VISUAL INVARIANCE FROM IMAGE-DRIVEN LOCAL DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce a computer vision framework where a continuous linear dynamical system (formulated as a PDE) transforms static image pixel arrays into temporal spectral functions for sampling, pooling, and classification. Evaluated on the MNIST handwritten digit dataset, our model is trained strictly on standard upright samples, yet demonstrates robust generalization to test samples subjected to severe, unseen geometric transformations—including rotation, anisotropic scaling, shearing, and elastic warping. Unlike prior approaches, our framework does not require a priori knowledge of the test-time transformations or brute-force data augmentation. Furthermore, continuous PDE dynamics offer a biologically plausible alternative to standard discrete spatial processing, suggesting that continuous dynamical evolution can establish stable feature trajectories across neural network architectures more efficiently than conventional iterative mechanisms.

1 INTRODUCTION

Baby animals do not learn visual objects from collections of static, standardized images such as those in the MNIST dataset. They encounter objects through continuously changing visual input, such as viewpoint, posture, motion, illumination, and the objects themselves produce substantial perturbations and deformations. To recognize an object, the visual system must therefore extract properties that remain stable across such changes. In conventional computer vision, an image is commonly represented on a fixed pixel lattice, with each intensity value associated with a global row and column position. However, under deformation, the same physical part of an object may occupy very different positions in this coordinate system, so invariance must be learned, engineered, or otherwise recovered from a representation that changes with the transformation. This problem is especially acute for biological vision because an animal does not know in advance what deformation will occur, and natural deformations are rarely restricted to simple global transformations such as uniform rotation, scaling, or shear. They are often local, nonrigid and only approximately describable by a predefined transformation group. Thus, there is a fundamental tension between representing an object by fixed global coordinates and discovering object properties that remain stable under unknown deformation.

This motivates us to ask whether visual input can instead be treated as a field that drives local dynamics, so that object identity is represented by stable properties of the resulting response trajectory rather than by the pixel array with absolute coordinates.

We consider a damped wave partial differential equation (PDE) that converts the spatial information of the input image to a spatiotemporal representation. In the following PDE, the forcing field $V(x,t)$ is the input image (possibly time-varying) and $u = u(x,t)$ is the system state. ∇^2 is the Laplacian operator. η is the damping coefficient. c and κ are adjustment parameters:

$$\frac{\partial^2 u}{\partial t^2} + \eta \frac{\partial u}{\partial t} = c^2 \nabla^2 u - \kappa \left[u - V(x,t) \right]. \tag{1}$$

Equivalently,

$$u_{tt} + \eta u_t - c^2 \nabla^2 u + \kappa u = \kappa V(x,t). \tag{2}$$

Equation (1) can be viewed as a driven damped-wave equation describing the evolution of a sensory field, with the input image continuously stimulating the field as it evolves over time.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To recognize an object, we pool and sample the response trajectory $u(x,t)$ before passing it to a shallow classifier. Spatial pooling substantially reduces dependence on the absolute global coordinate system, while sampling the evolving response at multiple times captures features of the trajectory that may remain stable under image transformations and deformations. This conversion from a static spatial pattern to an evolving spatiotemporal signal is also biologically plausible, since neural communication is inherently dynamical and sensory information is propagated through time rather than represented only as a static spatial array. The use of a Laplacian operator is likewise biologically motivated. Early visual processing begins with center–surround receptive fields in the retina, which are commonly modeled by a difference of Gaussians and can be closely approximated by a Laplacian-of-Gaussian operator.

In our algorithm, each image drives the above fixed second-order PDE with small internal translation and scale motion. Twenty PDE response states are summarized into a 442-D temporal descriptor. Eight object-size listeners independently feed a single 256-unit GELU/LayerNorm hidden layer and a 10-class linear output. The probabilities of the listeners are averaged. Training uses only upright MNIST images; no externally rotated, translated, scaled, sheared, blurred, or elastically deformed examples are shown during fitting; no radial descriptors are used. The hidden bias and LayerNorm affine parameters are frozen, so the main hidden object learned is the 256×442 direction matrix W. Henceforth, this algorithm is referred to as **W256**. In the untouched official test set, the mean accuracy over $0^\circ - 90^\circ$ rotations is 96.36%, and 9 of 12 unseen/unknown deformation conditions tested achieve at least 90% top-1 accuracy. We note that the present experiments were developed within a relatively short period, and substantial improvements may still be possible. Table 1 lists the preliminary results.

Table 1: Model performance under unseen test conditions.

Unseen Test Condition	**W256 Top-1**
identity	**96.73%**
rotation 30°	**96.31%**
rotation 45°	**95.95%**
rotation 60°	**96.20%**
rotation 90°	**96.73%**
shear 15°	**95.46%**
anisotropic $1.20/0.85$	**92.88%**
elastic 1.5 px	89.43%
contrast gain 1.30	**96.56%**
blur 0.8	87.92%
isotropic scale 0.80	88.61%
isotropic scale 1.20	**90.08%**

The MNIST used here is a controlled transformation laboratory, not an ecological model of natural vision. What we demonstrated is that a general-purpose sensory dynamics plus a shallow learned organizer can acquire a surprisingly transformation-stable artificial concept system from upright examples alone. We have not yet had the opportunity to investigate more realistic learning conditions for young animals in which object invariance must be acquired from naturally perturbed and deformed observations. Nevertheless, the results above provide encouraging evidence that the proposed dynamical representation may support such learning.

Standard MNIST provides a highly stable coordinate frame—upright, centered, similarly scaled digits on a fixed pixel lattice—so very high upright accuracy can benefit substantially from coordinate-specific regularities. But such accuracy may not be realistic for bio visual systems that do not use fixed coordinates. Our system gives up some performance on the highly standardized upright benchmark in exchange for a representation that transfers unusually well away from that benchmark coordinate frame. In our system, the spatial lattice is used to define local dynamics, but explicit image coordinates are largely discarded by local and global pooling before concept learning.

Pooling inevitably introduces the possibility that distinct spatial configurations produce similar responses. However, discrimination is based on a high-dimensional collection of responses across multiple spatial scales and dynamical times, making accidental collisions increasingly unlikely for natural structured inputs.

We present the colab-runnable python script used in the experiments in supplementary materials. The experiments were run on CPU and required approximately 22 minutes wall-clock time. This includes extraction of PDE features for eight listener scales, 60 epochs of upright-only fitting for each $442 \to 256 \to 10$ readout, and evaluation of the frozen ensemble on the full 10,000-image official test set under 12 test conditions. No GPU acceleration was used.

The broader implication of the present results may be that deep learning does not need to obtain all representational depth from backpropagation-trained layers. A fixed spatiotemporal dynamical system can itself generate a deep trajectory of intermediate states, and a shallow learned network can then organize this trajectory for recognition. In this view, PDE evolution partially substitutes dynamical depth for learned depth. Theoretical understanding of when such fixed dynamics can replace part of end-to-end representation learning may be important both for biologically motivated computation and for the design of more efficient learning systems.

2 THE ALGORITHM DESCRIPTION

The sensory front end used in the present experiments is a fixed recurrent spatiotemporal dynamical system. The image is not passed directly to the learned classifier. Instead, it continuously drives an internal field whose state is evolved for a finite number of time steps. The learned readout acts only on a compact description of the resulting dynamical trajectory.

We simulate the evolution of PDE (1) in discrete time steps. Let $u^n(x)$ denote the internal field at discrete time n, where x indexes the spatial lattice. The principal recurrence used is

$$u_{\mathrm{PDE}}^{n+1} = (2-\gamma)u^n - (1-\gamma)u^{n-1} + a\,\Delta_8 u^n, \tag{3}$$

where γ is a damping coefficient, a controls local spatial coupling, and Δ_8 is the fixed eight-neighbor lattice operator

$$\Delta_8 u^n(i) = \frac{1}{8}\sum_{j\in N_8(i)} u^n(j) - u^n(i). \tag{4}$$

In the experiments reported here,

$$\gamma = 0.04, \qquad a = 0.22. \tag{5}$$

The relation of the parameters to PDE (1) is

$$\eta = \lambda + (1-\lambda)\gamma,\ c^2 = (1-\lambda)a. \tag{6}$$

The field is not evolving independently of the sensory input. At each time step it is weakly pulled toward the current internally perturbed image V^{n+1}:

$$u^{n+1} = (1-\lambda)u_{\mathrm{PDE}}^{n+1} + \lambda V^{n+1}, \tag{7}$$

with

$$\lambda = 0.08. \tag{8}$$

Thus the recurrent dynamics retain temporal inertia while remaining continuously tethered to the observed stimulus.

Combining Eqs.(equation 3) and (equation 7), and writing $\alpha = 1-\lambda$, gives

$$u^{n+1} = \alpha(2-\gamma)u^n - \alpha(1-\gamma)u^{n-1} + \alpha a\,\Delta_8 u^n + \lambda V^{n+1}. \tag{9}$$

Rearranging,

$$\begin{aligned} u^{n+1} - 2u^n + u^{n-1} + [\lambda + \alpha\gamma]\left(u^n - u^{n-1}\right) \\ = \alpha a\,\Delta_8 u^n - \lambda\left(u^n - V^{n+1}\right). \end{aligned} \tag{10}$$

This form makes the continuous interpretation particularly transparent. The first term is a discrete second temporal derivative, the second is a first-order damping term, the third represents local spatial coupling, and the last term is a restoring force that relaxes the field toward the current sensory input.

2.1 TIME-DEPENDENT SENSORY FORCING

The forcing field $V(x,t)$ is generated from the same underlying image through a small deterministic internal motion. In the current baseline there is no internal rotation, while translation and scale vary sinusoidally over time. Suppressing interpolation details, the forcing field may be written as

$$V(x,t) = C(x)\, I(T_t x)\,, \tag{11}$$

where I is the input image, $C(x)$ is a fixed smooth spatial cover, and T_t is a time-dependent transformation.

The internal translation follows an approximately circular trajectory,

$$d_x(t) = A_T \sin(\omega t + \phi), \tag{12}$$

$$d_y(t) = A_T \sin\left(\omega t + \phi + \frac{\pi}{2}\right), \tag{13}$$

with

$$A_T = 3 \text{ pixels}. \tag{14}$$

The scale varies as

$$s(t) = 1 + A_S \sin\left(\omega t + \phi + \frac{\pi}{3}\right), \tag{15}$$

with

$$A_S = 0.20. \tag{16}$$

The phase ϕ is deterministic for each image. These perturbations are part of the internal sensory dynamics and should be distinguished from external data augmentation: the classifier is not trained on a collection of separately rotated, translated, or scaled examples. Rather, a single evolving field is continuously driven by slightly transformed views of the same input during its internal evolution.

2.2 WEAK SPATIAL REGULARIZATION

After each discrete dynamical update, a small fixed smoothing operation is applied,

$$u \leftarrow 0.99u + 0.01(G * u), \tag{17}$$

where

$$G = \frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}. \tag{18}$$

This operation is not learned. Since

$$u \leftarrow u + 0.01(G - I)u, \tag{19}$$

its continuum effect may be interpreted as a very weak additional local diffusion or numerical regularization. For conceptual clarity, we regard Eq. equation 1 as the principal dynamical equation, and the smoothing step as a small fixed correction to the discrete implementation.

2.3 DEEP DYNAMICS AND SHALLOW LEARNING

The recurrent field is evolved for 20 time steps,

$$u^0 \to u^1 \to \cdots \to u^{19}. \tag{20}$$

The representation supplied to the learned classifier is not just the final state. At each time step, global field statistics and spatial-gradient statistics are recorded. The final descriptor also contains successive temporal differences and low-frequency Fourier magnitudes of these response traces. Hence, the learned representation is a functional of the full solution trajectory,

$$q = \Phi\left(u^0, u^1, \ldots, u^{19}\right). \tag{21}$$

This suggests a useful distinction between dynamical depth and learned depth. The sensory computation is deep in time because the field undergoes many successive recurrent transformations before classification. However, the learned organizer is shallow: in the current baseline, the 442-dimensional trajectory descriptor is passed through only one 256-unit nonlinear hidden layer before

the final linear class readout. Thus, the architecture may be described as deep in dynamical time but shallow in learned parameter depth.

The phrase "solution trajectory" is used here rather than "sample path." The latter commonly refers to a realization of a stochastic process, whereas the present PDE evolution is deterministic once the input image and its deterministic internal phase are specified. Although diffusion and wave equations admit well-known stochastic path interpretations in other settings, the present implementation directly computes a deterministic recurrent field trajectory.

3 GLOBAL PDE MODEL

3.1 GLOBAL PDE PATHWAY TO EXPAND SCALING RANGE

When a young animal views an object, early visual mechanisms first estimate the basic viewing geometry, including the retinal scale, while ocular accommodation helps maintain a focused image. Neural processing can then approximately normalize the gross scale. From this common sensory representation, complementary local and global dynamical pathways may operate in parallel, each producing evidence about object identity. Then, their outputs can be combined according to their reliability to support a perceptual decision.

We develop a model to reflect this mechanism. Two complementary frozen dynamical pathways were evaluated under transformations absent from supervised training. The LOCAL8 pathway uses local eight-neighbor spatial dynamics as before and is particularly robust to rotation and local geometric deformation, whereas the GLOBAL pathway first normalizes gross spatial scale using a second-moment gauge and then applies global operator dynamics with internal translation and event-based temporal readout. Unusually strong scale evidence recruits the GLOBAL pathway, unusually asymmetric branch confidence selects the more reliable pathway, and otherwise the two probability estimates are fused equally. The preliminary results are reported in Table 2.

The global PDE pathway uses the following mathematical model.

$$\psi_{tt}(x,t) + \gamma\psi_t(x,t) = (\alpha\Delta + V(x))\,\psi(x,t), \tag{22}$$

$$h(t) = \int_{D_s} \psi(x,t)dx. \tag{23}$$

Here, the state of the sensor panel $\psi(x,t)$ is a function of the location vector x and the time t. ψ_{tt} and ψ_t are the second- and first order time derivatives of the state. $V(x)$ is the image and γ, α are real constants. D_s is a region for the sampler to gather information with the Laplacian Δ. This formulation converts the spatial signal $V(x)$ into a decaying oscillatory temporal response $h(t)$, which serves as a dynamic representation for similarity and recognition. Later we let $V(x)$ be a time varying $V(x,t)$ to enable motion concept extraction. The equation is based on the physical mechanisms. The light signal of the image is an operator that acts on the retina with the center-surrounding receptive field as a Laplacian. We note that according to the Lie product formula

$$e^{A+B} = \lim_{n\to\infty}(e^{A/n}e^{B/n})^n, \tag{24}$$

the two operators in $(\alpha\Delta + V(x))$ can be understood as acting alternatively on the resultant state of the previous step over small time intervals, which leads to the effect that they are acting simultaneously. In computer algorithms, we approximate $(\alpha\Delta + V(x))$ with the graph Laplacian matrix, leading to a discrete-time evolution over the image lattice for the dynamics of Eqs. (22) and (23).

In Table 2 experiment, both sensory pathways were fitted only with externally upright MNIST. No transformed image or official-test label was used to choose router parameters.

3.2 GLOBAL PDE FOR MOTION CONCEPT EXTRACTION FROM TIME VARYING IMAGES

In this experiment, we begin with a short sequence of image frames that may be interpreted as a time-dependent visual field. Rather than explicitly estimating motion quantities such as optical flow or correspondences, the sequence is modeled as a time-varying operator $V(x,t)$, or equivalently as a sequence of image-dependent graph operators $A_t = A(z_t)$, where each frame defines its own relational structure.

Table 2: Two-branch equal-fusion accuracy across the deformation suite. EQUAL denotes an untrained 50–50 arithmetic average of the LOCAL8 and GLOBAL class-probability vectors, followed by the usual argmax decision.

Deformation	**EQUAL top-1 (%)**
identity	98.49
rotation_30deg	96.58
rotation_45deg	94.36
rotation_60deg	93.48
rotation_90deg	92.36
shear_x_15deg	98.16
anisotropic_x1.20_y0.85	96.87
smooth_elastic_warp_1.5px	95.45
contrast_gain_0.70_zero_bg	98.26
contrast_gain_1.30_zero_bg	98.43
gaussian_blur_sigma0.8	93.98
isotropic_scale_0.60	90.17
isotropic_scale_0.80	97.20
isotropic_scale_1.00	98.49
isotropic_scale_1.20	98.14
isotropic_scale_1.40	97.28
isotropic_scale_1.60	97.15
isotropic_scale_1.80	96.70
isotropic_scale_2.00	96.72
isotropic_scale_2.20	96.66
isotropic_scale_2.40	96.65

For each frame, the image is first normalized to remove trivial global variations. The normalized image then defines a graph operator whose weights reflect local relational structure. Two initial fields are constructed: a raw normalized field and a mean-centered field that captures complementary aspects of the image intensity distribution. These fields serve as initial conditions for a short-time damped wave evolution governed by the operator.

The wave dynamics propagates relational information across the image, converting the spatial structure into temporal signals. From the evolving wave responses, local multi-scale statistics are extracted around a fixed set of anchors. These anchors define a consistent set of "ports" through which the relational state of the image is observed. The result is a sequence of frame-level descriptors

$$h_0, h_1, \cdots, h_T,$$

each summarizing the relational response of the corresponding frame.

Motion information is then extracted not by estimating displacement, but by analyzing the evolution of these descriptors. In particular, temporal differences and combinations are formed, including signed increments $h_{t+1} - h_t$, magnitude-based features $|h_{t+1} - h_t|$, and longer-range differences such as $|h_T - h_0|$. These features naturally separate into signed and magnitude components, encoding directional and non-directional aspects of motion.

The resulting temporal feature vector represents a path through a high-dimensional relational space. A shallow multilayer perceptron (MLP) is then sufficient to map this representation to motion concepts such as rotation, scaling, shear, translation, and their compositions. In particular, no explicit motion field is ever computed; instead, motion is encoded implicitly in the trajectory of operator-induced responses. The results of the experiment are presented in Table 3.

4 Related work

Our PDE-dynamic approach is related to several established lines of work on invariance, equivariance, graph-based learning, and spectral geometry, but it combines them from a different starting point. The central idea is to treat the image not primarily as a fixed pixel array from which increasingly abstract features are extracted, but as a signal that drives a structured dynamical evolution.

Table 3: Motion concept recognition under cross-digit generalization. Models are trained on digits $\{0, 1, 2, 3, 4\}$ and evaluated on $\{5, 6, 7, 8, 9\}$.

Task	**Train Digits**	**Test Digits**	**Accuracy (%)**
Rotation ($\pm 15°, \pm 30°, \pm 45°$)	0–4	5–9	98–100
Transform family (rot/scale/shear/trans)	0–4	5–9	$\sim$99
Compound motion	0–4	5–9	$\sim$98–99

The resulting temporal response is intended to preserve intrinsic image information while reducing dependence on arbitrary global coordinates; recognition is delegated to a comparatively simple readout.

- Reservoir Computing (RC) The present approach is related to reservoir computing (RC), in which fixed recurrent dynamics transform the input into a rich temporal state while learning is largely confined to the readout (6). Related biologically motivated ideas include liquid-state machines, where transient neural dynamics provide a computational substrate (7). RC has also been applied to MNIST recognition by sequentially presenting image columns to an echo-state network and using transformed image representations to improve classification (9). In contrast, our objective is different: externally transformed images are withheld during fitting, so invariance must arise from the internal dynamics and the subsequent organization of their responses. Our model also differs in using a spatially structured PDE as the dynamical medium rather than a randomly connected recurrent network, and in explicitly representing the temporal trajectory of the field generated by a sensory image.
- Capsule networks Capsule networks (5) address transformation robustness by representing entities together with pose variables and using routing-by-agreement to assemble parts into wholes. Matrix capsules, for example, associate an entity activation with a pose matrix and allow lower-level capsules to vote for higher-level poses through learned part–whole transformations.

  The connection is that both capsule networks and our PDE approach try to avoid premature scalar invariance: transformation-dependent information is retained before the final recognition decision. The difference is that capsules build explicit object- and pose-centered representations and use learned routing, whereas our approach seeks a more general signal conversion. The PDE field need not explicitly represent "part," "whole," or "pose." Instead, the image drives a covariant dynamical response, and invariant or relational information is extracted later by temporal and ensemble readouts.
- Group-, harmonic-, and gauge-equivariant networks Group-equivariant CNNs and harmonic networks impose known transformation laws directly on the architecture. Group convolutions extend weight sharing from translations to discrete rotations and reflections, while harmonic networks use circular harmonic filters to obtain continuous rotation-equivariant feature maps.

  Gauge-equivariant (1; 2) networks go further by requiring computations to be independent of arbitrary local reference-frame choices. This is especially close to our concern about coordinate baggage. Gauge-equivariant CNNs formulate convolution on manifolds so that changing local frames changes feature coordinates covariantly but does not change the represented object.

  The commonality is the separation of invariant content from covariant transformation behavior. The distinction is that most equivariant networks begin by specifying the relevant symmetry group or representation law. Our PDE program asks whether useful equivariance and invariance can emerge from an image-driven operator and its temporal response, without encoding a complete transformation group into every learned layer. In this sense, it is less a group-convolution architecture than an operator-dynamics hypothesis.
- Spectral shape analysis and diffusion signatures

  Spectral shape descriptors (8; 10) such as the Heat Kernel Signature also derive representations from Laplacian-generated dynamics. Our formulation differs in the role assigned to the sensory signal. The image may determine not only the initial field but also the operator governing its evolution. For a static image-dependent operator L(I), its eigenvalues determine the temporal frequencies of the response, while projections onto its covariantly

transforming eigenvectors determine the observed modal amplitudes and phases. This provides a natural separation between invariant spectral content and covariant spatial mode structure. Moreover, replacing V(x)by V(x,t)yields a PDE in which a video continuously modifies the propagation operator. Motion is then represented through frequency modulation, changing eigenmodes, and intermode coupling within a persistent dynamical state, rather than through post hoc comparison of independently processed frames.

- Graph convolutional and message-passing networks Graph convolutional networks (3; 4) propagate and transform node features according to a graph adjacency or Laplacian. The classical GNN formulation learns hidden node representations by combining local graph structure with node features.

  The main distinction is the role of propagation. In a conventional GNN, propagation is usually one learned network layer among many. In the PDE approach, propagation is interpreted as an explicit dynamical system whose full trajectory is itself the representation. Time is therefore not merely network depth. It is a controlled physical or mathematical evolution, with interpretable damping, frequencies, initial states, forcing, and sampling strategies.

- Data augmentation Data augmentation is the standard empirical strategy for transformation robustness: rotated, translated, scaled, or otherwise modified examples are added during training so that the network learns approximately invariant predictions.

  Our PDE-dynamic approach uses small jitters during training as opposed to the "real" data augmentation. Small training data jitter is analogous to microsaccades in bio visual systems.

5 Future Work

Several directions remain for further investigation.

First, scalability to natural images and complex scenes is an important next step. The current experiments are conducted on MNIST to isolate representational properties, and extending the framework to larger datasets would confirm its robustness in realistic settings. We note that The computational cost of the proposed framework is dominated by the short-time dynamic evolution and involves a small, fixed number of steps. In contrast to deep convolutional or transformer-based architectures, which scale with depth and feature dimensionality, the proposed formulation separates representation generation from readout, potentially enabling more efficient scaling. Furthermore, the operator-based structure suggests a natural decomposition of large images into regions with shared relational structure, enabling hierarchical or divide-and-conquer extensions. A full evaluation of scalability to large-scale datasets remains an important direction for future work.

Second, extension to continuous-time and continuous-space formulations may provide a closer connection to biological vision and physical systems. More refined discretizations of the operator dynamics could improve both efficiency and interpretability.

Third, integrating system state trajectory based feedback mechanisms into a unified system for perception and control presents an important opportunity.

Finally, the proposed spatial-to-temporal conversion via image-driven dynamics suggests a possible extension to other domains such as language. In large language models, token embeddings interact through attention mechanisms that define a relational structure over sequences. One may view this structure as inducing an operator whose dynamics could generate temporal trajectories analogous to those studied in this work.

From this perspective, concept formation in language models may be interpreted as arising from the evolution of relational states rather than solely from static transformations across layers. While this connection remains speculative, it points toward a broader principle in which learning systems operate by converting structured inputs into dynamical representations for concept extraction. In this sense, both vision and language may be understood as instances of relational systems whose structure becomes meaningful through dynamics.

6 CONCLUSION

We introduced a representation framework in which images are driving a dynamic system that induce short-time trajectories on a sensor state. Within this formulation, spatial structure is transformed into temporal responses, and visual concepts are extracted from the resulting relational trajectories.

Through controlled experiments, we demonstrated that invariant recognition for unseen/unknown deformed objects can emerge from the same underlying mechanism. In particular, invariance was achieved without explicit transformation.

A central theme underlying the proposed framework is a shift from coordinate-dependent representations to relational ones. Conventional vision models operate on pixel grids or token sequences whose structure is tied to explicit spatial coordinates or ordering. Although effective, such representations require that invariance be learned through architectural design or data augmentation.

In contrast, the present approach constructs an image-driven dynamic system that encodes local relations among pixels and derives representations from the system state evolution trajectory. As a result, the representation depends primarily on the relational structure rather than on the particular coordinate system in which the image is expressed. Invariance to unknown transformations then arises naturally from the properties of the induced dynamics, rather than being imposed externally.

This perspective aligns with a broader principle that meaningful representations should be defined by relations rather than by arbitrary coordinate choices. In this sense, perception can be viewed as the extraction of relational invariants through dynamical processes, with coordinates serving only as an initial indexing mechanism rather than as a fundamental component of the representation.

REFERENCES

[1] Taco S. Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral CNN. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2019.

[2] Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pages 2990–2999. PMLR, 2016.

[3] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

[4] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR, 2017.

[5] Geoffrey E. Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with EM routing. In *International Conference on Learning Representations (ICLR)*, 2018.

[6] Mantas Lukoševičius and Herbert Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009.

[7] Wolfgang Maass, Thomas Natschläger, and Henry Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14(11):2531–2560, 2002.

[8] Martin Reuter, Franz-Erich Wolter, and Norbert Peinecke. Laplace–beltrami spectra as "shape-dna" of surfaces and solids. *Computer-Aided Design*, 38(4):342–366, 2006.

[9] Nils Schaetti, Michel Salomon, and Raphaël Couturier. Echo state networks-based reservoir computing for MNIST handwritten digits recognition. In *2016 IEEE 19th International Conference on Computational Science and Engineering (CSE)*, pages 484–491. IEEE, 2016.

[10] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Symposium on Geometry Processing*, pages 1383–1392, 2009.

7 STATEMENTS

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for editing writing and script construction for ideas given by the author. We have not used generative AI tools for other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work for potential plagiarism since all the technical ideas are original by the author. LLM-generated code was verified and tested for correctness by the author. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We uploaded the colab runnable python script for the Table 1 experiments in the supplementary material window.