Common visual learning constraints in transformers and newborn brains: Evidence from line drawings

Anonymous Author(s)
Affiliation
Address
email

Abstract

A core goal in artificial intelligence (AI) is to build machines that learn like brains. Many AI systems, including convolutional neural networks (CNNs) and vision transformers (ViTs), rival human adults on visual recognition tasks. But, do these AI systems actually learn like brains? If so, AI systems should produce the same learning outcomes as brains when trained with the same data. Here, we tested whether AI systems learn the same object recognition skills as newborn chicks when trained in the same visual environments as chicks. We performed digital twin studies of prior controlled-rearing experiments, evaluating whether CNNs and ViTs produce the same pattern of successes and failures as chicks. When ViTs were equipped with a biologically inspired temporal learning objective, the ViTs showed the same learning patterns as chicks: both learned object recognition when reared with normal objects, but failed to learn object recognition when reared with line drawings. Conversely, when CNNs were equipped with the same temporal learning objective, the CNNs showed a different pattern from chicks: CNNs learned object recognition whether exposed to normal objects or line drawings. These results show that transformers can be accurate image-computable models of visual learning.

1 Introduction

A major scientific and engineering goal is to build machines that learn like brains. For science, this would provide working models for simulating how brains learn to perceive and understand the world. For engineering, this would provide systems that learn with the same power and efficiency as brains. The past decade has produced dozens of success stories in which machines match or exceed the abilities of humans. For example, convolutional neural networks (CNNs) and vision transformers (ViTs) can achieve high levels of performance on a range of tasks, including object recognition [1, 2], action recognition [3], scene perception [4], object segmentation [5], optic flow perception [6], and navigation [7].

But, do AI systems actually learn like brains (i.e., produce the same learning outcomes when trained with the same data)? As many researchers have pointed out [8–11], the training regimes faced by animals and machines differ radically. AI systems are typically trained on massive datasets (e.g., millions of images and videos across thousands of object categories and environments), whereas animals spend their postnatal lives in one environment surrounded by a handful of objects and caregivers. On face, there seems to be a massive mismatch between the volume and variety of training data needed by machines versus animals. Accordingly, AI systems are often regarded as data hungry systems, "gorging on hundreds of terabytes of data," whereas brains are thought to be "efficient and even elegant systems that operate with small amounts of information" [12]. From this perspective, learning in brains and machines is nothing alike.

The view that AI systems are data hungry rests on the assumption that the visual experiences of newborns are impoverished and noisy, a "blooming, buzzing confusion" [13]. However, recent work from

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

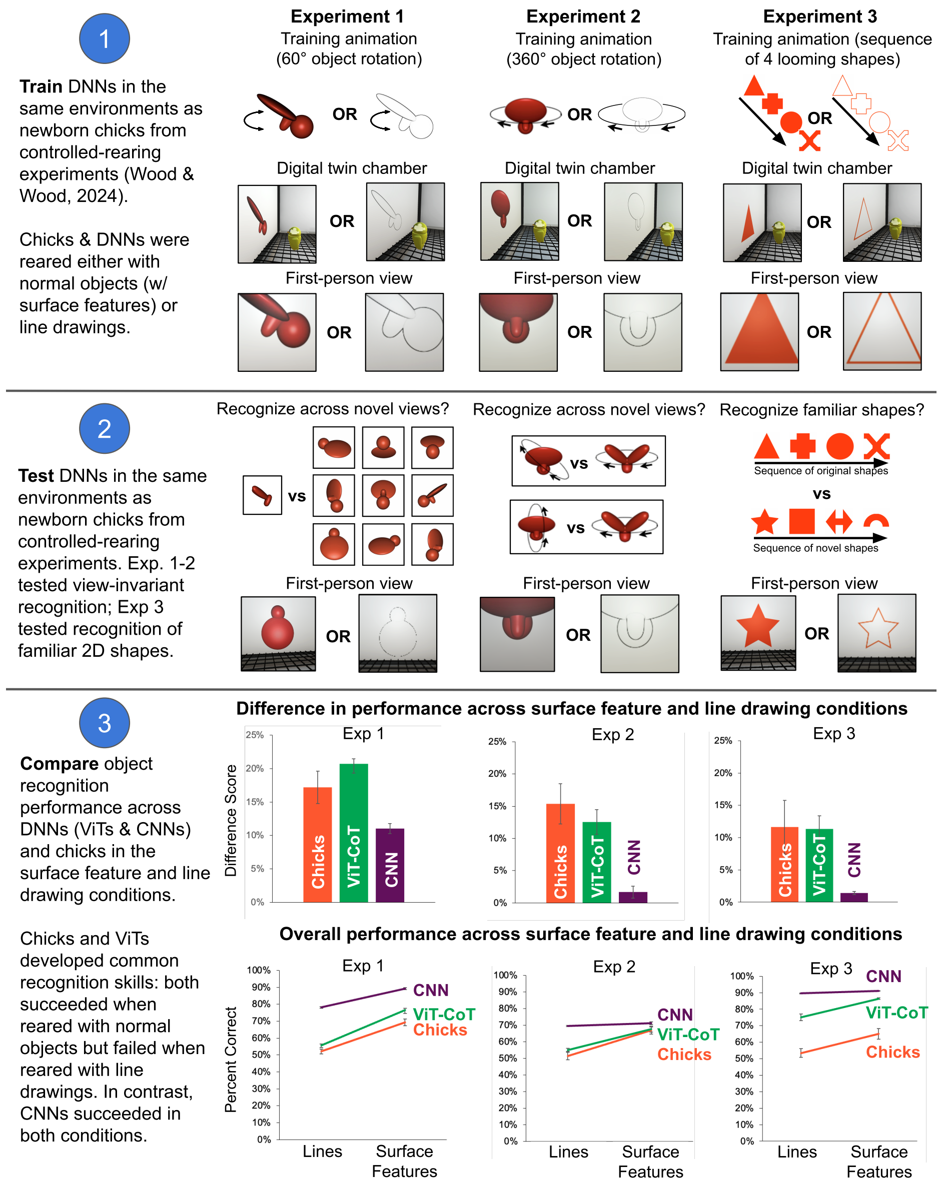

Figure 1: Design and Results. (1) Deep neural networks (DNNs) and newborn chicks were reared in the same visual environments, containing either normal objects or line drawings. (2) DNNs and chicks were tested with the same object recognition tasks. (3) Chicks and vision transformers (ViTs) showed common patterns of development: both learned object recognition when reared with normal objects, and both failed to learn object recognition when reared with line drawings. In contrast, convolutional neural networks (CNNs) learned object recognition in both conditions.

developmental psychology suggests that the opposite is true: The first-person views acquired by babies are temporally and spatially rich [14]. By moving their bodies and heads, babies produce large numbers of diverse, high-quality object views that are well suited for learning[10, 15]. In fact, when first-person views from babies are used to train CNNs and ViTs, the models learn core visual skills [9, 11]. Thus, the gap between human and machine vision might not be as great as previously thought [16].

Human infants acquire a rich visual diet filled with many objects, people, and environments. However, biological visual systems can learn effectively even in impoverished worlds. For example, newborn chicks learn object perception in worlds that contain just a single object [17–20]. Can CNNs and ViTs learn in the same impoverished environments faced by newborn animals?

Digital twin studies [21] are designed to tackle this question, by raising animals and machines in the same environments and testing them with the same tasks. Researchers control and match the training data from which brains and machines learn, allowing for direct comparison of learning outcomes. Prior digital twin studies show that AI algorithms show common learning successes as newborn animals. When CNNs [22, 23] or ViTs [8] are trained on first-person views of agents exploring virtual animal chambers that mimic the rearing conditions of chicks, CNNs and ViTs learn the same object recognition skills as chicks. These results contradict the view that AI algorithms are more data hungry than brains.

The discovery that both CNNs and ViTs can learn effectively in the impoverished environments faced by newborn animals raises a new challenge: how do we distinguish between these model classes? On one hand, CNNs might be the more accurate model class because CNNs and newborn visual systems are both hierarchically and retinotopically organized [24, 25]. On the other hand, the visual system's receptive field structure could be an emergent property of even more foundational (and generic) learning machinery. For instance, fully connected neural networks learn convolutional structures when trained on data with non-Gaussian, higher-order local structure [26]. During prenatal development, brains are shaped by spontaneous retinal waves, which have a non-Gaussian, higher-order local structure [27]. Thus, brains could learn a hierarchical and retinotopic organization during prenatal development, powered by more generic learning mechanisms. If so, the core learning machinery driving visual intelligence would not be CNN-like; rather, it might be more like a transformer. Indeed, ViTs learn CNN-like receptive field structures when trained on natural images [28]. The core computations in transformers also closely match those in the neuron–astrocyte network in the brain [29], raising the possibility that ViTs are the more accurate model of the core learning algorithm in brains.

We tested this hypothesis by evaluating whether CNNs and ViTs show the same successes *and failures* as newborn chicks. Newborn chicks learn better from some experiences than others, so by examining whether CNNs and ViTs show the same pattern of successes and failures as newborn chicks across studies, we can measure which model class learns more like brains. We focused on visual learning from normal objects versus line drawings (objects lacking surface features, Fig. 1). If a chick's visual environment contains normal objects with surface features, then chicks learn to recognize objects across familiar and novel viewpoints [17]. But, if a chick's environment contains line drawings, then chicks fail to develop object recognition [30]. For newborn brains, a visual diet of line drawings is insufficient to learn object recognition.

Line drawings for studying vision. Line drawings have been used for decades to study object recognition. Many studies show that human adults can readily recognize objects depicted in line drawings (e.g., [31–34]). This ability develops rapidly. Infants show enhanced attention to lines that depict corners and edges in the first year of life [35], and young children use lines to depict objects in their earliest attempts to draw the world [36]. Humans have used line drawings to depict scenes since prehistoric times [37, 38]. There is also evidence that nonhuman animals can recognize line drawings, including chimpanzees [39, 40] and pigeons [41].

None of these studies, however, tested humans or animals at the beginning of life. All of the subjects had already acquired months to years of visual experience with real-world objects before they were tested. To explore whether newborn brains can recognize line drawings, Wood and Wood [30] used controlled rearing. The researchers raised newborn chicks in automated controlled-rearing chambers that contained a single object, then tested the chicks' ability to recognize that object across novel viewpoints. When chicks were reared with an object that had surface features, the chicks developed view-invariant object recognition. However, when chicks were reared with a line drawing of an object, the chicks failed to develop object recognition. Do CNNs and ViTs show this same learning pattern? We address this question through digital twin experiments, raising CNNs and ViTs in the same visual environments as chicks and testing them with the same tasks.

2 Methods

Architecture. We used two architectures (CNNs and ViTs) because both are high-performing model classes on a range of visual recognition tasks [42–44] and because the models differ in terms of their hardcoded inductive biases. CNNs have a strong spatial bias. The convolutional operation reflects the spatial structure of natural images, allowing CNNs to generalize well from small datasets and learn useful feature hierarchies that capture the structure of visual images [45, 46]. Conversely, ViTs are generic learning algorithms that do not have hardcoded knowledge about objects or space [43]. Instead, ViTs learn through flexible (learned) allocation of attention that does not assume any spatial (or object) structure.

Objective Function. For each experiment, we performed comparisons between CNNs and ViTs that had the same temporal learning objective function. Based on decades of empirical and theoretical work in neuroscience, we hypothesized that unsupervised temporal learning (UTL) drives visual development in the brain [47–52]. According to UTL models, brains build object representations by adapting to the spatiotemporal statistics of the animal's visual environment. The key assumption underlying UTL is that distal scene variables (e.g., curvature, depth, orientation, texture, shape) vary slowly over time in natural visual environments. Thus, in principle, brains could learn distal scene variables by encoding statistical regularities across successive changes in proximal retinal images (see Appendix A.4 for details).

Our models were initially untrained (no pre-training), and during training, the models were trained on simulated first-person visual experiences from chicks. Like the chicks, the models' visual diet was limited to a single object in a controlled-rearing chamber.

Training Data We used behavioral benchmarks from newborn chicks because chicks can be raised in strictly controlled environments from the onset of vision, providing strict control of all visual experiences (training data) acquired by the animal [53]. This control over training data is essential for directly comparing learning across animals and machines. Chicks can also inform our understanding of human vision because avian brains have similar cells and circuitry as mammalian brains [54–56], as well as a similar large-scale organization, including a hierarchy of sensory information processing, hippocampus regions, and associative areas.

To simulate the visual experiences of newborn chicks, we created realistic digital twins of the controlled-rearing chambers in a video game engine (Unity 3D). Then, we simulated the visual diet available in the chick's environment by recording the first-person images acquired by an agent moving through the virtual chambers. We collected 80,000 first-person images in each of the rearing conditions and used those images to train the CNNs and ViTs (see Appendix section A.2 for details).

3 Results

3.1 Experiment 1: 60° object rotation experience

In Experiment 1 (Fig. 1, *left*), we focused on the view-invariant object recognition task and data reported in Wood [17] and Wood & Wood [30]. Newborn chicks were hatched in darkness, then raised singly in automated controlled-rearing chambers that measured each chick's behavior continuously (24/7) during the first two weeks of life. The chambers were equipped with two display walls (LCD monitors) for displaying object stimuli. The chambers did not contain any objects other than the virtual objects projected on the display walls, providing control over all object experiences acquired by the animal from the onset of vision.

During the training phase, chicks were reared in an environment containing a single virtual object rotating through a 60° viewpoint range. This virtual object was the only object in the chick's environment. The chicks were raised in this environment for 1 week, allowing the critical period on filial imprinting to close. The chicks were raised and tested with either line drawings or objects with surface features.

During the test phase, the chicks were tested on their ability to recognize the imprinted object across 12 in-depth viewpoint changes. On each test trial, the imprinted object appeared on one display wall and an unfamiliar object appeared on the opposite display wall. Test trials were scored as correct when the chicks spent a greater proportion of time with their imprinted object and incorrect when the chicks spent a greater proportion of time with the unfamiliar object. The viewpoint changes introduced large, novel, and complex changes in the object's appearance. Nevertheless, as shown in Fig. 1 (*Panel 3, left*), the chicks reared with surface-feature objects successfully recognized their imprinted object across the novel viewpoints. From a visual diet of a single object, chicks can learn view-invariant object representations. In contrast, when chicks were reared with line drawings of that same object, the chicks never learned to recognize objects. The chicks reared with the line drawings performed at chance level, despite acquiring over 100 hours of visual experience with the line drawings during the training phase.

To compare learning across chicks, CNNs, and ViTs, we performed matching controlled-rearing experiments on CNNs and ViTs (Fig. 1, *Panels 1 & 2*). We created digital twins of the controlled-rearing chambers, then simulated the visual diet in those chambers and used those simulated data streams to train CNNs and ViTs. We then tested the models with the same stimuli used to test the chicks. The chicks and models were trained in the same visual environment and tested on the same task, allowing for direct comparison of their learning outcomes.

Fig. 1 (*Panel 3, left*), shows the performance of CNNs (SimCLR-CLTT) and ViTs (ViT-CoT) in the surface feature and line drawing conditions. SimCLR-CLTT succeeded at the task, learning view-invariant object representations in both conditions. In contrast, ViT-CoT showed the same learning pattern as chicks: ViT-CoT succeeded when learning from normal objects, but failed when learning from line drawings.

3.2 Experiment 2: 360° of object rotation experience

To validate this conclusion under different conditions, we performed a second digital twin experiment of prior controlled-rearing studies [30, 57](Fig. 1, *middle*). Rather than presenting the objects from a 60° viewpoint range, the objects moved through a 360° viewpoint range, completing an in-depth rotation every 15 seconds. The chicks, CNNs, and ViTs were thus exposed to six times as many unique views of the object during the training phase. In the test phase, we tested whether the models could recognize the imprinted object across novel viewpoints, by rotating the object around novel axes of rotation (Fig. 1, *Panels 1 & 2*).

As shown in Fig. 1 (*Panel 3, middle*), when chicks were reared with an object with surface features, the chicks built view-invariant representations that generalized across large, novel, and complex changes in the object's appearance [57]. When chicks were reared with line drawings, they performed at chance level, despite acquiring over 100 hours of visual experience with the line drawings during the training phase [30]. Fig. 1 (*Panel 3, middle*) shows the performance of SimCLR-CLTT and ViT-CoT in the surface feature and line drawing conditions. Again, SimCLR-CLTT succeeded on the task, learning view-invariant object representations in both conditions. In contrast, ViT-CoT showed the same learning pattern as the chicks. ViT-CoT succeeded when learning from normal objects, but failed when learning from line drawings.

3.3 Experiment 3: Looming 2D shapes

To validate our results with different object stimuli, we performed a third digital twin experiment of prior controlled-rearing studies [30, 58]. These studies used simple two-dimensional objects, rather than complex three-dimensional objects. During the training phase, the chicks were presented with a sequence of four looming shapes (Fig. 1, *right*). During the test phase, the chicks were tested on their ability to distinguish familiar shapes from novel shapes.

When chicks were reared with a sequence of shapes containing surface features, they reliably distinguished familiar from novel shapes [58]. In contrast, when reared with a sequence of line drawing shapes, the chicks failed to distinguish familiar from novel shapes [30]. Fig. 1 (*Panel 3, right*) shows the performance of SimCLR-CLTT and ViT-CoT in the surface feature and line drawing conditions. SimCLR-CLTT performed equally well in the surface features and lines conditions. Conversely, ViT-CoT, like the chicks, showed impaired recognition when learning from line drawings.

4 Conclusion

Do AI systems learn like brains? We trained CNNs and ViTs on simulated visual experiences from newborn chicks, and found that temporal learning ViTs (ViT-CoT) showed the same learning patterns as chicks. Both ViT-CoT and chicks learned object recognition when reared with normal objects, but failed to learn object recognition when reared with line drawings. Conversely, CNNs equipped with the same temporal learning objective as the ViTs (SimCLR-CLTT) did not show this pattern: SimCLR-CLTT learned object recognition from both normal objects and line drawings. Appendix A.1 contains additional experiments using alternative architectures and objective functions, and Appendix A.7 contains a more detailed discussion of the limitations and broader impacts of this work. Transformers, but not CNNs, showed the same visual learning pattern as chicks. We conclude that transformers can be accurate image-computable models of visual learning in newborn brains.

References

[1] Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. *arXiv preprint arXiv:2309.16779*, 2023.

[2] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.

[3] Anwaar Ulhaq, Naveed Akhtar, Ganna Pogrebna, and Ajmal Mian. Vision transformers for action recognition: A survey. *arXiv preprint arXiv:2209.05700*, 2022.

[4] Aditya Jonnalagadda and Miguel Eckstein. A foveated vision-transformer model for scene classification. *Journal of Vision*, 22(14):4440–4440, 2022.

[5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[6] Daniel M Bear, Kevin Feigelis, Honglin Chen, Wanhee Lee, Rahul Venkatesh, Klemen Kotar, Alex Durango, and Daniel LK Yamins. Unifying (machine) vision via counterfactual world modeling. *arXiv preprint arXiv:2306.01828*, 2023.

[7] Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. Vint: A foundation model for visual navigation. *arXiv preprint arXiv:2306.14846*, 2023.

[8] Lalit Pandey, Samantha Marie Waters Wood, and Justin Newell Wood. Are vision transformers more data hungry than newborn visual systems? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[9] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, and Daniel L. K. Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021. doi: 10.1073/pnas.2014196118.

[10] Saber Sheybani, Himanshu Hansaria, Justin Wood, Linda Smith, and Zoran Tiganj. Curriculum learning with infant egocentric videos. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[11] A Emin Orhan and Brenden M Lake. Learning high-level visual representations from a child's perspective without strong inductive biases. *Nature Machine Intelligence*, pages 1–13, 2024.

[12] Noam Chomsky, Ian Roberts, and Jeffrey Watumull. Noam chomsky: The false promise of chatgpt. *The New York Times*, 8, 2023.

[13] W. James. *The Principles of Psychology*. Number v. 1 in American science series–Advanced course. H. Holt, 1890.

[14] Linda B Smith, Swapnaa Jayaraman, Elizabeth Clerkin, and Chen Yu. The developing infant creates a curriculum for statistical learning. *Trends in Cognitive Sciences*, 22(4):325–336, 2018.

[15] Sven Bambach, David Crandall, Linda Smith, and Chen Yu. Toddler-inspired visual object learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[16] Justin N Wood. Artificial intelligence tackles the nature–nurture debate. *Nature Machine Intelligence*, pages 1–2, 2024.

[17] Justin N Wood. Newborn chickens generate invariant object representations at the onset of visual object experience. *Proceedings of the National Academy of Sciences*, 110(34):14000–14005, 2013.

[18] Justin N. Wood. Newly hatched chicks solve the visual binding problem. *Psychological Science*, 25(7):1475–1481, 2014.

[19] Justin N Wood and Samantha MW Wood. One-shot learning of view-invariant object representations in newborn chicks. *Cognition*, 199:104192, 2020.

[20] Samantha M. W. Wood and Justin N. Wood. One-shot object parsing in newborn chicks. *Journal of Experimental Psychology. General*, 2021.

[21] Justin N Wood, Lalit Pandey, and Samantha MW Wood. Digtal twin studies for reverse engineering the origins of visual intelligence. *Annual Review of Vision Science*, 2024. In press.

[22] Lalit Pandey, Donsuk Lee, Samantha MW Wood, and Justin N Wood. Parallel development of object recognition in newborn chicks and deep neural networks. Under review.

[23] Donsuk Lee, Pranav Gujarathi, and Justin N Wood. Controlled-rearing studies of newborn chicks and deep neural networks. *arXiv preprint arXiv:2112.06106*, 2021.

[24] Michael J Arcaro and Margaret S Livingstone. On the relationship between maps and domains in inferotemporal cortex. *Nature Reviews Neuroscience*, 22(9):573–583, 2021.

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

[26] Alessandro Ingrosso and Sebastian Goldt. Data-driven emergence of convolutional structure in neural networks. *Proceedings of the National Academy of Sciences*, 119(40):e2201854119, 2022.

[27] Mark V Albert, Adam Schnabel, and David J Field. Innate visual learning through spontaneous activity patterns. *PLoS Computational Biology*, 4(8):e1000137, 2008.

[28] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

[29] Leo Kozachkov, Ksenia V. Kastanenka, and Dmitry Krotov. Building transformers from neurons and astrocytes. *Proceedings of the National Academy of Sciences*, 120(34):e2219150120, 2023. doi: 10.1073/pnas.2219150120.

[30] Justin N Wood and Samantha MW Wood. The development of object recognition requires experience with the surface features of objects. *Animals*, 14(2):284, 2024.

[31] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94(2):115, 1987.

[32] Irving Biederman and Ginny Ju. Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, 20(1):38–64, 1988.

[33] Alumit Ishai, Leslie Ungerleider, Alex Martin, and James Haxby. The representation of objects in the human occipital and temporal cortex. *Journal of Cognitive Neuroscience*, 12 Suppl 2: 35–51, 11 2000. doi: 10.1162/089892900564055.

[34] Dirk B. Walther, Barry Chai, Eamon Caddigan, Diane M. Beck, and Li Fei-Fei. Simple line drawings suffice for functional mri decoding of natural scene categories. *Proceedings of the National Academy of Sciences*, 108(23):9661–9666, 2011. doi: 10.1073/pnas.1015666108.

[35] Albert Yonas and Martha E Arterberry. Infants perceive spatial structure specified by line junctions. *Perception*, 23(12):1427–1435, 1994. doi: 10.1068/p231427. PMID: 7792132.

[36] Jacqueline Goodnow. *Children drawing*. Harvard University Press, 1977.

[37] Jean Clottes. Chauvet cave (ca. 30,000 bc), 2000.

[38] John M Kennedy and Abraham S Ross. Outline picture perception by the songe of papua. *Perception*, 4(4):391–406, 1975.

[39] Shoji Itakura. Recognition of line-drawing representations by a chimpanzee (pan troglodytes). *The Journal of General Psychology*, 121(3):189–197, 1994.

[40] Masayuki Tanaka. Recognition of pictorial representations by chimpanzees (pan troglodytes). *Animal cognition*, 10:169–179, 2007.

[41] Edward A Wasserman, Joseph L Gagliardi, Brigette R Cook, Kim Kirkpatrick-Steger, Suzette L Astley, and Irving Biederman. The pigeon's recognition of drawings of depth-rotated stimuli. *Journal of Experimental Psychology: Animal Behavior Processes*, 22(2):205, 1996.

[42] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems*, 35:10078–10093, 2022.

[43] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[45] Yun-Hao Cao and Jianxin Wu. A random cnn sees objects: One inductive bias of cnn and its applications. In *Proceedings Of The AAAI Conference On Artificial Intelligence*, volume 36, pages 194–202, 2022.

[46] Shuying Liu and Weihong Deng. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian conference on pattern recognition (ACPR)*, pages 730–734. IEEE, 2015.

[47] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

[48] Jacob Feldman and Patrice D Tremoulet. Individuation of visual objects over time. *Cognition*, 99(2):131–165, 2006.

[49] Peter Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3(2): 194–200, 1991.

[50] Edmund T Rolls. Invariant visual object and face recognition: neural and computational bases, and a model, visnet. *Frontiers in Computational Neuroscience*, 6:35, 2012.

[51] James V Stone. Learning perceptually salient visual parameters using spatiotemporal smoothness constraints. *Neural Computation*, 8(7):1463–1492, 1996.

[52] Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002.

[53] Samantha MW Wood and Justin N Wood. A chicken model for studying the emergence of invariant object recognition. *Frontiers in Neural Circuits*, 9:7, 2015.

[54] Onur Güntürkün and Thomas Bugnyar. Cognition without cortex. *Trends in Cognitive Sciences*, 20(4):291–303, 2016.

[55] Erich D Jarvis, Onur Güntürkün, Laura Bruce, András Csillag, Harvey Karten, Wayne Kuenzel, Loreta Medina, George Paxinos, David J Perkel, Toru Shimizu, et al. Avian brains and a new understanding of vertebrate brain evolution. *Nature Reviews Neuroscience*, 6(2):151–159, 2005.

[56] Harvey J Karten. Neocortical evolution: neuronal circuits arise independently of lamination. *Current Biology*, 23(1):R12–R15, 2013.

[57] Justin N Wood and Samantha MW Wood. The development of newborn object recognition in fast and slow visual worlds. *Proceedings of the Royal Society B: Biological Sciences*, 283 (1829):20160166, 2016.

[58] Samantha MW Wood, Scott P Johnson, and Justin N Wood. Automated study challenges the existence of a foundational statistical-learning ability in newborn chicks. *Psychological Science*, 30(11):1592–1602, 2019.

[59] David D Cox, Philip Meier, Nadja Oertelt, and James J DiCarlo. 'breaking'position-invariant object recognition. *Nature Neuroscience*, 8(9):1145–1147, 2005.

[60] Taosheng Liu. Learning sequence of views of three-dimensional objects: The effect of temporal coherence on object memory. *Perception*, 36(9):1320–1333, 2007.

[61] Guy Wallis, Benjamin T Backus, Michael Langer, Gesche Huebner, and Heinrich Bülthoff. Learning illumination-and orientation-invariant representations of objects through temporal association. *Journal of Vision*, 9(7):6–6, 2009.

[62] Nuo Li and James J DiCarlo. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science*, 321(5895):1502–1507, 2008.

[63] Travis Meyer and Carl R Olson. Statistical learning of visual transitions in monkey inferotemporal cortex. *Proceedings of the National Academy of Sciences*, 108(48):19401–19406, 2011.

[64] Yasushi Miyashita. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, 335(6193):817–820, 1988.

[65] Giulio Matteucci and Davide Zoccolan. Unsupervised experience with temporal continuity of the visual environment is causally involved in the development of v1 complex cells. *Science Advances*, 6(22):eaba3742, 2020.

[66] Justin N Wood. A smoothness constraint on the development of object recognition. *Cognition*, 153:140–145, 2016.

[67] Justin N Wood, Aditya Prasad, Jason G Goldman, and Samantha MW Wood. Enhanced learning of natural visual sequences in newborn chicks. *Animal Cognition*, 19:835–845, 2016.

[68] Justin N Wood and Samantha MW Wood. The development of invariant object recognition requires visual experience with temporally smooth objects. *Cognitive Science*, 42(4):1391–1406, 2018.

[69] Samantha MW Wood and Justin N Wood. One-shot object parsing in newborn chicks. *Journal of Experimental Psychology: General*, 150(11):2408, 2021.

[70] Arthur Aubret, Markus Ernst, Céline Teulière, and Jochen Triesch. Time to augment self-supervised visual representation learning. *arXiv preprint arXiv:2207.13492*, 2022.

[71] Felix Schneider, Xia Xu, Markus R Ernst, Zhengyang Yu, and Jochen Triesch. Contrastive learning through time. In *SVRHM 2021 Workshop@ NeurIPS*, 2021.

[72] Gerald M Edelman. Neural darwinism: selection and reentrant signaling in higher brain function. *Neuron*, 10(2):115–125, 1993.

[73] Olaf Sporns and Gerald M Edelman. Solving bernstein's problem: A proposal for the development of coordinated movement by selection. *Child Development*, 64(4):960–981, 1993.

[74] Andrew Shtulman. Why people do not understand evolution: an analysis of the cognitive barriers to fully grasping the unity of life. *Skeptic (Altadena, CA)*, 16(3):41–46, 2011.

[75] Richard C Lewontin. The units of selection. *Annual review of ecology and systematics*, 1(1): 1–18, 1970.

[76] Stephen Jay Gould. Darwinism and the expansion of evolutionary theory. *Science*, 216(4544): 380–387, 1982.

[77] Uri Hasson, Samuel A Nastase, and Ariel Goldstein. Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron*, 105(3):416–434, 2020.

[78] Donsuk Lee, Samantha MW Wood, and Justin N Wood. Development of collective behavior in newborn artificial agents. *arXiv preprint arXiv:2111.03796*, 2021.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes] ,

   Justification: Our main claim is that ViTs that learn with contrastive learning through time show the same pattern of successes and failures as newborn chick (when trained with the same training environment and tested with the same tasks). This is supported, in particular, by Figure 2.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, we discuss the limitations of our work in Appendix Section A.7.2 (Limitations).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our work does not include any theoretical proofs or mathematical derivations.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide detailed instructions for each experiment in Sections 3.1, 3.2, and 3.3, and instructions for generating the datasets in Appendix A.2. Additionally, information on training and testing the models can be found in Appendix Sections A.3, A.4, A.5, and A.6. Code and data for reproducibility are provided in Appendix Section A.8.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
    (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
    (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
    (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
    (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide open access to our datasets and the models with detailed instructions on our GitHub page in Appendix Section A.8.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all experimental details relevant to replicating our results in the Appendix. Additionally, we also provide information on our computational models (architecture and objective function) in Section 2 (Methods).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We include error bars (showing Standard Error) in all of our bar charts (reported in Fig 1, Panel 3).

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide all the relevant information in Appendix Section A.6 (Training Details).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, all author(s) in this paper have reviewed the NeurIPS Code of Ethics and abide by its rules.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We include a Broader Impacts section, but our societal impacts are limited because we are reporting foundational research that is not tied to particular applications, let alone deployments.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The datasets and models used in this study pose no risks. The datasets consist of virtual 3D and 2D objects that do not pose any safety concerns.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Yes, all the original assets used in this study (if any) are properly cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide open access to all the assets used in this paper (models, datasets, and scripts) in Appendix Section A.8.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our study does not involve research with Human Subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our study does not involve research with Human Subjects. Our study does not require IRB Approvals or Equivalent.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

A Appendix

A.1 Experiment 4: Comparing Different Objective Functions

In Experiments 1-3, we compared CNNs and ViTs that had the same temporal learning objective. We used contrastive learning through time because it implements the UTL principle discovered in neuroscience and behavioral experiments. In Experiment 4, we assessed the contribution of the temporal learning objective by comparing CNNs and ViTs across other objective functions. If the CNNs and ViTs still show the same pattern of performance (i.e., ViTs, but not CNNs, are impaired when learning from line drawings), then the architecture alone would be the main contributing factor for mimicking visual learning in chicks. However, if the pattern changes, then both the architecture and the objective function would be essential for mimicking learning in chicks.

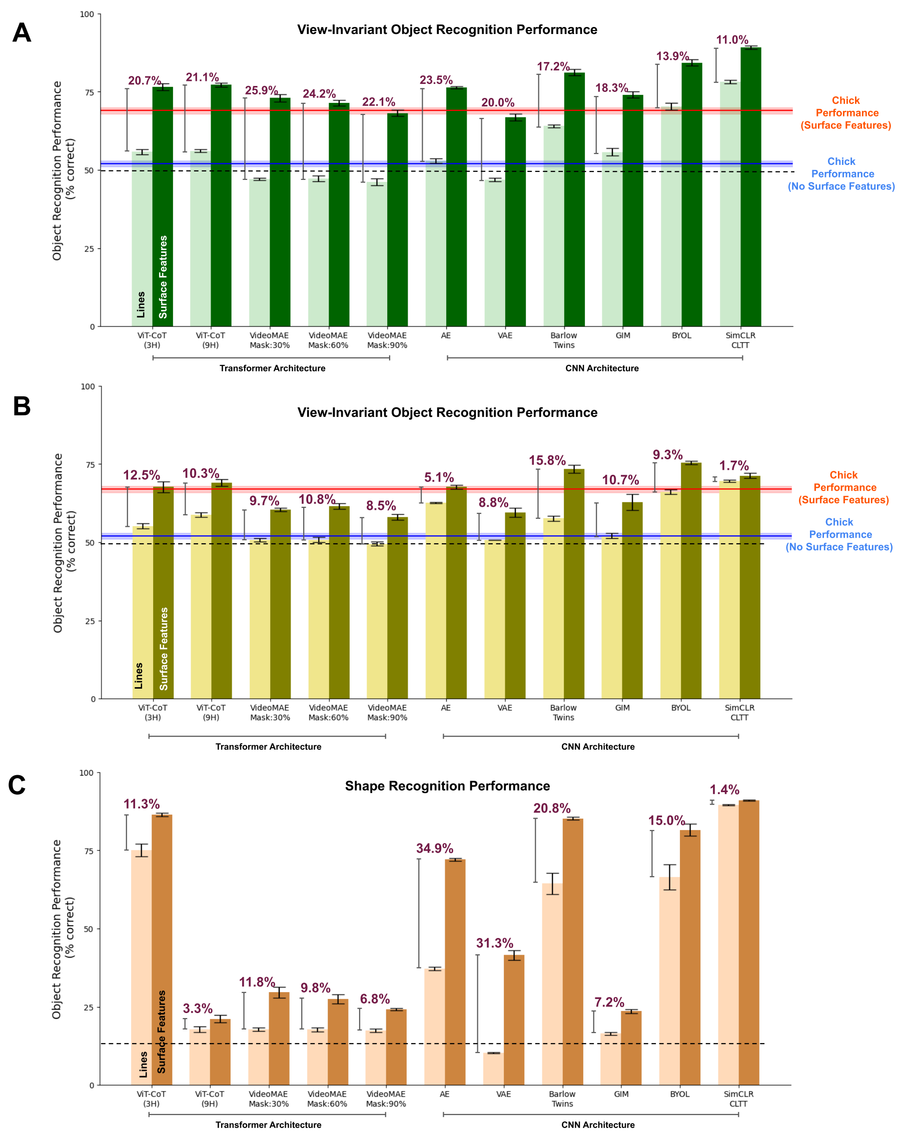

Figure S 1: Object recognition performance of a range of CNNs and ViTs in (A) Experiment 1, (B) Experiment 2, and (C) Experiment 3.

We repeated Experiments 1-3 with four additional ViT models and five additional CNN models. All of the models used different self-supervised objective functions. As shown in Fig. S1, most of the

ViT and CNN models were significantly impaired when learning from line drawings compared to learning from objects with surface features. Yet, many of the CNN models still succeeded in both conditions, learning object recognition even from line drawings (unlike chicks).

Overall, Experiment 4 shows that researchers will need to consider both the architecture and the objective function to build models of visual learning. We show that, by precisely characterizing both the architecture (transformer) *and* the objective function (temporal learning), deep neural networks can serve as accurate image-computable models of visual learning.

A.2 Data Generation

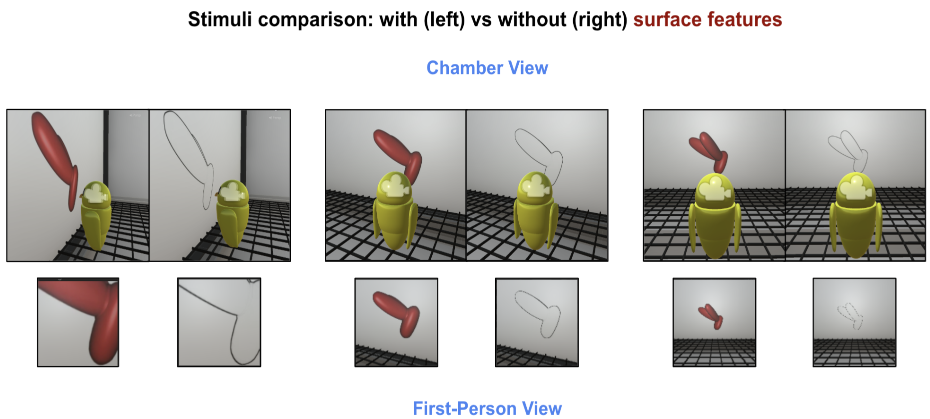

Figure S 2: The virtual chamber in the normal object and line drawing conditions. (Top) The agent visually explores the chamber, randomly moving from place to another. (Bottom) First-person images captured from the camera attached to the agent's head. We use the first-person images to train the ViTs and CNNs.

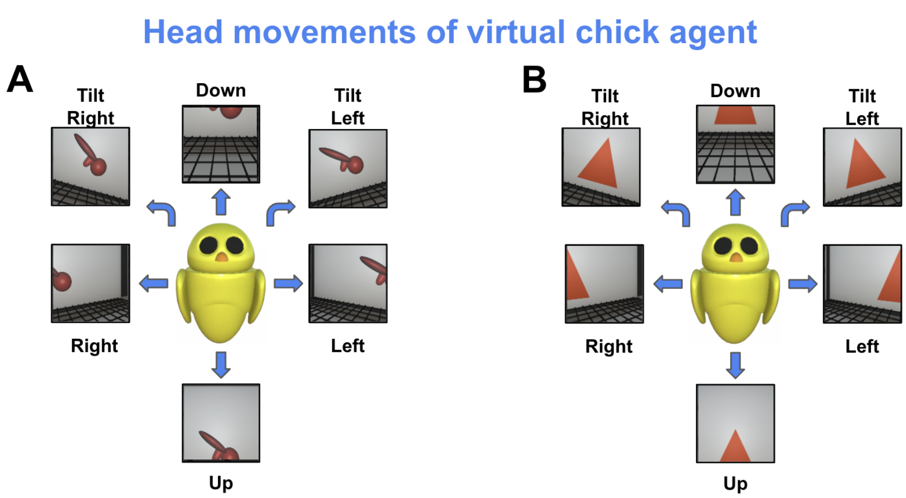

Figure S 3: Head movements of the virtual agent across the three axes of rotations (yaw, roll, and tilt). The agent moved its head 60° on each axis. The images show how head movements provide a form of natural data augmentation.

We created a virtual animal chamber in the Unity Game Engine (Fig. S2). The virtual chamber had two 19" LCD monitors on opposite sides, while the other two sides of the chamber were white walls. The LCDs were used to display virtual 3D or 2D objects moving on a white background at the center of the screens. The floor of the virtual chamber was constructed with black wire mesh and had a provision for food and water next to one of the chamber walls. The dimensions of the virtual chamber were 66 cm (L) x 42 cm (W) x 69 cm (H). The chamber also contained a virtual chick agent; the dimensions of the chick agent were 3.5 units (H) x 1.2 units (L).

This virtual chamber was equipped with two cameras. One camera was placed in the position of the agent's eyes to capture first-person RGB images, simulating the visual experiences of newborn chicks. The second camera was placed on the chamber's ceiling to capture a top view of the agent's movement. To simulate the visual diet available in the chambers, the agent moved to random locations inside the chamber, at a speed of 1.5 units per second. While moving, the agent maintained a constant gaze at the object. Once it reached its destination, the agent then moved its head along all three axes (yaw, roll, and tilt) in a random order (Fig. S3). These head movements lasted for 9.5 seconds. This cycle was repeated until 80,000 first-person images were collected in each rearing condition. The same method was used to collect test data, except the agent kept their gaze fixed on the object.

This simulation approach canvassed the range of visual experiences that chicks could acquire in the chamber. The approach did not directly simulate a specific chick's visual experiences. The approach also did not capture views chicks may have seen of their own bodies (e.g., wings, feet). Our virtual agent could not see its body, so its visual diet was limited to views of the chamber. As such, this approach established a baseline of what could be learned when a model has access to the same visual environment as newborn chicks.

A.3 Architectures

We report all the model architectures and their hyperparameters in Table 1.

Table 1: Architectures and Hyperparameters for various self-supervised learning models

Model	Parameters (M)	Attention Heads	Layers	Learning Objective	Batch Size
ViT-3H	16.9	3	3	Contrastive Learning Through Time	128
ViT-9H	59.4	9	9	Contrastive Learning Through Time	128
VideoMAE-0.3	53.9	6	6	Video Reconstruction	32x8(GPUs)
VideoMAE-0.6	53.9	6	6	Video Reconstruction	32x8(GPUs)
VideoMAE-0.9	53.9	6	6	Video Reconstruction	32x8(GPUs)
SimCLR-CLTT	7.9	NA	10	Contrastive Learning Through Time	512
BYOL	15.9	NA	10	Asymmetric Embedding	512
Barlow Twins	7.9	NA	10	Joint Embedding	512
AE	15.5	NA	10	Image Reconstruction	128
VAE	15.6	NA	10	Image Reconstruction	128
GIM	16.5	NA	10	Non-Backpropagation Contrastive Learning	32

A.3.1 ViT-CoT

We systematically varied the number of attention heads and transformer layers to create different architecture sizes for ViTs. For instance, we used three attention heads and layers to create ViT-3H. For ViT-9H, we increased the number of attention heads and layers to nine. The last layer of the ViT-CoTs generated a 512-dimensional embedding, which was then passed through the loss function. Each architecture was trained using self-supervised learning with a contrastive learning through time objective function. To preserve the temporal relationships between consecutive frames, we did not shuffle the frames in the dataset. Additionally, to avoid hardcoding spatial knowledge in the

ViT-CoTs, we did not use any convolutional layers to generate image patches. The models were trained using images of size 64x64 and a patch size of 8x8. A constant learning rate of 0.0001 was used to train the models.

A.3.2 VideoMAE

In the VideoMAE architecture, both the encoder and decoder blocks had six layers and attention heads. The VideoMAEs were trained by sampling 16 frames from the training set with a temporal stride of 1. Each batch sample had dimensions of (16x3x64x64), where 16 represents the temporal window and 3x64x64 indicates the image dimensions. Subsequently, a random mask of spatial dimension 8x8 and temporal dimension of 2 (2x8x8) was applied to the training batch. The visible patches (non-masked patches) were encoded by the VideoMAE encoder and passed on to the VideoMAE decoder. The decoder combined the encoded features and the masked patches to reconstruct the entire sequence of temporal frames. We experimented with three masking ratios: 30%, 60%, and 90%.

A.3.3 CNN

For the CNN models, we created a custom ResNet architecture (ResNet-10). Each architecture consisted of two residual blocks, totaling of 10 convolutional layers. We used the same bridge connections between the residual blocks as implemented in default ResNets. Similar to the ViT-CoTs, the last layer of the CNNs generated a 512-dimensional embedding, which was then passed through the loss function. To train SimCLR-CLTT, we used a learning rate scheduler with the warm-up epochs set to 5. Additionally, to preserve the temporal relationships between consecutive frames, we did not shuffle the frames in the dataset.

A.4 Objective Function

Many behavioral studies provide evidence that human adults use UTL to learn object representations [59–61]. UTL has also been found on the neurophysiological level in adult monkeys [62–64]. There is even evidence that newborn animals (including chicks) use UTL to build their first object representation [65, 66, 57, 67–69]. These findings suggest that UTL is foundational to visual learning.

To incorporate UTL in our models, we used a temporal learning algorithm, Contrastive Learning Through Time (CLTT), that can be implemented in both CNNs [70, 71] and ViTs [8]. CLTT leverages the temporal structure of natural visual experience, without relying on supervision or labeled data (see Fig. S4). The algorithm contrasts temporally adjacent instances (positive examples) against non-adjacent instances (negative examples), thereby learning representations that capture the underlying dynamics, context, and patterns across time.

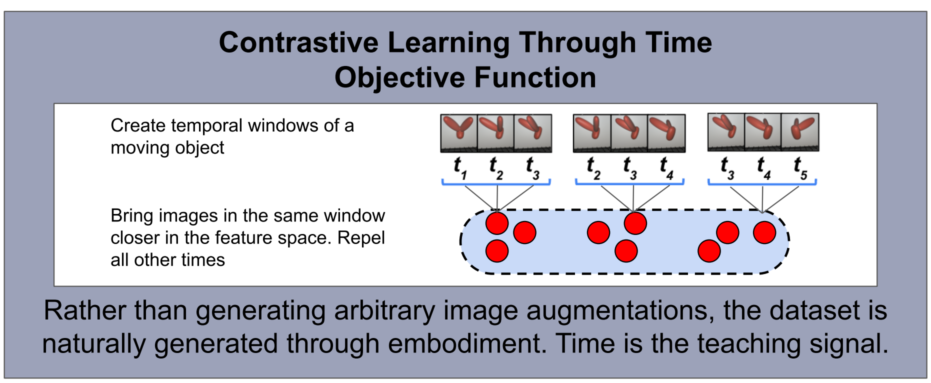

Figure S 4: Contrastive Learning Through Time (CLTT) objective function used with the SimCLR-CLTT (CNN) and ViT-COT (transformer) models. The algorithm pushes together features that occur in the same temporal window (300 ms time window), akin to the 100-400 ms spike-timing-dependent plasticity temporal learning window in brains.

A.5 Evaluation

After training the models (encoders), we evaluated their classification performance using the test stimuli. Task performance was assessed by removing the last fully connected layer of the network, adding a new fully connected linear readout layer on top of the last layer of each trained encoder, and then training only the parameters of the readout layer on the object classification task. The linear classifiers contained 512 neurons, each of which received input from one of the 512 neurons in the final layer of the model. The linear readout layers were optimized for binary cross-entropy loss.

To train and test the linear classifiers, we used the test images collected from the agents moving through the virtual chambers (10,000 images for each of two objects across 12 viewpoint ranges, see Fig. S5). When training the linear classifiers, the object identities were used as the ground-truth labels. Since the encoder weights were frozen, the supervised training of the linear classifiers did not change the features learned by the model.

To evaluate whether the features learned by the models could generalize across novel viewpoints, we used a cross-validated K-fold analysis to train/test the linear classifiers, where each fold contained images from one of the 12 viewpoint ranges. Specifically, the test images were divided into 12 folds, with each fold containing images of each object rotating through 1 viewpoint range. The linear classifiers were cross-validated by training on 11 folds (11 viewpoint ranges) and testing on the held-out fold (1 viewpoint range).

The linear classifiers were trained on 11,000 total images. During training, we used a batch size of 128 for 100 epochs. Transfer performance was evaluated by first fitting the parameters of the linear classifier on the training set and then measuring classification accuracy on the held-out test set. We report average cross-validated performance on the held-out images not used to train the linear readout layer. Thus, all of our results reflect the generalization performance of the models across novel viewpoints.

In Experiment 2, we reused the same linear classifier design from Experiment 1 to conduct binary classification between the two objects. In both experiments, the linear classifier was always trained on 10,000 samples.

In Experiment 3, the linear classifier had 8 output neurons, each corresponding to an object class. We used softmax and categorical cross-entropy loss to train the linear classifier. The training dataset consisted of 8 object classes with each class having 2,500 samples. To construct a test set, we split the training set in half by selecting the initial 1,250 samples from each class. This way, the linear classifier could be trained and evaluated on 10,000 samples (1250 samples x 8 classes).

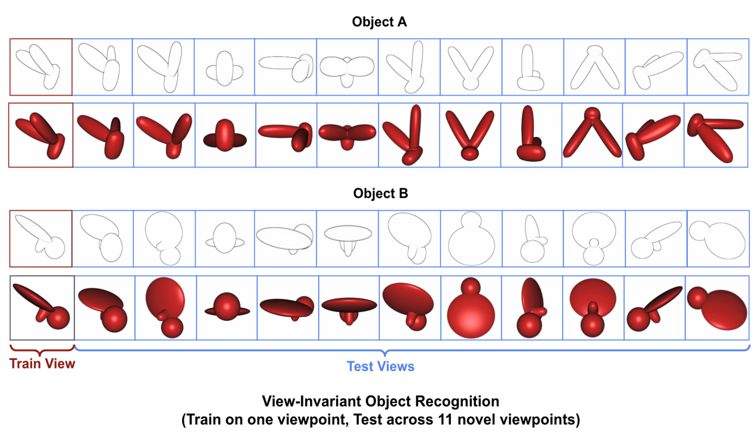

Figure S 5: Viewpoints used in Experiment 1 for the view-invariant object recognition task. The encoder was trained on a single viewpoint and tested on 11 novel viewpoints, using a 12-fold cross-validation design with a linear classifier. The images show object images for the line drawings (top) and normal objects with surface features (bottom).

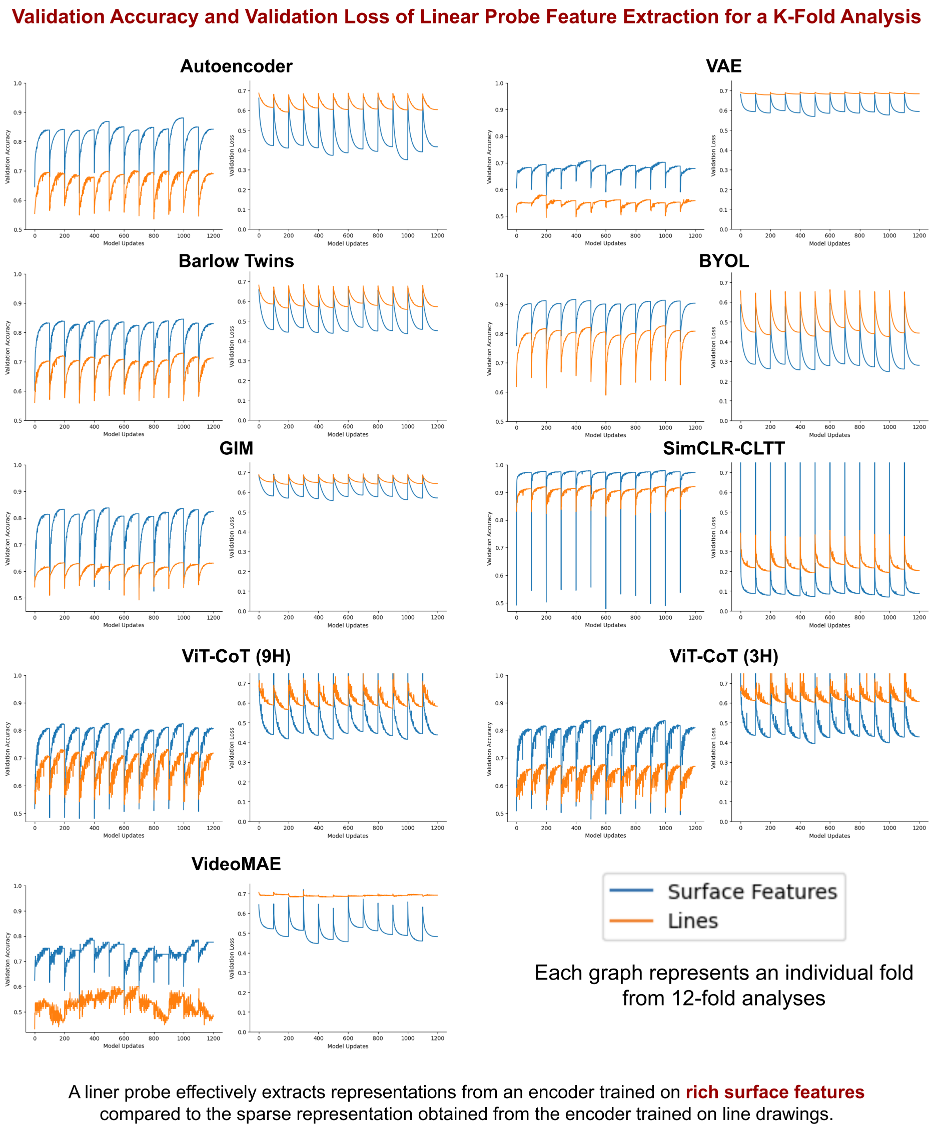

Figure S 6: Linear classifier evaluation results for Experiment 4. The plots show the validation accuracy and loss for the linear classifiers attached to nine different visual encoders. Normal objects with surface features consistently provide stronger and more robust learning signals than line drawings.

A.5.1 Evaluation Results

In Fig. S6, we present the validation accuracy and validation loss data for the linear classifiers trained on frozen encoders in Experiment 1. We observed a consistent pattern across all models: the validation accuracy was high when the linear classifier was evaluated on a frozen encoder trained on normal objects, but it was low when evaluated on line drawings. Similarly, the validation loss was low for encoders trained on normal objects compared to those trained on line drawings. This indicates that linear classifiers can effectively extract rich surface features, but struggle to disentangle features when the encoders are trained on line drawings.

Fig. S7 compares the validation accuracy and validation loss for ViT and SimCLR, both having the same objective function (contrastive learning through time) but different architectures in case

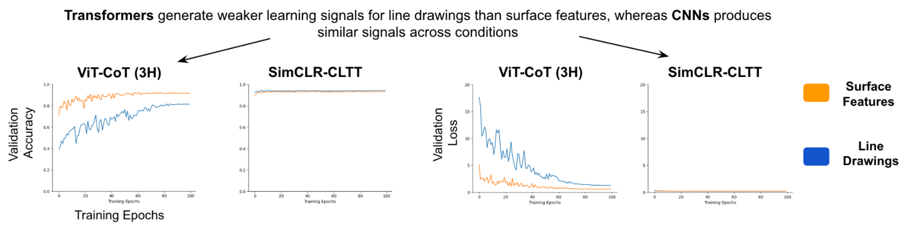

Figure S 7: Linear classifier evaluation results for Experiment 3. The plots compare ViTs and CNNs that have the same contrastive learning through time objective function. Transformers generate weaker learning signals when reared with line drawings versus normal objects (like chicks), whereas CNNs produce similar learning signals for normal objects and line drawings.

of experiment 3. The transformers provide strong training signals to the linear classifier when the encoder is trained on surface features, as opposed to line drawings. In contrast, the CNN backbone provides consistent training signals regardless of whether it is trained on surface features or line drawings.

A.6 Training Details

We trained each model using 3 different seeds and 100 epochs. All models, except VideoMAEs, were trained on a single NVIDIA A10 GPU. VideoMAEs were trained using multi-GPU distributed training across 8 NVIDIA A10 GPUs. Each GPU had 24 gigabytes of memory. We report the number of trainable parameters for each model in Table 1.

A.7 Discussion

Our study provides a new form of guidance for building brain-like AI systems. Researchers have long attempted to build machines that learn like brains, but almost all prior studies compared animals and machines that were raised (trained) in different environments. If animals and machines learn from different training data, then it is impossible to determine whether machines learn like brains (i.e., differences in performance could be due to the algorithm, training data, or some combination of the factors). We overcome this barrier by performing parallel controlled-rearing experiments on newborn chicks and AI algorithms, matching training data across animals and machines. This allowed us to distinguish between candidate model classes (ViTs vs. CNNs) and discover AI systems (ViTs) that show the same learning outcomes as newborn brains. We found that transformers, which are typically considered to be less "brain-like" than CNNs, are the more accurate model of visual learning.

A.7.1 Theoretical simulations of the origins of vision

There is a long history of attempts to characterize the core learning machinery underlying intelligence. Our work expands earlier techniques [72, 73] that used theoretical simulations to study whether blind, evolution-like fitting processes can explain the rapid, self-organized development of visual intelligence. Earlier simulations were limited by compute power, so researchers could not run *image computable* simulations testing whether core visual skills really are learnable via generic fitting machinery. Image computable simulations are essential for testing fitting theories of brain development because the outcomes of evolutionary processes can be counter intuitive [74].

Transformers are ideal models for running simulations of evolution-like fitting processes. The machinery underlying evolution involves blind, brute-force fitting, in which a generic high-dimensional combinatorial medium (the genetic code) adapts to the environment [75, 76]. Likewise, transformers are blind, brute-force fitting systems, in which a generic high-dimensional combinatorial medium (neural networks) adapts to the data distributions in the environment. Both natural selection and transformers start from scratch and produce complex animal forms (natural selection) and mental skills (transformers). Since evolution and transformers operate by common general fitting principles [77], they can be united under a common framework.

We show that transformers, which start from scratch (no prior knowledge of objects or space) and learn through blind fitting, are sufficient to account for successes and failures of visual object learning in newborn chicks. Based on these (and other [8, 78, 77]) findings, we speculate that learning in the brain can be understood in evolutionary terms, as a dynamic high-dimensional system adapting (fitting) to the spatiotemporal data distributions underlying sensory experiences. Under this view, object recognition is not a hard-coded system, structure, primitive, module, or program; rather, it is an emergent property of generic temporal fitting machinery adapting to the embodied visual data streams acquired by newborn animals.

A.7.2 Limitations

One limitation of our study is the models were trained passively, learning from batches of images in a pre-specified order. Newborn animals, in contrast, interact with their environment to produce their own training data. Future studies could close this gap between animals and machines by embodying CNNs and ViTs in artificial agents that collect their own training data from the environment. A second limitation is we do not know *why* the objects with surface features provide better learning signals than line drawings. In Appendix Section A.5.1, we provide preliminary results showing that objects with surface features provide more robust learning signals than line drawings.

A.7.3 Broader Impact

This paper tackles a question at the heart of cognitive science: What are the core learning algorithms in brains? By demonstrating that transformers produce similar learning outcomes as newborn animals, our work shows that transformers can be powerful modeling tools for studying how brains learn to perceive and understand the world. Our work also provides an important step towards building "naturally intelligent" learning systems. Naturally intelligent learning algorithms are an untapped goldmine for inspiring the next generation of machine learning systems.

A.8 Data and Code Availability

The code and data needed to reproduce these findings will be available upon publication.