is preserved by orthogonality:

$$r_c(Q_\rho\eta) = \frac{\|Q_\rho\eta - c\|}{w} = \frac{\|\eta - Q_\rho^{-1}c\|}{w} = r_{Q_\rho^{-1}c}(\eta).$$

The angular coordinate changes by a predictable phase or sign, so the basis factor transforms as

$$V_\lambda^c(Q_\rho\eta) = \chi_m(\mathfrak{g}_{\mathrm{geo},\rho})^{-1} V_\lambda^{Q_\rho^{-1}c}(\eta).$$

Hence

$$\begin{aligned}\nu_\lambda[\mathfrak{g}_{\mathrm{geo},\rho} \cdot f](c) &= \frac{1}{w^2} \iint_{B(Q_\rho^{-1}c,w)} \big(V_\lambda^{Q_\rho^{-1}c}(\eta)\big)^* \chi_m(\mathfrak{g}_{\mathrm{geo},\rho})\, f(\eta)\, d\eta \\ &= \chi_m(\mathfrak{g}_{\mathrm{geo},\rho})\, \frac{1}{w^2} \iint_{B(Q_\rho^{-1}c,w)} \big(V_\lambda^{Q_\rho^{-1}c}(\eta)\big)^* f(\eta)\, d\eta \\ &= \chi_m(\mathfrak{g}_{\mathrm{geo},\rho})\, \nu_\lambda[f](Q_\rho^{-1}c).\end{aligned}$$

Since $Q_\rho^{-1}c = (\mathfrak{g}_{\mathrm{geo},\rho})^{-1}c$, the result follows. For a pure rotation by angle α, one has

$$\chi_m(\mathfrak{g}_{\mathrm{geo},\alpha}) = \exp(-\mathrm{j} m\alpha).$$

For reflections, one obtains an analogous unimodular factor. Since the local geometric operator uses the magnitude, only the condition $|\chi_m| = 1$ is needed in the sequel. □

Lemma A.3 (Scaling property)**.** *For a scaling transformation $\mathfrak{g}_{\mathrm{scale},s}$ with factor $s > 0$,*

$$\nu_\lambda[\mathfrak{g}_{\mathrm{scale},s} \cdot f](c) = \nu_{\lambda_s}[f](s^{-1}c), \qquad \lambda_s = (n, m, w/s).$$

Proof. By the pullback action,

$$[\mathfrak{g}_{\mathrm{scale},s} \cdot f](\xi) = f(s^{-1}\xi).$$

Substituting into Eq. (A.1),

$$\nu_\lambda[\mathfrak{g}_{\mathrm{scale},s} \cdot f](c) = \frac{1}{w^2} \iint_{B(c,w)} \big(V_\lambda^c(\xi)\big)^* f(s^{-1}\xi)\, d\xi.$$

Let $\eta = s^{-1}\xi$, equivalently $\xi = s\eta$. Then $d\xi = s^2\, d\eta$, and

$$\xi \in B(c, w) \quad \Longleftrightarrow \quad \eta \in B(s^{-1}c, w/s).$$

Moreover,

$$r_c(s\eta) = \frac{\|s\eta - c\|}{w} = \frac{\|\eta - s^{-1}c\|}{w/s},$$

and similarly

$$\theta_c(s\eta) = \theta_{s^{-1}c}(\eta).$$

Hence

$$V_\lambda^c(s\eta) = V_{\lambda_s}^{s^{-1}c}(\eta).$$

Therefore,

$$\begin{aligned}\nu_\lambda[\mathfrak{g}_{\mathrm{scale},s}\cdot f](c) &= \frac{1}{w^2}\iint_{B(s^{-1}c,w/s)} \big(V_{\lambda_s}^{s^{-1}c}(\eta)\big)^* f(\eta)\, s^2\, d\eta \\ &= \frac{1}{(w/s)^2}\iint_{B(s^{-1}c,w/s)} \big(V_{\lambda_s}^{s^{-1}c}(\eta)\big)^* f(\eta)\, d\eta \\ &= \nu_{\lambda_s}[f](s^{-1}c).\end{aligned}$$

□

Corollary A.1 ($\mathfrak{G}_{\mathrm{geo}}$-equivariance of $\mathcal{M}^\lambda$)**.** *For any* $\mathfrak{g}_{\mathrm{geo}} \in \mathfrak{G}_{\mathrm{geo}}$,

$$\mathcal{M}^\lambda(\mathfrak{g}_{\mathrm{geo}}\cdot f) = \mathfrak{g}_{\mathrm{geo}}\cdot\mathcal{M}^\lambda(f).$$

Proof. For translation, Lemma A.1 gives

$$\mathcal{M}^\lambda(\mathfrak{g}_{\mathrm{geo},\Delta}\cdot f)(c) = \big|\nu_\lambda[f](c-\Delta)\big| = [\mathfrak{g}_{\mathrm{geo},\Delta}\cdot\mathcal{M}^\lambda(f)](c).$$

For rotation or flipping, Lemma A.2 gives

$$\mathcal{M}^\lambda(\mathfrak{g}_{\mathrm{geo},\rho}\cdot f)(c) = \big|\chi_m(\mathfrak{g}_{\mathrm{geo},\rho})\nu_\lambda[f]((\mathfrak{g}_{\mathrm{geo},\rho})^{-1}c)\big|.$$

Since $|\chi_m(\mathfrak{g}_{\mathrm{geo},\rho})| = 1$,

$$\mathcal{M}^\lambda(\mathfrak{g}_{\mathrm{geo},\rho}\cdot f)(c) = \big|\nu_\lambda[f]((\mathfrak{g}_{\mathrm{geo},\rho})^{-1}c)\big| = [\mathfrak{g}_{\mathrm{geo},\rho}\cdot\mathcal{M}^\lambda(f)](c).$$

Since $\mathfrak{G}_{\mathrm{geo}}$ is generated by translations, rotations, and reflections, the conclusion follows. □

Corollary A.2 ($\mathfrak{G}_{\mathrm{scale}}$-covariance of $\mathcal{M}^\lambda$)**.** *For any scaling* $\mathfrak{g}_{\mathrm{scale},s} \in \mathfrak{G}_{\mathrm{scale}}$,

$$\mathcal{M}^\lambda(\mathfrak{g}_{\mathrm{scale},s}\cdot f) = \mathfrak{g}_{\mathrm{scale},s}\cdot\mathcal{M}^{\lambda_s}(f), \qquad \lambda_s = (n, m, w/s).$$

Proof. Using Lemma A.3,

$$\mathcal{M}^\lambda(\mathfrak{g}_{\mathrm{scale},s}\cdot f)(c) = \big|\nu_\lambda[\mathfrak{g}_{\mathrm{scale},s}\cdot f](c)\big| = \big|\nu_{\lambda_s}[f](s^{-1}c)\big|.$$

On the other hand,

$$[\mathfrak{g}_{\mathrm{scale},s}\cdot\mathcal{M}^{\lambda_s}(f)](c) = \mathcal{M}^{\lambda_s}(f)(s^{-1}c) = \big|\nu_{\lambda_s}[f](s^{-1}c)\big|.$$

Hence

$$\mathcal{M}^\lambda(\mathfrak{g}_{\mathrm{scale},s}\cdot f) = \mathfrak{g}_{\mathrm{scale},s}\cdot\mathcal{M}^{\lambda_s}(f).$$

□

A.3 Proof of Hierarchical Invariants

Lemma A.4 ($\mathfrak{G}_{\mathrm{geo}}$-equivariance of the cascade)**.** *Let* $f_{[0]} = f$ *and define recursively*

$$f_{[l]} = \mathcal{M}^{\lambda_{[l]}}(f_{[l-1]}), \qquad l = 1, \dots, L.$$

Then for any $\mathfrak{g}_{\mathrm{geo}} \in \mathfrak{G}_{\mathrm{geo}}$,

$$\mathcal{M}^{\lambda_{[L]}}\circ\cdots\circ\mathcal{M}^{\lambda_{[1]}}(\mathfrak{g}_{\mathrm{geo}}\cdot f) = \mathfrak{g}_{\mathrm{geo}}\cdot\big(\mathcal{M}^{\lambda_{[L]}}\circ\cdots\circ\mathcal{M}^{\lambda_{[1]}}(f)\big).$$

Proof. The proof is by induction on L.

For $L = 1$, the claim is exactly Corollary A.1.

Assume it holds for depth $L-1$. Then

$$\begin{aligned}&\mathcal{M}^{\lambda_{[L]}}\circ\cdots\circ\mathcal{M}^{\lambda_{[1]}}(\mathfrak{g}_{\mathrm{geo}}\cdot f) \\ &= \mathcal{M}^{\lambda_{[L]}}\Big(\mathcal{M}^{\lambda_{[L-1]}}\circ\cdots\circ\mathcal{M}^{\lambda_{[1]}}(\mathfrak{g}_{\mathrm{geo}}\cdot f)\Big) \\ &= \mathcal{M}^{\lambda_{[L]}}\Big(\mathfrak{g}_{\mathrm{geo}}\cdot(\mathcal{M}^{\lambda_{[L-1]}}\circ\cdots\circ\mathcal{M}^{\lambda_{[1]}}(f))\Big) \\ &= \mathfrak{g}_{\mathrm{geo}}\cdot\mathcal{M}^{\lambda_{[L]}}\Big(\mathcal{M}^{\lambda_{[L-1]}}\circ\cdots\circ\mathcal{M}^{\lambda_{[1]}}(f)\Big),\end{aligned}$$

where the second equality uses the induction hypothesis and the third uses Corollary A.1. This completes the induction. □

Lemma A.5 ($\mathfrak{G}_{\mathrm{scale}}$-covariance of the cascade)**.** *For any* $\mathfrak{g}_{\mathrm{scale},s} \in \mathfrak{G}_{\mathrm{scale}}$,

$$\begin{aligned}&\mathcal{M}^{\lambda_{[L]}}\circ\cdots\circ\mathcal{M}^{\lambda_{[1]}}(\mathfrak{g}_{\mathrm{scale},s}\cdot f) \\ &= \mathfrak{g}_{\mathrm{scale},s}\cdot\big(\mathcal{M}^{\lambda_{[L],s}}\circ\cdots\circ\mathcal{M}^{\lambda_{[1],s}}(f)\big),\end{aligned}$$

where

$$\lambda_{[l],s} = (n_{[l]}, m_{[l]}, w_{[l]}/s).$$

Proof. The proof is again by induction on L.

For $L = 1$, the statement is exactly Corollary A.2.

Assume it holds for depth $L-1$. Then

$$\begin{aligned}&\mathcal{M}^{\lambda_{[L]}}\circ\cdots\circ\mathcal{M}^{\lambda_{[1]}}(\mathfrak{g}_{\mathrm{scale},s}\cdot f) \\ &= \mathcal{M}^{\lambda_{[L]}}\Big(\mathcal{M}^{\lambda_{[L-1]}}\circ\cdots\circ\mathcal{M}^{\lambda_{[1]}}(\mathfrak{g}_{\mathrm{scale},s}\cdot f)\Big) \\ &= \mathcal{M}^{\lambda_{[L]}}\Big(\mathfrak{g}_{\mathrm{scale},s}\cdot(\mathcal{M}^{\lambda_{[L-1],s}}\circ\cdots\circ\mathcal{M}^{\lambda_{[1],s}}(f))\Big) \\ &= \mathfrak{g}_{\mathrm{scale},s}\cdot\mathcal{M}^{\lambda_{[L],s}}\Big(\mathcal{M}^{\lambda_{[L-1],s}}\circ\cdots\circ\mathcal{M}^{\lambda_{[1],s}}(f)\Big),\end{aligned}$$

where the second equality uses the induction hypothesis and the third uses Corollary A.2. This proves the claim. □

Proposition A.1 ($\mathfrak{G}_{\mathrm{inv}}$-invariance of the hierarchical representation)**.** *Let*

$$\mathcal{R}_p(f) = \mathcal{I}\circ\mathcal{M}^{\lambda_{[L]}}\circ\cdots\circ\mathcal{M}^{\lambda_{[1]}}(f).$$

Assume that the readout $\mathcal{I}$ *is defined on the final scalar field and satisfies*

$$\mathcal{I}(\tau_{\mathfrak{g}_{\mathrm{inv}}}h) = \mathcal{I}(h), \qquad \forall h \text{ in the domain of } \mathcal{I}, \quad \forall \mathfrak{g}_{\mathrm{inv}} \in \mathfrak{G}_{\mathrm{inv}}, \tag{A.2}$$

where $\tau_{\mathfrak{g}_{\mathrm{inv}}}$ *denotes the induced transformation on the final field entering* $\mathcal{I}$*. Then*

$$\mathcal{R}_p(\mathfrak{g}_{\mathrm{inv}}\cdot f) = \mathcal{R}_p(f), \qquad \forall \mathfrak{g}_{\mathrm{inv}} \in \mathfrak{G}_{\mathrm{inv}}.$$

Proof. Let

$$f_{[L]} = \mathcal{M}^{\lambda_{[L]}}\circ\cdots\circ\mathcal{M}^{\lambda_{[1]}}(f).$$

By Lemmas A.4 and A.5, the cascade transforms under $\mathfrak{g}_{\mathrm{inv}} \in \mathfrak{G}_{\mathrm{inv}}$ by a predictable induced action on the final field, denoted abstractly by $\tau_{\mathfrak{g}_{\mathrm{inv}}}$. Hence

$$\mathcal{M}^{\lambda_{[L]}}\circ\cdots\circ\mathcal{M}^{\lambda_{[1]}}(\mathfrak{g}_{\mathrm{inv}}\cdot f) = \tau_{\mathfrak{g}_{\mathrm{inv}}}f_{[L]}.$$

Applying the readout $\mathcal{I}$ and using Eq. (A.2), we obtain

$$\begin{aligned}\mathcal{R}_p(\mathfrak{g}_{\mathrm{inv}}\cdot f) &= \mathcal{I}\Big(\mathcal{M}^{\lambda_{[L]}}\circ\cdots\circ\mathcal{M}^{\lambda_{[1]}}(\mathfrak{g}_{\mathrm{inv}}\cdot f)\Big) \\ &= \mathcal{I}(\tau_{\mathfrak{g}_{\mathrm{inv}}}f_{[L]}) \\ &= \mathcal{I}(f_{[L]}) \\ &= \mathcal{R}_p(f).\end{aligned}$$

□

APPENDIX B
PROOF OF THE MAIN THEOREM

In this appendix, we give a complete proof of Theorem 1. Consistent with Sec. 4, we use F for generic multi-channel feature maps in the operator-level analysis. The proof proceeds in three steps: locality yields a finite-order differential representation; localized dilations eliminate all positive-order terms; the remaining pointwise coefficient is then shown to be constant on the connected domain.

B.1 Auxiliary Setup

Let

$$\mathfrak{G}_{\mathrm{probe}} := \mathrm{Diff}_c(\Omega)$$

be the group of diffeomorphisms with compact support in Ω. As in the main text, the group acts by pullback on scalar-like vector-valued feature maps:

$$[\mathfrak{g}_{\mathrm{probe}} \cdot F](x) = F(\mathfrak{g}_{\mathrm{probe}}^{-1} x), \qquad \forall \mathfrak{g}_{\mathrm{probe}} \in \mathfrak{G}_{\mathrm{probe}}.$$

Thus, channels are scalar-like in the sense that the action is entirely on spatial coordinates and does not mix channel indices.

Throughout this appendix, we study a continuous linear local operator

$$\mathcal{T} : D_{N_{\mathrm{in}}}(\Omega) \to D_{N_{\mathrm{out}}}(\Omega),$$

and assume $\mathfrak{G}_{\mathrm{probe}}$-equivariance:

$$\mathcal{T}(\mathfrak{g}_{\mathrm{probe}} \cdot F) = \mathfrak{g}_{\mathrm{probe}} \cdot \mathcal{T}(F), \quad \forall F \in D_{N_{\mathrm{in}}}(\Omega),\ \forall \mathfrak{g}_{\mathrm{probe}} \in \mathfrak{G}_{\mathrm{probe}}.$$

Definition B.1 (Pointwise operator)**.** *A linear operator*

$$\mathcal{T} : D_{N_{\mathrm{in}}}(\Omega) \to D_{N_{\mathrm{out}}}(\Omega)$$

is called pointwise *if there exists a smooth matrix field*

$$\mathbf{W} \in C^\infty(\Omega, \mathbb{R}^{N_{\mathrm{out}} \times N_{\mathrm{in}}})$$

such that

$$(\mathcal{T}F)(x) = \mathbf{W}(x)F(x), \qquad \forall F,\ \forall x \in \Omega.$$

If $\mathbf{W}(x)$ is independent of x, then $\mathcal{T}$ is called a constant pointwise mixing operator.

Definition B.2 (Localized dilation)**.** *Fix $x_0 \in \Omega$. A family*

$$\{\mathfrak{g}_{\mathrm{probe},a}\}_{a \in J} \subset \mathfrak{G}_{\mathrm{probe}}, \qquad J \subset (0, \infty),$$

is called a localized dilation about x_0 *if $1 \in J$, and there exists a neighborhood U of x_0 such that*

$$\mathfrak{g}_{\mathrm{probe},a}(x) = x_0 + a(x - x_0), \qquad \forall x \in U,\ \forall a \in J.$$

Lemma B.1 (Peetre-type reduction)**.** *Assume that*

$$\mathcal{T} : D_{N_{\mathrm{in}}}(\Omega) \to D_{N_{\mathrm{out}}}(\Omega)$$

is continuous, linear, and local. Then, for every point $x_0 \in \Omega$, there exist a neighborhood $U \ni x_0$, an integer $r \geq 0$, and matrix-valued coefficient functions

$$\mathbf{A}_\alpha : U \to \mathbb{R}^{N_{\mathrm{out}} \times N_{\mathrm{in}}}, \qquad |\alpha| \leq r,$$

whose entries belong to $C^\infty(U, \mathbb{R})$, such that

$$(\mathcal{T}F)(x) = \sum_{|\alpha| \leq r} \mathbf{A}_\alpha(x)\, \partial^\alpha F(x), \qquad \forall x \in U.$$

In particular, after possibly shrinking the neighborhood, we may regard $\mathcal{T}$ locally as a finite-order linear differential operator.

Proof. This is a standard consequence of the linear Peetre theorem [38], [39], [40]: every continuous linear local operator between spaces of smooth sections is locally a finite-order differential operator. Applying this componentwise to the input and output channels yields the stated matrix-valued representation, with each entry of $\mathbf{A}_\alpha$ lying in $C^\infty(U, \mathbb{R})$. □

For the remainder of this appendix, fix a point $x_0 \in \Omega$ and work on a neighborhood where Lemma B.1 applies. Thus,

$$(\mathcal{T}F)(x) = \sum_{|\alpha| \leq r} \mathbf{A}_\alpha(x)\, \partial^\alpha F(x)$$

holds locally for some finite r.

B.2 Localized Dilations and Jet Realization

Lemma B.2 (Existence of localized dilations)**.** *For every $x_0 \in \Omega$, there exist a neighborhood U of x_0 and an interval $J \subset (0, \infty)$ containing 1 such that a localized dilation family*

$$\{\mathfrak{g}_{\mathrm{probe},a}\}_{a \in J} \subset \mathfrak{G}_{\mathrm{probe}}$$

exists.

Proof. Choose $r_0 > 0$ such that $\overline{B(x_0, 2r_0)} \subset \Omega$. Let

$$\eta \in D(\Omega)$$

satisfy

$$0 \leq \eta \leq 1, \qquad \eta \equiv 1 \text{ on } B(x_0, r_0), \qquad \mathrm{supp}(\eta) \subset B(x_0, 2r_0).$$

Define the smooth compactly supported vector field

$$\zeta(x) = \eta(x)(x - x_0).$$

Let Φ_t denote the flow generated by ζ, i.e., the local one-parameter family of diffeomorphisms satisfying

$$\frac{d}{dt}\Phi_t(x) = \zeta(\Phi_t(x)), \qquad \Phi_0 = \mathrm{id}.$$

Since ζ is smooth and compactly supported, this flow is well defined, and $\Phi_t \in \mathrm{Diff}_c(\Omega)$ for all sufficiently small $|t|$. On $B(x_0, r_0)$, one has $\zeta(x) = x - x_0$, so the flow solves

$$\frac{d}{dt} z(t) = z(t) - x_0, \qquad z(0) = x,$$

whose solution is

$$z(t) = x_0 + e^t(x - x_0).$$

Hence, for all sufficiently small $|t|$,

$$\Phi_t(x) = x_0 + e^t(x - x_0), \qquad x \in B(x_0, r_0).$$

Now set

$$a = e^t$$

and define

$$\mathfrak{g}_{\mathrm{probe},a} := \Phi_{\log a}.$$

Then for a in some interval J containing 1, one has

$$\mathfrak{g}_{\mathrm{probe},a} \in \mathfrak{G}_{\mathrm{probe}}$$

and

$$\mathfrak{g}_{\mathrm{probe},a}(x) = x_0 + a(x - x_0) \qquad \text{for all } x \in U := B(x_0, r_0).$$

This is precisely a localized dilation about x_0. □

Lemma B.3 (Realization of a prescribed jet component)**.** *Fix an integer $m \geq 0$, a multi-index β with $|\beta| = m$, and a vector*

$$v \in \mathbb{R}^{N_{\mathrm{in}}}.$$

Then there exists

$$H \in D_{N_{\mathrm{in}}}(\Omega)$$

such that

$$\partial^\beta H(x_0) = v, \qquad \partial^\alpha H(x_0) = 0 \quad \textit{for all } |\alpha| \le m,\ \alpha \ne \beta.$$

Proof. Choose

$$\chi \in D(\Omega)$$

such that $\chi \equiv 1$ in a neighborhood of x_0. Define

$$H(x) = \chi(x)\,\frac{(x-x_0)^\beta}{\beta!}\,v.$$

Since $\chi \equiv 1$ near x_0, the derivatives of H at x_0 agree with those of the polynomial

$$x \mapsto \frac{(x-x_0)^\beta}{\beta!}\,v.$$

Therefore,

$$\partial^\beta H(x_0) = v, \qquad \partial^\alpha H(x_0) = 0 \quad \text{for all } |\alpha| \le m,\ \alpha \ne \beta.$$

□

B.3 Elimination of Positive-Order Terms

Lemma B.4 (Vanishing of the highest-order coefficient at a point)**.** *Fix $x_0 \in \Omega$. Suppose that*

$$(\mathcal{T}F)(x) = \sum_{|\alpha|\le r} \mathbf{A}_\alpha(x)\,\partial^\alpha F(x)$$

holds locally near x_0, and assume $\mathfrak{G}_{\mathrm{probe}}$-equivariance:

$$\mathcal{T}(\mathfrak{g}_{\mathrm{probe}} \cdot F) = \mathfrak{g}_{\mathrm{probe}} \cdot \mathcal{T}(F), \qquad \forall \mathfrak{g}_{\mathrm{probe}} \in \mathfrak{G}_{\mathrm{probe}}.$$

Let

$$m = \max\{\,|\alpha| : \mathbf{A}_\alpha(x_0) \ne 0\,\}.$$

If $m \ge 1$, then for every multi-index β with $|\beta| = m$,

$$\mathbf{A}_\beta(x_0) = 0.$$

Consequently, no nonzero positive-order highest coefficient can exist at x_0.

Proof. Assume $m \ge 1$. By Lemma B.2, choose a localized dilation family

$$\{\mathfrak{g}_{\mathrm{probe},a}\}_{a\in J}$$

about x_0. Since

$$\mathfrak{g}_{\mathrm{probe},a}(x_0) = x_0,$$

equivariance gives

$$\begin{aligned}\mathcal{T}(\mathfrak{g}_{\mathrm{probe},a} \cdot F)(x_0) &= (\mathfrak{g}_{\mathrm{probe},a} \cdot \mathcal{T}(F))(x_0)\\ &= \mathcal{T}(F)(x_0), \qquad \forall F,\ \forall a \in J.\end{aligned}$$

Using the local differential representation at x_0,

$$\sum_{|\alpha|\le r} \mathbf{A}_\alpha(x_0)\,\partial^\alpha(\mathfrak{g}_{\mathrm{probe},a} \cdot F)(x_0) = \sum_{|\alpha|\le r} \mathbf{A}_\alpha(x_0)\,\partial^\alpha F(x_0).$$

Near x_0, the inverse map is exactly affine:

$$\mathfrak{g}^{-1}_{\mathrm{probe},a}(x) = x_0 + a^{-1}(x - x_0).$$

Hence

$$\partial^\alpha(\mathfrak{g}_{\mathrm{probe},a} \cdot F)(x_0) = a^{-|\alpha|}\partial^\alpha F(x_0).$$

Therefore,

$$\sum_{|\alpha|\le r} \big(a^{-|\alpha|} - 1\big)\mathbf{A}_\alpha(x_0)\,\partial^\alpha F(x_0) = 0.$$

By definition of m, all terms with $|\alpha| > m$ vanish at x_0, so

$$\sum_{|\alpha|\le m} \big(a^{-|\alpha|} - 1\big)\mathbf{A}_\alpha(x_0)\,\partial^\alpha F(x_0) = 0. \tag{B.1}$$

Now fix any multi-index β with $|\beta| = m$, and any vector

$$v \in \mathbb{R}^{N_{\mathrm{in}}}.$$

By Lemma B.3, choose H so that

$$\partial^\beta H(x_0) = v, \qquad \partial^\alpha H(x_0) = 0 \quad \text{for all } |\alpha| \le m,\ \alpha \ne \beta.$$

Substituting this H into (B.1) yields

$$(a^{-m} - 1)\mathbf{A}_\beta(x_0)v = 0, \qquad \forall a \in J.$$

Since $m \ge 1$, the factor $a^{-m} - 1$ is not identically zero on J. Hence

$$\mathbf{A}_\beta(x_0)v = 0 \qquad \forall v \in \mathbb{R}^{N_{\mathrm{in}}},$$

so

$$\mathbf{A}_\beta(x_0) = 0.$$

As β was arbitrary among all multi-indices of the order m, every coefficient of order m vanishes at x_0. □

Corollary B.1 (Elimination of all positive-order coefficients)**.** *Under the assumptions of Lemma B.4, for every multi-index α with $|\alpha| \ge 1$,*

$$\mathbf{A}_\alpha(x_0) = 0.$$

Since x_0 is arbitrary, it follows that in every local differential representation of $\mathcal{T}$, one has

$$\mathbf{A}_\alpha \equiv 0 \qquad \textit{for all } |\alpha| \ge 1.$$

Proof. If some positive-order coefficient were nonzero at x_0, let

$$m = \max\{\,|\alpha| : \mathbf{A}_\alpha(x_0) \ne 0\,\} \ge 1.$$

Lemma B.4 then implies that every coefficient of order m vanishes at x_0, contradicting the definition of m. Therefore all positive-order coefficients vanish at x_0. Since x_0 was arbitrary, the conclusion follows pointwise everywhere. □

Lemma B.5 (Pointwise reduction)**.** *Under the assumptions above, the operator $\mathcal{T}$ is pointwise. More precisely, there exists a smooth matrix field*

$$\mathbf{W} \in C^\infty(\Omega, \mathbb{R}^{N_{\mathrm{out}} \times N_{\mathrm{in}}})$$

such that

$$(\mathcal{T}F)(x) = \mathbf{W}(x)F(x), \qquad \forall F,\ \forall x \in \Omega.$$

Proof. By Lemma B.1, the operator is locally of finite differential order:

$$(\mathcal{T}F)(x) = \sum_{|\alpha|\le r} \mathbf{A}_\alpha(x)\,\partial^\alpha F(x).$$

By Corollary B.1, all coefficients with $|\alpha| \ge 1$ vanish identically. Hence only the zero-order term remains:

$$(\mathcal{T}F)(x) = \mathbf{A}_0(x)F(x).$$

Defining

$$\mathbf{W}(x) := \mathbf{A}_0(x)$$

proves the claim. □

B.4 Constancy of the Pointwise Coefficient

Lemma B.6 (Compactly supported transport)**.** *For any two points $x, y \in \Omega$, there exists*

$$\mathfrak{g}_{\mathrm{probe}} \in \mathfrak{G}_{\mathrm{probe}}$$

such that

$$\mathfrak{g}_{\mathrm{probe}}(x) = y.$$

Proof. Because Ω is connected and open, there exists a smooth embedded path

$$\gamma : [0,1] \to \Omega$$

with $\gamma(0) = x$ and $\gamma(1) = y$. Since $\gamma([0,1])$ is compact and contained in Ω, one may choose an open tubular neighborhood U with compact closure contained in Ω. Standard extension arguments (see, e.g., **[?]**) provide a smooth time-dependent vector field ζ_t supported in U such that

$$\zeta_t(\gamma(t)) = \dot{\gamma}(t), \qquad t \in [0,1].$$

Let $\Phi_{s,t}$ denote the flow generated by the time-dependent vector field $\zeta_\cdot$, i.e.,

$$\frac{\partial}{\partial t}\Phi_{s,t}(x) = \zeta_t(\Phi_{s,t}(x)), \qquad \Phi_{s,s} = \mathrm{id}.$$

Because ζ_t is compactly supported in U, the time-one map

$$\mathfrak{g}_{\mathrm{probe}} := \Phi_{0,1}$$

belongs to $\mathfrak{G}_{\mathrm{probe}} = \mathrm{Diff}_c(\Omega)$. By construction, the trajectory starting at $\gamma(0) = x$ follows the path $\gamma(t)$, hence

$$\mathfrak{g}_{\mathrm{probe}}(x) = \gamma(1) = y.$$

□

Lemma B.7 (Constancy of the pointwise coefficient)**.** *If*

$$(\mathcal{T}F)(x) = \mathbf{W}(x)F(x)$$

and $\mathcal{T}$ is $\mathfrak{G}_{\mathrm{probe}}$-equivariant, then $\mathbf{W}(x)$ is constant on the connected domain Ω.

Proof. Fix arbitrary

$$x, y \in \Omega.$$

By Lemma B.6, choose $\mathfrak{g}_{\mathrm{probe}} \in \mathfrak{G}_{\mathrm{probe}}$ such that

$$\mathfrak{g}_{\mathrm{probe}}(x) = y.$$

Let

$$v \in \mathbb{R}^{N_{\mathrm{in}}}$$

be arbitrary, and choose

$$H \in D_{N_{\mathrm{in}}}(\Omega)$$

such that $H \equiv v$ on a neighborhood of x. Then

$$\mathfrak{g}_{\mathrm{probe}} \cdot H \equiv v$$

on a neighborhood of y. Therefore,

$$\mathcal{T}(\mathfrak{g}_{\mathrm{probe}} \cdot H)(y) = \mathbf{W}(y)v.$$

On the other hand, equivariance gives

$$\mathcal{T}(\mathfrak{g}_{\mathrm{probe}} \cdot H)(y) = (\mathfrak{g}_{\mathrm{probe}} \cdot \mathcal{T}(H))(y) = \mathcal{T}(H)(x) = \mathbf{W}(x)v.$$

Hence

$$\mathbf{W}(y)v = \mathbf{W}(x)v \qquad \forall v \in \mathbb{R}^{N_{\mathrm{in}}},$$

so

$$\mathbf{W}(y) = \mathbf{W}(x).$$

Thus, $\mathbf{W}$ is constant on Ω. □

B.5 The Main Theorem

Theorem B.1 (Classification theorem for local linear diffeomorphism-equivariant operators)**.** *Let*

$$\mathcal{T} : D_{N_{\mathrm{in}}}(\Omega) \to D_{N_{\mathrm{out}}}(\Omega)$$

be continuous, linear, and local. Assume that $\mathcal{T}$ is $\mathfrak{G}_{\mathrm{probe}}$-equivariant with respect to the pullback action of

$$\mathfrak{G}_{\mathrm{probe}} = \mathrm{Diff}_c(\Omega),$$

that is,

$$\mathcal{T}(\mathfrak{g}_{\mathrm{probe}} \cdot F) = \mathfrak{g}_{\mathrm{probe}} \cdot \mathcal{T}(F), \forall \mathfrak{g}_{\mathrm{probe}} \in \mathfrak{G}_{\mathrm{probe}}, \forall F \in D_{N_{\mathrm{in}}}(\Omega).$$

Assume further that channels are scalar-like, so the group acts only on spatial coordinates. Then there exists a constant matrix

$$\mathbf{W} \in \mathbb{R}^{N_{\mathrm{out}} \times N_{\mathrm{in}}}$$

such that

$$(\mathcal{T}F)(x) = \mathbf{W}F(x), \qquad \forall F \in D_{N_{\mathrm{in}}}(\Omega),\ \forall x \in \Omega.$$

Proof. By Lemma B.1, $\mathcal{T}$ is locally a finite-order differential operator. By Lemma B.5, all positive-order derivative terms vanish, so

$$(\mathcal{T}F)(x) = \mathbf{W}(x)F(x)$$

for some smooth matrix field $\mathbf{W}(x)$. By Lemma B.7, the field $\mathbf{W}(x)$ is constant on the connected domain Ω. Therefore, there exists

$$\mathbf{W} \in \mathbb{R}^{N_{\mathrm{out}} \times N_{\mathrm{in}}}$$

such that

$$(\mathcal{T}F)(x) = \mathbf{W}F(x)$$

for all F and all $x \in \Omega$. □

Corollary B.2 (Kernel form)**.** *Let $K(x,y)$ denote the Schwartz kernel of $\mathcal{T}$. Under the assumptions of Theorem B.1,*

$$K(x,y) = \mathbf{W}\,\delta(x-y),$$

where

$$\mathbf{W} \in \mathbb{R}^{N_{\mathrm{out}} \times N_{\mathrm{in}}}$$

is the constant matrix from Theorem B.1.

Proof. By Theorem B.1,

$$(\mathcal{T}F)(x) = \mathbf{W}F(x).$$

In distributional kernel form, this is exactly

$$(\mathcal{T}F)(x) = \int_\Omega \mathbf{W}\,\delta(x-y)\,F(y)\,dy.$$

Hence the kernel is

$$K(x,y) = \mathbf{W}\,\delta(x-y).$$

□

Remark B.1 (Discrete interpretation)**.** *On a discrete image lattice, constant pointwise channel mixing is exactly a shared 1×1 convolution. Thus, the continuous-domain classification above provides a structural explanation for why rich geometric compatibility collapses admissible local linear spatial mixing to the 1×1 form.*

TABLE B.1
GeoFormer Architecture Details on CIFAR-100.

Item	Setting
Backbone type	Bottleneck Geometric MetaFormer
Embedding	Shared 1×1 convolution
Number of stages	4
Stage widths	$[32,\ 96,\ 288,\ 864]$
Block allocation	$[2,\ 4,\ 6,\ 2]$
Stage transition	AvgPool(2×2, stride 2)
Block form	Bottleneck geometric block
Bottleneck width	$B = \max(16, \lfloor C/4 \rfloor)$ for stage width C
Basis functions	Polar cosine transform (PCT)
Local geometric operator sizes	$5,\ 7$
Local geometric operator orders	$(m, n),\ m, n \in \{0, 1, 2\}$
Readout type	Global geometric pooling
Global geometric pooling orders	$(0,0), (1,0), (0,1), (2,0), (0,2)$
Classifier head	Dropout(0.5) $\rightarrow$ FC(100)
Loss	Label-smoothed cross-entropy with factor 0.1

TABLE B.2
GeoFormer Training Details on CIFAR-100.

Item	Setting
Optimizer	SGD with momentum
Momentum	0.9
Initial learning rate	0.01
Learning-rate schedule	Piecewise decay
Learning-rate drop factor	0.1
Learning-rate drop period	Every 50 epochs
Weight decay	0.001
Mini-batch size	64
Training epochs	100

Remark B.2 (Architectural implication)**.** *Theorem B.1 implies that, under rich scalar-channel geometric equivariance, generic local linear spatial mixing cannot carry nontrivial spatial interaction. Therefore, the spatial interaction mechanism should be delegated to a separate geometry-aware token mixer, while channel interaction remains pointwise. This provides a geometric route to the token/channel decoupling characteristic of MetaFormer-style backbones.*

Remark B.3 (Scope, minimal probe conditions, and the role of $\mathfrak{G}_{\mathrm{probe}} = \mathrm{Diff}_c(\Omega)$)**.** *Theorem B.1 should be read as a structural classification result within the present setting: local linear operators, scalar-like channels, and equivariance under a sufficiently rich local probe class. In particular, it is not a universal classification of all geometry-compatible vision operators. If channels transform under nontrivial geometric representations, if nonlocal operators are permitted, or if equivariance is imposed only with respect to a smaller transformation class, then additional equivariant operators may exist and the conclusion need not hold.*

The use of the full probe group $\mathfrak{G}_{\mathrm{probe}} = \mathrm{Diff}_c(\Omega)$ is sufficient but stronger than strictly necessary. Inspecting the proof shows that the classification argument relies only on two structural ingredients.

First, for every point $x_0 \in \Omega$, the isotropy of the probe class must contain localized isotropic rescalings: namely, for some neighborhood $U \ni x_0$ and some interval $J \subset (0, \infty)$ containing 1, there exist probe transformations $\{\mathfrak{g}_a\}_{a \in J}$ such that

$$\mathfrak{g}_a(x_0) = x_0, \qquad \mathfrak{g}_a(x) = x_0 + a(x - x_0) \quad \text{for all } x \in U.$$

These localized isotropic dilations eliminate all positive-order differential terms, since derivatives of order $m \geq 1$ scale nontrivially under such rescalings, whereas the zeroth-order term does not.

TABLE B.3
Scale-specific GeoFormer Configurations on ImageNet-1K.

Scale	Stage widths	Block allocation	Params (M)	MACs (G)	Extra mixer params (K)
S12	$[64,\ 128,\ 320,\ 512]$	$[2,\ 2,\ 6,\ 2]$	11.9	1.9	3.3
S24	$[64,\ 128,\ 320,\ 512]$	$[4,\ 4,\ 12,\ 4]$	21.3	3.5	6.7
S36	$[64,\ 128,\ 320,\ 512]$	$[6,\ 6,\ 18,\ 6]$	30.8	5.2	10.0
M36	$[96,\ 192,\ 384,\ 768]$	$[6,\ 6,\ 18,\ 6]$	56.1	9.1	13.2
M48	$[96,\ 192,\ 384,\ 768]$	$[8,\ 8,\ 24,\ 8]$	73.3	12.0	17.7

TABLE B.4
GeoFormer Architecture Details on ImageNet-1K.

Item	Setting
Backbone type	Width-preserving Geometric MetaFormer
Embedding	Standard PoolFormer-style embedding
Number of stages	4
Stage widths	Scale-dependent; see Table B.3
Block allocation	Scale-dependent; see Table B.3
Stage transition	Standard PoolFormer-style stage transitions
Basis functions	Polar cosine transform (PCT)
Local op. size	5
Local op. orders	$(m, n),\ m, n \in \{0, 1, 2\}$
Anchor branch	Fixed depthwise PCT kernel with order $(0, 0)$
Detail branches	Up to 8 branches, subject to channel split; min. 8 channels per branch
Branch aggregation	Concatenate detail branches, then add to anchor response
Detail scaling	Learnable per-channel α_{detail}, initialized to 3×10^{-3}
Normalization	GroupNorm1
MLP activation	GELU
Readout type	Global avg. pooling, normalization, fully connected classifier

Second, the probe class must act transitively on the connected domain Ω: for any $x, y \in \Omega$, there must exist a probe transformation sending x to y. This is precisely what forces the remaining pointwise coefficient $\mathbf{W}(x)$ to be independent of position.

Accordingly, the reduction to constant pointwise channel mixing is driven not by the full richness of $\mathrm{Diff}_c(\Omega)$ as such, but by the combination of localized isotropic dilation in the point stabilizer and transitive transport across the domain. We retain $\mathrm{Diff}_c(\Omega)$ in the theorem statement because it provides a standard, self-contained, and technically convenient maximal probe under which these two ingredients are immediate.

Remark B.4 (Vision-facing interpretation via localized similarities)**.** *Although $\mathfrak{G}_{\mathrm{probe}} = \mathrm{Diff}_c(\Omega)$ is mathematically convenient, it is stronger than the transformation families usually invoked as geometric priors in vision. From an application-oriented viewpoint, a more interpretable probe is given by a localized similarity class: compactly supported transports together with localized translations, rotations/reflections, and isotropic dilations.*

This smaller probe is closer to the geometric transformations discussed in the main text—translation, rotation, reflection, and scale—and therefore provides a more vision-facing reading of the theorem. Moreover, it still retains the essential mechanism of the proof, provided that the two key structural ingredients in Remark B.3 remain available: localized isotropic rescaling around every point and transport between arbitrary points in the connected domain.

Under this interpretation, the role of $\mathrm{Diff}_c(\Omega)$ is diagnostic rather than descriptive. We do not claim that natural-image

TABLE D.1
GeoFormer Training Details on ImageNet-1K.

Item	Setting	Item	Setting
Model sizes	S12/S24/S36/M36/M48	Per-GPU batch size	256
Stochastic depth drop rate	$0.1/0.1/0.2/0.3/0.4$	Global batch size	2048 (8 GPUs $\times$ 256)
LayerScale initialization	$10^{-5}/10^{-5}/10^{-6}/10^{-6}/10^{-6}$	Peak learning rate	1.5×10^{-3}
Data augmentation	RandAugment	Optional LR rule	$\mathrm{lr} = 10^{-3} \times \frac{\mathrm{batch}}{1024}$
Repeated augmentation	Off	Warmup learning rate	10^{-6}
Input resolution	224	Minimum learning rate	10^{-6}
Epochs	300	Learning-rate decay	Cosine
Warmup epochs	5	Optimizer	AdamW
Hidden dropout	0	Adam ϵ	10^{-8}
GELU dropout	0	Adam (β_1, β_2)	$(0.9,\ 0.999)$
Classification dropout	0	Weight decay	0.05
Random erasing prob.	0.25	Gradient clipping	None
CutMix α	1.0	Label smoothing	0.1
Mixup α	0.8	CutMix–Mixup switch prob.	0.5

data are literally symmetric under arbitrary compactly supported diffeomorphisms. Rather, we use a sufficiently rich local coordinate probe to test how restrictive strong geometric compatibility becomes when channels are kept scalar-like. In this sense, the theorem identifies the structural price of demanding rich local geometric compatibility, while the localized-similarity viewpoint explains why the resulting architectural consequence remains relevant to practical vision priors.

APPENDIX C
IMPLEMENTATION DETAILS ON CIFAR-100

For completeness, Tables B.1 and B.2 summarize the GeoFormer instantiation used in the CIFAR-100 experiments. The implementation is chosen to match the experimental purpose and the theoretical design of Sec. 4 and Sec. 5: shared 1×1 embeddings avoid introducing additional unconstrained spatial mixing; bottleneck blocks allow each reduced channel to be processed by the full family of geometric branches; average-pooling stage transitions provide a geometry-friendlier hierarchical reduction; and the final readout remains geometric.

APPENDIX D
IMPLEMENTATION DETAILS ON IMAGENET-1K

For completeness, Tables B.3, B.4, and D.1 summarize the GeoFormer instantiation used in the ImageNet-1K experiments. The implementation is intentionally designed to preserve the standard MetaFormer scaffold as much as possible, so that the main architectural difference relative to the internal comparison baselines lies in the token mixer rather than in the surrounding backbone template.