Unifying Gestalt Principles Through Inference-Time Prior Integration

Anonymous Author(s)
Affiliation
Address
email

Abstract

Gestalt principles such as closure, similarity, continuation, and figure–ground segregation have often been characterized as distinct rules guiding perceptual organization. We demonstrate that these diverse phenomena emerge from a single computational mechanism: inference-time prior integration. Our algorithm, Prior-Guided Drift Diffusion (PGDD), repurposes the same feedback pathways used during backpropagation to refine neural activations during inference, enabling networks to integrate learned statistical regularities with sensory input. Applied to pre-trained networks, PGDD reproduces illusory contours, perceptual grouping, and figure-ground segregation without additional training or architectural modifications. These effects depend on appropriate learned priors, for example, networks trained on object-centric datasets generate Kanizsa illusions, while those trained only on faces or scenes fail to do so. Our results suggest that Gestalt principles are not hardwired perceptual rules but emergent consequences of how neural networks can dynamically combine learned statistical knowledge with incoming sensory evidence. This computational framework bridges artificial and biological vision by showing how inference-time optimization can account for fundamental aspects of human perceptual organization.

1 Introduction

Gestalt psychology identified a set of robust phenomena of perceptual organization: closure, similarity, continuation, and figure–ground segregation [1]. These effects have been central to vision science for more than a century. They demonstrate that perception is not a direct reflection of the stimulus, but instead reflects internal constraints that group elements into coherent wholes. Several theoretical frameworks have attempted to explain these effects. Structural information theory formalized Gestalt laws through preferences for simpler encodings [2]. Bayesian approaches treated grouping as probabilistic inference over common causes [3]. Predictive coding models proposed that feedback carries expectations to generate illusory contours and modulate grouped responses [4, 5]. Despite these efforts, no existing account provides a single mechanistic explanation that unifies all classical Gestalt principles within the same computational process that also underlies typical visual perception.

Neural networks trained for object recognition show alignment with human behavioral and neural activity patterns on natural images [6, 7]. However, they still fall short on many perceptual phenomena tied to Gestalt principles. Prior studies have found that deep CNNs often fail to perceive illusory contours and do not reliably discriminate based on global shape cues, instead relying more heavily on local texture or outlines [8–10]. Some Gestalt-like effects have been reported in convolutional neural networks. For example, Kim et al. [11] showed that closure emerges spontaneously in networks trained on natural images, and Biscione & Bowers [12] reported mixed evidence for grouping. However, these effects do not follow the rules observed in biological systems. In the brain,

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Gestalt phenomena such as illusory contours and figure–ground segregation are characterized by delayed neural responses, and causal dependence on intact feedback pathways [1, 13–15]. While representations in standard feedforward CNNs can sometimes be interpreted as Gestalt-like, they do not exhibit the temporal delays or feedback dependence observed in biological systems, suggesting that they may rely on a different underlying mechanism.

Inspired by the biological evidence that Gestalt phenomena depend on cortical feedback (Fig. 1 A), we hypothesized that the same feedback signals used during learning could also be used during inference. In this work, we propose that the feedback pathways used to propagate error gradients during backpropagation may also serve as a channel for accessing the implicit priors a classifier acquires during training. We devised an algorithm, *Prior-Guided Drift Diffusion* (PGDD), that reuses these learning feedback signals to update activations at inference time, without any additional training or new feedback connections. PGDD applied to both convolutional and transformer pre-trained networks reproduces classic Gestalt demonstrations—including Kanizsa figures, similarity grouping, contour continuation, and figure–ground segregation. Each of these specific stimuli has been the subject of decades of psychological and neurophysiological experiments[1, 13, 14, 16–20], providing a direct link between our computational results and established experimental findings. This demonstrates that all of these phenomena can be unified under a single mechanism: the integration of learned priors with incoming sensory input.

2 Prior-Guided Drift Diffusion Algorithm

Our central hypothesis is that the same feedback signals used for learning can be repurposed during inference to integrate priors with incoming sensory input. To test this idea in generic neural networks, we developed an inference procedure that makes this mechanism explicit. We call this procedure *Prior-Guided Drift Diffusion* (PGDD). PGDD refines activations iteratively by drifting away from noisy or incomplete representations and diffusing toward states consistent with the network's learned data distribution (Fig. 1 B). We implemented PGDD as a feedback-mediated inference-time algorithm to update activations. Given a network f_θ trained on natural images under a robust-to-noise classification objective, PGDD iteratively refines activations to align with learned priors:

$$\mathcal{L}_{\text{PGDD}} = \|r(x_t) - r(x_{\text{noisy}})\|^2 \tag{1}$$

where $r(\cdot)$ denotes representations at a chosen layer, $x_{\text{noisy}} = x + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, and the constraint $\|x_{t+1} - x_{\text{sensory}}\| \leq \epsilon_f$ prevents runaway updates. The update rule is:

$$x_{t+1} = x_t + \eta \nabla_{x_t} \mathcal{L}_{\text{PGDD}} + \xi_t, \qquad \xi_t \sim \mathcal{N}(0, I). \tag{2}$$

This procedure encourages activations to move toward regions of representation space that are more consistent with the network's training distribution while maintaining fidelity to sensory input. A theoretical analysis, provided in the Supplementary (Section A.1), shows how the PGDD update connects formally to gradients of the data distribution, providing a principled justification for its ability to recover Gestalt-like structure.

3 Results

We presented the exact same stimuli used in decades of psychological and neurophysiological experiments on Gestalt perception. Unless otherwise noted, all results were obtained with a robustly trained ResNet-50 on ImageNet [21]; in the Supp.A.4, we further show that a probabilistic variant of PGDD applied to standard (non-robust) networks produces qualitatively similar outcomes. **Gestalt Grouping: Similarity and Continuation** PGDD implemented classical Gestalt grouping principles tested in monkeys [14] through the same mechanism (Fig. 1 C,D). For similarity grouping, elements sharing color or shape characteristics were progressively linked through iterative feedback updates. Good continuation was demonstrated using interrupted line segments. PGDD completed contours across gaps, extending fragmented visual information into smooth, continuous patterns.

Figure-ground segregation principle In primate V1, neurons show enhanced responses to figure regions compared to uniform backgrounds even when local features are identical [13]. PGDD reproduced this texture-defined figure-ground effect by progressively amplifying boundaries while suppressing background patterns, matching both the spatial selectivity and temporal dynamics observed in cortical recordings. PGDD applied to face-vase illusion, often framed as a instance of

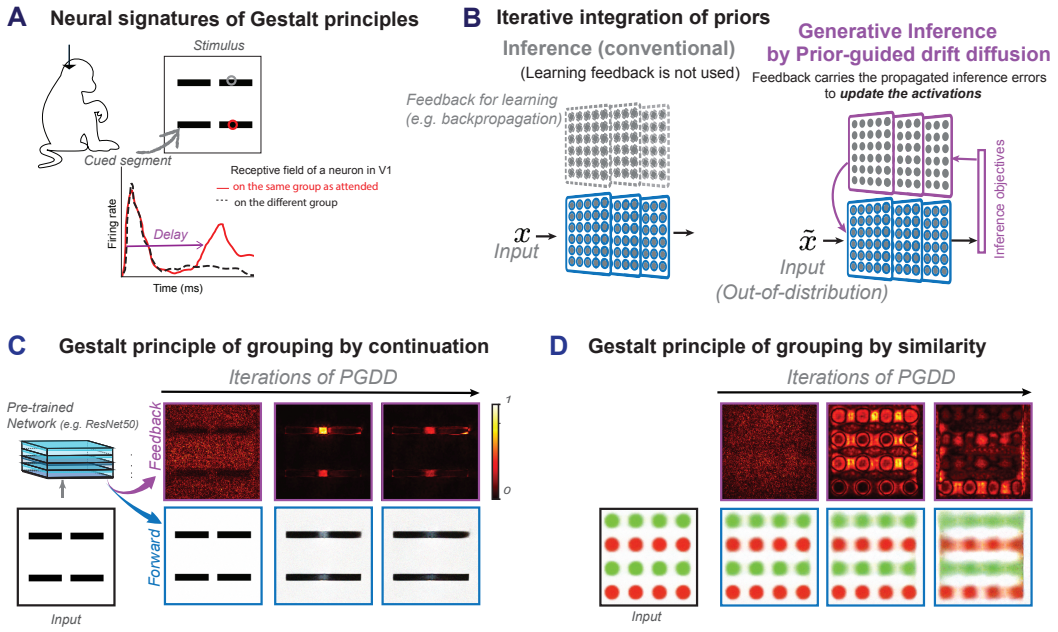

Figure 1: PGDD unifies Gestalt principles through feedback-driven prior integration. **(A)** Neural signatures of Gestalt principles in biological vision show characteristic delayed responses when neurons' receptive fields fall on grouped versus ungrouped elements, indicating feedback-mediated processing [1, 17], see also [13, 14, 16, 19, 20]. **(B)** PGDD repurposes learning feedback pathways during inference. While conventional inference leaves learning feedback connections unused, PGDD reactivates these same pathways (e.g., backpropagation) to iteratively update activations using inference objectives, enabling integration of learned priors with out-of-distribution inputs (Supp.A.2). **(C)** Gestalt principle of grouping by continuation: PGDD applied to the same stimulus in [1] progressively completes segments across iterations, with feedback activation patterns (top row) showing enhanced connectivity while forward activations (bottom row) display completed continuous lines. **(D)** Gestalt principle of grouping by similarity: PGDD groups elements sharing visual features (color, shape) through iterative feedback updates (same stimulus used in [1]).

figure-ground segregation, revealed how different learned priors shape perceptual interpretation (Fig. 2) B. Networks trained for object recognition generated vase-like features through PGDD, while those trained for face recognition produced face-like patterns. Control experiments with simple rectangular shapes showed that both appropriate sensory input and relevant priors are required for structured interpretation.

Closure principle: Kanizsa Figures When presented with Kanizsa triangle inducers, PGDD generated clear illusory contours completing the missing edges (Figure 2). Initial feedforward processing showed no evidence of triangle completion, but iterative prior accumulation progressively structured activations to represent the completed figure. Control experiments confirmed the prior-dependence of this effect. Networks trained on faces or places failed to generate Kanizsa squares, while those trained on object-centric datasets succeeded, highlighting the role of learned statistical regularities. Across all tested Gestalt principles, PGDD operated through the same computational process: iterative accumulation of learned statistical regularities through feedback-mediated inference. The specific manifestation depended on the network's training history and the choice of layer for implementing the PGDD parameters (Supp. A.3), but the underlying mechanism remained the same.

Limitations A key constraint in our current implementation stems from using pretrained neural networks (both robust and standard), which restricts our activation updates to the input layer. In the brain, sensory input remains unchanged while early cortical areas show delayed, feedback-driven responses. Our approach of updating input activations serves as a computational proxy for this biological process. We predict that training neural networks where perturbations are applied to initial layer (rather than inputs) would yield similar Gestalt phenomena while more closely matching the biological implementation. This architectural modification would allow internal representations to

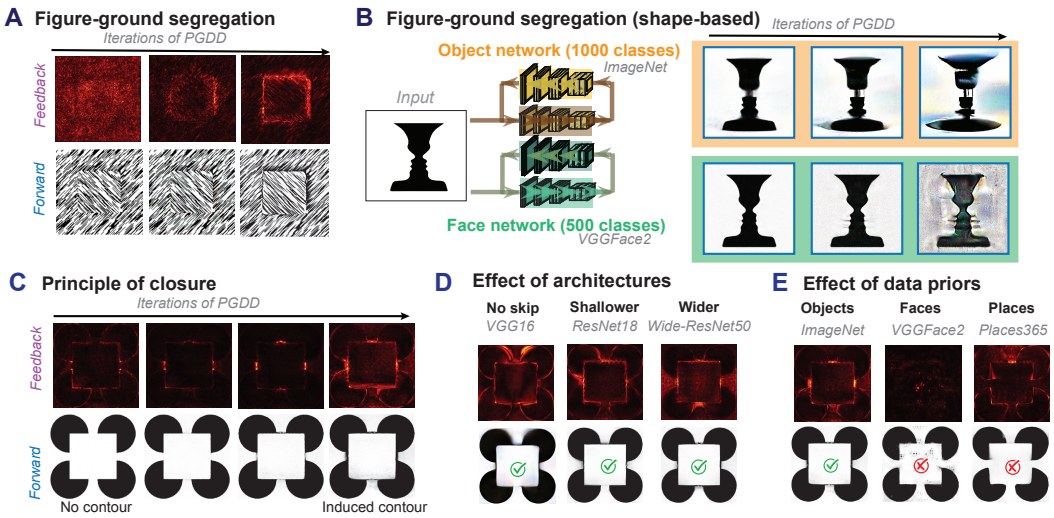

Figure 2: PGDD accounts for instances of Gestalt principles of figure-ground segregation and closure. **(A)** Figure-ground segregation: PGDD applied to texture-defined stimuli (the same stimulus in [13]) progressively enhances figure boundaries through iterative feedback updates, consistent with delayed V1 responses observed in primate studies. **(B)** Prior-dependent interpretation demonstrates the role of learned statistical regularities in ambiguous perception. When presented with the Rubin face-vase [18] stimulus, networks with different training regiems produce distinct interpretations: object networks (ImageNet, top) generate vase-like interpretations while face networks (VGGFace2, bottom) produce face-like patterns through the same algorithm (PGDD) accounting for perception-dependent activity in human's early visual areas [18]. **(C)** Closure through prior integration: Kanizsa square inducers processed by PGDD show progressive emergence of illusory contours. Feedback activations (top) reveal the iterative strengthening of missing edges, while forward activations (bottom) display the completed square structure emerging over iterations, matching the delayed responses to illusory contours in early visual cortex [15]. **(D)** Architectural robustness: PGDD produces consistent closure effects across different network architectures (VGG16, ResNet18, Wide-ResNet50), demonstrating that the mechanism operates independently of specific architectural features. **(E)** Prior-specificity: Closure effects depend critically on appropriate learned priors. Networks trained on object-rich datasets (ImageNet) successfully complete Kanizsa figures, while networks trained exclusively on faces (VGGFace2) or scenes (Places365) fail to generate illusory contours, confirming that statistical regularities from training data drive perceptual completion.

be refined through feedback while maintaining fixed sensory input, better aligning with the layered processing observed in visual cortex.

4 Conclusion

We have shown that classic Gestalt phenomena—including closure, similarity grouping, good continuation, and figure–ground segregation—can all be understood as expressions of a single mechanism: feedback-driven prior integration. By reusing the same feedback pathways that support learning, our theory-grounded algorithm *Prior-Guided Drift Diffusion* (PGDD) enables standard neural networks to reproduce both the perceptual effects and the feedback dependence observed in biological vision. This unification suggests that Gestalt principles are not independent heuristics but emergent consequences of learned priors integrated at inference time. The approach highlights that perceptual organization, long considered a gap between natural and artificial vision, can be realized without new architectures or additional training. More broadly, our results point toward inference-time computations inspired by human and animal brain, where feedback not only enables learning but also actively shapes perception. Such dual use of feedback suggests a bridge between perceptual inference and continual learning, where the same mechanism that guides moment-to-moment interpretation may also support adaptation and knowledge accumulation across experience.

References

[1] Anja Wannig, Liviu Stanisor, and Pieter R. Roelfsema. Automatic spread of attentional response modulation along gestalt criteria in primary visual cortex. *Nature Neuroscience*, 14(10):1243–1244, 2011.

[2] E. Leeuwenberg and P. A. van der Helm. Structural information theory: Towards a theory of perceptual organization. *Psychological Review*, 98(2):267–285, 1991.

[3] V. Froyen, J. Feldman, and M. Singh. Bayesian hierarchical grouping: Perceptual grouping as mixture estimation. *Psychological Review*, 122(4):575–597, 2015.

[4] Rajesh PN Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999.

[5] Zhaoyang Pang, Callum Biggs O'May, Bhavin Choksi, and Rufin VanRullen. Predictive coding feedback results in perceived illusory contours in a recurrent neural network. *Neural Networks*, 144:164–175, 2021. doi: 10.1016/j.neunet.2021.08.024.

[6] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, May 2014. ISSN 1091-6490. doi: 10.1073/pnas.1403112111.

[7] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.*, 10(11):e1003915, November 2014.

[8] N. Baker, H. Lu, G. Erlikhman, and P. J. Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, 14(12):e1006613, 2018. doi: 10.1371/journal.pcbi.1006613.

[9] N. Baker, P. J. Kellman, G. Erlikhman, and H. Lu. Deep convolutional networks do not perceive illusory contours. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, Madison, WI, 2018.

[10] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

[11] Been Kim, Emily Reif, Martin Wattenberg, Samy Bengio, and Michael C. Mozer. Neural networks trained on natural scenes exhibit gestalt closure. *Computational Brain & Behavior*, 4 (3):251–263, 2021.

[12] V. Biscione and J. S. Bowers. Mixed evidence for gestalt grouping in deep neural networks. *arXiv preprint arXiv:2203.07302*, 2022.

[13] V A Lamme. The neurophysiology of figure-ground segregation in primary visual cortex. *J. Neurosci.*, 15(2):1605–1615, February 1995.

[14] U. H. Schnabel, C. Bossens, J. A. M. Lorteije, et al. Figure-ground perception in the awake mouse and neuronal activity elicited by figure-ground stimuli in primary visual cortex. *Scientific Reports*, 8:17800, 2018. doi: 10.1038/s41598-018-36087-8. Received: 13 March 2018; Accepted: 09 November 2018; Published: 12 December 2018.

[15] Tai Sing Lee and My Nguyen. Dynamics of subjective contour formation in the early visual cortex. *Proceedings of the National Academy of Sciences*, 98(4):1907–1911, 2001. doi: 10.1073/pnas.98.4.1907.

[16] D H Grosof, R M Shapley, and M J Hawken. Macaque V1 neurons can signal 'illusory' contours. *Nature*, 365(6446):550–552, October 1993.

[17] Tai Sing Lee and David Mumford. Hierarchical bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448, 2003.

[18] Lauri Parkkonen, Jesper Andersson, Matti Hämäläinen, and Riitta Hari. Early visual brain areas reflect the percept of an ambiguous scene. *Proceedings of the National Academy of Sciences*, 105(51):20500–20504, 2008. doi: 10.1073/pnas.0810966105.

[19] Alexandr Pak, Esther Ryu, Claudia Li, and Alexander A. Chubykin. Top-down feedback controls the cortical representation of illusory contours in mouse primary visual cortex. *The Journal of Neuroscience*, 40(3):648–660, December 2019. ISSN 1529-2401.

[20] H. Shin, M. B. Ogando, L. Abdeladim, et al. Recurrent pattern completion drives the neocortical representation of sensory inference. *bioRxiv*, 2023.

[21] Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Image synthesis with a single (robust) classifier. 2019.

[22] Harris Drucker and Yann Le Cun. Improving generalization performance using double back-propagation. In *Advances in Neural Information Processing Systems*, volume 5, pages 145–151, 1992.

[23] Andrew Slavin Ross and Finale Doshi-Velez. Improving neural network generalization by reducing input gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[24] Judy Hoffman, Avinash Ravichandran, and Oncel Tuzel. Robust learning with jacobian regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4349–4357, 2019.

[25] Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. In *International Conference on Learning Representations*, 2018.

[26] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning*, pages 854–863, 2017.

[27] Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: Global convergence guarantees for training shallow neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[28] Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Léon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2018.

[29] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smooth-grad: removing noise by adding noise. In *Proceedings of the 34th International Conference on Machine Learning Workshops*, 2017.

[30] Jonas Kubilius, Martin Schrimpf, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. In H. Wallach, H. Larochelle, A. Beygelzimer, F. D'Alché-Buc, E. Fox, and R. Garnett, editors, *Neural Information Processing Systems (NeurIPS)*, pages 12785—-12796. Curran Associates, Inc., 2019. URL `http://papers.nips.cc/paper/9441-brain-like-object-recognition-with-high-performing-shallow-recurrent-anns`.

[31] Kuan Han, Haiguang Wen, Yizhen Zhang, Di Fu, Eugenio Culurciello, and Zhongming Liu. Deep predictive coding network with local recurrent processing for object recognition, 2018. URL `https://arxiv.org/abs/1805.07526`.

A Supplementary Material

A.1 Theoretical intuition: PGDD as Implicit Denoising

The Prior-Guided Drift Diffusion (PGDD) algorithm operates through a mathematically principled mechanism that leverages implicit denoising capabilities embedded in robust neural networks. Here we provide the theoretical justification for why PGDD successfully accesses learned statistical priors.

Gradient Structure in PGDD Given the PGDD objective:

$$\mathcal{L}_{\text{PGDD}} = \|r(\mathbf{x}) - r(\mathbf{x} + \boldsymbol{\epsilon})\|^2 \tag{3}$$

where $r(\cdot)$ represents network activations at layer ℓ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, the gradient with respect to input $\mathbf{x}$ becomes:

$$\nabla_{\mathbf{x}}\mathcal{L}_{\text{PGDD}} = 2\mathbf{J}(\mathbf{x})^T(\mathbf{r}(\mathbf{x}) - \mathbf{r}(\mathbf{x} + \boldsymbol{\epsilon})) \tag{4}$$

where $\mathbf{J}(\mathbf{x}) = \nabla_{\mathbf{x}} r(\mathbf{x})$ is the Jacobian of representations with respect to input.

Using the linearization $r(\mathbf{x} + \boldsymbol{\epsilon}) \approx r(\mathbf{x}) + \mathbf{J}(\mathbf{x})\boldsymbol{\epsilon}$, this simplifies to:

$$\nabla_{\mathbf{x}}\mathcal{L}_{\text{PGDD}} \approx 2\mathbf{J}(\mathbf{x})^T\mathbf{J}(\mathbf{x})\boldsymbol{\epsilon} = 2\mathbf{J}^T\mathbf{J}\boldsymbol{\epsilon} \tag{5}$$

Low-Rank Structure and Denoising Robust neural networks are known to exhibit reduced Jacobian norms and improved local Lipschitz constants [22–24]. Adversarial training can be understood as operator-norm regularization on $\mathbf{J}$ [25, 26], which reshapes the singular value spectrum. Empirical studies report that most singular values of $\mathbf{J}$ are strongly suppressed while a few dominant modes remain [27, 28]. This produces a low *effective* rank structure: sensitivity is concentrated in a small discriminative subspace while class-insensitive directions are attenuated.

For a low-rank Jacobian $\mathbf{J}$, the operator $\mathbf{J}^T\mathbf{J}$ inherits this structure and acts as a projection onto the span of the network's learned feature directions. When applied to noise $\boldsymbol{\epsilon}$:

$$\mathbf{J}^T\mathbf{J}\boldsymbol{\epsilon} = \mathbf{J}^T\mathbf{J}(\boldsymbol{\epsilon}_{\|} + \boldsymbol{\epsilon}_{\perp}) \tag{6}$$

where $\boldsymbol{\epsilon}_{\|}$ lies in the span of learned features and $\boldsymbol{\epsilon}_{\perp}$ is orthogonal to this span. Due to the low-rank structure:

$$\mathbf{J}^T\mathbf{J}\boldsymbol{\epsilon} \approx \mathbf{J}^T\mathbf{J}\boldsymbol{\epsilon}_{\|} \tag{7}$$

This selective filtering preserves class-relevant perturbations while suppressing orthogonal noise components, effectively implementing a learned denoising operator.

Connection to Score-Based Models The denoising operator $\mathbf{J}^T\mathbf{J}$ in PGDD parallels the score function $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ (estimated by denoiser autoencoders) used in diffusion models. Both operators guide iterative refinement toward higher-probability regions of the learned data distribution. However, unlike explicit generative models, PGDD accesses this denoising capability as an emergent property of robust classification training.

The iterative update rule:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \eta\mathbf{J}^T\mathbf{J}\boldsymbol{\epsilon} + \boldsymbol{\xi}_t \tag{8}$$

where $\boldsymbol{\xi}_t \sim \mathcal{N}(0, \sigma^2_{\text{diff}}\mathbf{I})$ provides stochastic exploration, implements a form of Langevin dynamics guided by the implicit score function encoded in the robust classifier's gradient structure. This theoretical framework explains why PGDD can successfully extract learned priors from networks trained solely for classification, without requiring additional generative objectives or architectural modifications.

A.2 PGDD algorithm

Algorithm 1 Prior-Guided Drift Diffusion Objective

Input: Image x_{input}, model f, target layer ℓ, constraint ϵ, step size α, noise ratio σ, iterations T
Output: Refined representations $\{x_t\}_{t=0}^{T}$
// Step 1: Feedforward pass
$x_0 \leftarrow \text{normalize}(x_{\text{input}})$
$f_\ell \leftarrow \text{extract\_layers}(f, \ell)$ {Extract model up to layer ℓ}
$x_{\text{noisy}} \leftarrow x_0 + \sigma \cdot \mathcal{N}(0, I)$
$r_{\text{anti-target}} \leftarrow f_\ell(x_{\text{noisy}})$ {Generate noisy reference representation}
for $t = 1$ to T **do**
  // Step 2: Inference objective selection
  anti-target $\leftarrow r_{\text{anti-target}}$ {Use noisy reference as target}
  // Step 3: Feedback error propagation
  $h_t \leftarrow f_\ell(x_{t-1})$ {Forward pass through target layers}
  $\mathcal{L}_t \leftarrow \|h_t - r_{\text{anti-target}}\|^2$ {MSE loss in representation space}
  $g_t \leftarrow \nabla_{x_{t-1}} \mathcal{L}_t$ {Gradient via feedback pathways}
  // Step 4: Constrained activation update
  $\tilde{g}_t \leftarrow \alpha \cdot g_t / (\|g_t\| + \text{1e-10})$ {Normalize gradient}
  $\eta_t \leftarrow \text{diffusion\_noise\_ratio} \cdot \mathcal{N}(0, I)$ {Add stochastic noise}
  $x'_t \leftarrow x_{t-1} + \tilde{g}_t + \eta_t$ {Move away from representation of noisy input (anti-target)}
  $x_t \leftarrow \text{project}(x'_t, x_0, \epsilon)$ {Enforce $\|x_t - x_0\|_\infty \leq \epsilon$}
  // Step 5: Iteration control
  {Continue to next iteration}
end for

return $\{x_t\}_{t=0}^{T}$

A.3 PGDD parameters

Note: These parameters represent working values for the demonstrations shown. PGDD operates effectively over broader parameter ranges not detailed here.

Table 1: PGDD parameters for different Gestalt principle instances

Phenomenon	Layer	Noise σ^2	Iterations	Step Size (update rate)
Closure: Kanizsa figures	Last	0.1	50	0.5
Similarity grouping	Layer 3	0.1	50	1.0
Good continuation	Layer 3	0.1	50	0.4
Figure-ground	Layer 3	0.5	100	0.8
Rubin's vase	Last	0.5	100	0.8

A.4 sPGDD: probabilistic variant of PGDD for standard neural networks

While PGDD works optimally with networks trained for robustness to adversarial perturbations [21], PGDD can still be extended to standard networks trained without explicit robustness objectives, we developed a probabilistic variant called smooth Prior-Guided Drift Diffusion (sPGDD). sPGDD leverages averaging gradients over multiple noise-corrupted versions of an input [29]. At each iteration, sPGDD generates multiple noise-corrupted versions of the current state, computes gradients for each sample, and averages these gradients to update the input activation.

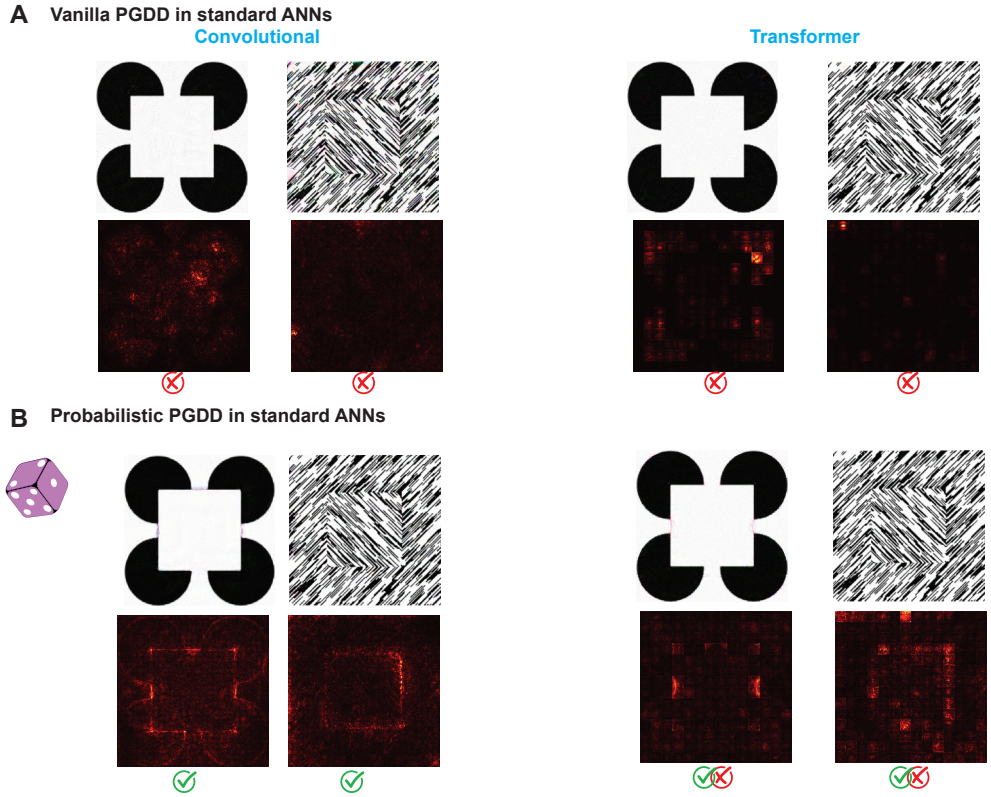

Figure S1: sPGDD extends PGDD to standard neural networks through probabilistic inference. (A) Standard generative inference fails in conventional networks: Both convolutional (ResNet-50) and transformer (ViT-Large) architectures trained without robustness show minimal illusory contour formation when using direct PGDD on Kanizsa squares and figure-ground stimuli. (B) Probabilistic PGDD (or smoothed PGDD) succeeds through gradient averaging [29]: The same networks generate clear illusory contours and figure-ground effects when using sPGDD with gradient averaging over multiple noise samples.

A.5 End-to-end training of feedback as in RNNs and predictive coding

To demonstrate the specificity of PGDD, we investigate alternative frameworks that incorporate feedback processing but use separate feedback for learning and inference. We tested two prominent architectures that explicitly include recurrent or feedback connections trained end-to-end for visual recognition: CORnet-S [30] and PredNetBpE [31].

CORnet-S implements recurrent processing within cortical areas through fixed recurrent weights during inference, effectively extending processing depth but not directly leveraging the same pathways used for error propagation during learning. PredNetBpE implements predictive coding principles with feedback connections that carry top-down predictions and bottom-up prediction errors during both training and inference. Both models were trained on ImageNet for object recognition and tested on the same Gestalt stimuli used to evaluate PGDD. Despite incorporating feedback mechanisms, neither model reliably reproduced the range of Gestalt phenomena observed with PGDD, highlighting that end-to-end training of recurrent feedback (through backpropagation through time) is insufficient to account for these perceptual and neural phenomenon.

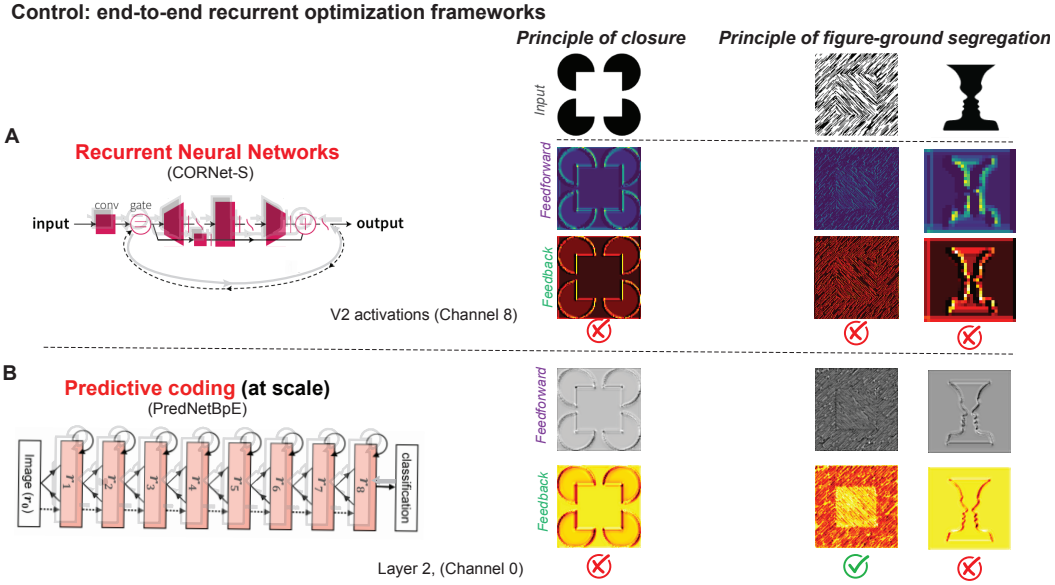

Figure S2: Alternative recurrent frameworks show limited Gestalt processing compared to PGDD. (A) CORnet-S, a biologically-inspired recurrent neural network, fails to generate clear closure effects for Kanizsa figures (red X) and shows weak figure-ground segregation despite its recurrent architecture. (B) PredNetBpE, implementing predictive coding at scale, captures some figure-ground segregation (green checkmark) but fails at closure generation.