Scientific Research Publishing

Open Journal of Philosophy, 2026, 16(1), 95-111
https://www.scirp.org/journal/ojpp
ISSN Online: 2163-9442
ISSN Print: 2163-9434

Machine Cognition and Prior Knowledge: A Study Based on Computer Vision Models

Jianwei Sun

School of Philosophy, Beijing Normal University, Beijing, China
Email: sjwneu@icloud.com

How to cite this paper: Sun, J. W. (2026). Machine Cognition and Prior Knowledge: A Study Based on Computer Vision Models. *Open Journal of Philosophy, 16,* 95-111.
https://doi.org/10.4236/ojpp.2026.161006

Received: December 28, 2025
Accepted: January 24, 2026
Published: January 27, 2026

Abstract

By reviewing the evolution and applications of computer-vision models, this paper systematically analyzes the beneficial impact of prior knowledge on machine cognition—namely, improved data efficiency, enhanced robustness, and increased interpretability. Vision models exploit a rich set of a priori image properties—spatial locality, translational invariance, and hierarchical organization—by embedding these priors, explicitly or implicitly, into network architecture, preprocessing pipelines, and regularization terms. Such incorporation enables models to achieve high accuracy and clearer internal representations even when training datasets are scarce. In parallel, we examine the cognitive constraints and potential risks of relying on a priori knowledge. Overly strong priors can restrict the expressive range of machine cognition, introduce subjectivity, and pose difficulties in formally representing the acquired knowledge. This dual perspective underscores both the promise and the pitfalls of integrating prior knowledge into visual-perception systems.

Keywords

Machine Cognition, Prior Knowledge, Computer Vision, Convolutional Neural Networks, Vision Transformer, Hyperparameters, Inductive Bias

1. Introduction

The successive emergence of Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and large-scale self-supervised pre-trained models has dramatically broadened computer-vision capabilities, covering image classification, object detection, semantic segmentation, and even cross-modal generation. In many benchmark settings, these systems now approach—or even surpass—human performance. This rapid progress has sparked intense debate about the relationship between innate knowledge and empirical learning (Cameron, 2023).

Analogous to human cognition, which depends both on innate mechanisms such as memory and imagination and on accumulated experience with the external world, machine vision—viewed as a form of machine cognition—embeds diverse a priori knowledge via inductive biases, thereby imposing structural constraints and providing semantic guidance that shapes the learning process.

Inductive bias refers to prior assumptions that a model makes about the solution space; the model tends to learn certain types of functional relationships (e.g., local correlation, sequence order) while ignoring other irrelevant patterns. Inductive bias is rooted in both the model architecture and the learning algorithm; therefore, model selection and its associated training procedure are essentially equivalent to choosing an inductive bias that aligns with the structure and requirements of the target problem.

It should be noted that inductive bias does not constitute a validation of rationalist nativism. Instead, it demonstrates that more flexible, domain-general neuro-inspired modeling methods can significantly improve performance on certain problems—precisely those problems where domain-specific, manually coded symbolic systems have repeatedly failed (Cameron, 2023).

Current machine cognition still faces several core bottlenecks: the model's heavy reliance on massive labeled data leads to low sample efficiency and high learning costs, making it difficult to adapt to real-world application scenarios with scarce annotations; the lack of commonsense reasoning and causal modeling renders the system vulnerable when encountering distribution shifts, extreme lighting conditions, occlusions, or even abstract concepts, often being able to perform only pattern matching rather than achieve genuine semantic understanding; the black-box nature inherent in models limits the feasibility of error diagnosis and safety auditing, while also hindering long-term cross-task learning and knowledge accumulation.

For this reason, the value of a priori knowledge in machine cognition has become increasingly prominent. A priori knowledge provides prior constraints on the laws governing the real world, and can inject additional signals into models when data is insufficient or disturbed by noise, thereby enhancing generalization ability, robustness, and sample efficiency. Meanwhile, serving as an interpretive bridge, a priori knowledge maps perceptual results to human-understandable concepts, causal relationships, and physical laws, assisting systems in conducting logical reasoning, knowledge fusion, and interpretable decision-making, and advancing machine cognition toward genuine semantic understanding and commonsense reasoning.

Through a comprehensive review of the evolution and applications of computer-vision models, this paper systematically examines how a priori knowledge is implemented and exploited in vision systems. It also investigates the cognitive constraints and potential risks that such priors introduce, offering both theoretical insight and practical guidance for their deep integration into machine-learning pipelines.

The remainder of the paper is organized into five sections:

1) Conceptual Foundations: Clarify the notions of prior knowledge and machine cognition.

2) Evolution of Convolutional Neural Networks: Review the development of CNNs as computer-vision models and analyze how prior knowledge has been incorporated at each stage of their evolution.

3) Forms and Encoding of Prior Knowledge: Examine the various representations of a priori knowledge in vision models and the mechanisms by which it is encoded (e.g., architectural design, preprocessing pipelines, regularization terms).

4) Limitations and Risks: Discuss the cognitive constraints and potential hazards that arise when prior knowledge is embedded in machine-cognitive systems.

5) Implications for Machine Cognition: Summarize the significance of prior knowledge for advancing machine cognition and outline directions for its deep integration with machine-learning methods.

2. Prior Knowledge and Machine Cognition

2.1. Prior Knowledge in Human Cognition

Kant defined a priori knowledge as "the cognitive conditions that precede experience and are independent of it" (Kant, 1998). The brain does not passively receive stimuli during the perceptual process; instead, it constantly compares external inputs with internal a priori knowledge through predictive coding, and the discrepancies (i.e., prediction errors) are used to update internal models or direct attention. From the perspective of cognitive science, a priori knowledge plays a crucial role in human learning and reasoning. It is understood as the "prior assumptions" formed by the innate structures of the brain and the products of long-term learning. These assumptions constitute the knowledge systems, conceptual frameworks, rules, or hypotheses that cognitive subjects possess before encountering new information, enabling people to understand, interpret, and predict new situations more efficiently. Such knowledge manifests itself in innate structures (e.g., the edge detection mechanism of the visual system) or internally constructed models formed through long-term learning (e.g., the grammatical rules of language). Andrew Brook hailed Kant as the intellectual godfather of cognitive science, arguing that Kant's ideas have exerted and continue to exert a profound influence on contemporary philosophy of mind and cognitive science (Schlicht, 2022). The foundation of human cognition is more aligned with what Kant referred to as "synthetic a priori judgments"—that is, humans are inherently endowed with certain cognitive capacities about the world. These capacities are distinctly different from those of other biological species, a distinction now generally regarded as part of the subjectivity of cognition. This innate difference encoded in DNA serves as its ontological foundation (Dong, 2023).

Piaget regarded schema as the basic unit of cognitive development, arguing that children continuously incorporate new experiences into existing schemas through assimilation and accommodation. When a new situation arises, children first at-

tempt to fit it into their current schemas (Piaget, 1970). Assimilation constitutes the direct manifestation of a priori knowledge in learning: existing schemas provide the starting point for interpreting new experiences, determining which features children notice, how they categorize new information, and how they respond to it.

The predictive coding model posits that hierarchical cortical layers update internal models by constantly comparing the error signals generated between predictions and actual inputs. This error-driven learning mechanism enables the brain to interpret complex stimuli in an extremely limited time frame (Clark, 2013; Friston, 2010). Therefore, a priori knowledge plays the role of compressing information, reducing uncertainty, and accelerating interpretation in the cognitive process, allowing humans to instantly transform sparse, noise-ridden sensory data into coherent situational perception.

2.2. Machine Cognition

Within the framework of cognitive science, machine cognition is regarded as the computational realization and simulation of human cognitive processes. It encompasses not only traditional perception-feature extraction and pattern classification, but also the active organization of information, the encoding of short-term and long-term memory, reasoning based on internal models, and the purposeful regulation of behavior in response to the environment. Cognitive scientists define cognition as "a systematic process of acquiring, storing, transforming, and utilizing information" (Anderson, 1983). Machine cognition maps these processes onto artificial neural networks, reinforcement learning, or symbolic-subsymbolic hybrid systems, enabling artificial systems to form a closed-loop cycle between sensory input, internal representation, and action output. For instance, the perceptual module is responsible for converting raw signals into high-level feature representations; the working memory module maintains representations relevant to the current task; the reasoning module performs causal or conceptual deduction on these representations based on existing structural assumptions; and the action module generates behavioral strategies according to the results of reasoning. This complete closed loop is precisely the perception-action coupling emphasized in cognitive science (Clark, 2013; Friston, 2010). Within this framework, the goal of machine cognition is no longer mere function approximation, but rather the construction of a cognitive system that is capable of self-regulation and of interpreting new information by leveraging prior structures.

The a priori plays a decisive role in machine cognition. In cognitive science, the Kantian concept of the a priori is embodied as the brain's predictive models, categorical structures, and the sensible forms of time and space—these structures are precisely the "innate constraints" that machine systems require prior to learning. The translational invariance and local correlation of convolutional neural networks constitute a geometric a priori, enabling the networks to autonomously learn hierarchical feature representations even when labeled data are scarce. A

priori principles such as scale, viewpoint, and illumination invariance further help models maintain robustness under multi-scale, multi-viewpoint, and complex illumination conditions. If deeper geometric constraints (e.g., projection models and depth consistency) or domain knowledge (e.g., object shapes, materials, and physical interaction modes) are explicitly embedded into network architectures or loss functions, they can often significantly enhance the capacity for scene structure understanding, occlusion recovery, and cross-domain transfer.

Through the design of network architectures, the regularization terms of objective functions, and the encoding of parameter distributions obtained from large-scale pre-training, the a priori can provide powerful constraints in scenarios with scarce data or severe noise, enabling models to achieve rapid adaptation even with extremely limited labeled samples. Meanwhile, the a priori also supports cross-task transfer and zero-shot reasoning, as the internal conceptual framework can map new scenarios to existing representational spaces (Lake et al., 2017).

From the perspective of predictive coding, machine systems update their internal a priori distributions by comparing the errors between model predictions and actual perceptual inputs. This error-driven learning mechanism is the core of cognitive systems' ability to achieve self-calibration (Rao & Ballard, 1999). Therefore, the a priori not only enhances the learning efficiency and generalization ability of machine cognition, but also endows the system to a certain extent with the potential to reason about and imagine unknown scenarios.

3. Evolution of CNN and Priors

Since the advent of AlexNet, CNNs have emerged as the core driving force of computer vision and currently serve as the preferred model for tasks such as image classification and object detection. Their core advancements are reflected in architectural innovation, efficiency optimization, and task expansion, while their key value lies in automatic hierarchical feature extraction, efficient processing of complex high-dimensional images, supporting the implementation of full-stack visual tasks, and promoting the industrial popularization of deep learning and visual AI.

The development of convolutional neural networks can be traced back to 1958, when David Hubel and Torsten Wiesel, through experiments on the visual cortex of cats, first observed that neurons in the primary visual cortex are sensitive to moving edge stimuli. They defined simple and complex cells and revealed the hierarchical processing mechanism of the visual system. Their findings demonstrated that as visual information is transmitted from the retina to the cortex, it is gradually abstracted from low-level features (edges) to high-level features (shapes, objects) through multiple layers of neurons. This proof that human vision is formed by the gradual integration of basic features (edges) became the source of inspiration for the hierarchical structure of machine vision.

In the 1970s, David Marr proposed a computational framework for visual processing, emphasizing that visual information needs to undergo hierarchical pro-

cessing (such as edge detection, stereopsis, and shape representation) to gradually construct an understanding of the world. This framework provided theoretical support for the design of the hierarchical architecture of convolutional neural networks, clarifying that visual computation must follow the logic of "from low-level features to high-level representations" and laying down the core framework for subsequent architectural design.

In 1980, Kunihiko Fukushima proposed a neural network architecture incorporating convolutional and pooling layers: the Neocognitron. It mimicked the hierarchical structure of the visual cortex, comprising simple and complex cells, and achieved digit recognition through alternating convolution and pooling operations. The Neocognitron established the rudiments of an architectural a priori, simulating the local receptive fields and feature invariance inherent to human vision. However, it did not incorporate the backpropagation algorithm—an approach independently introduced by Rumelhart, Hinton, and others in 1986. Overly reliant on manually engineered features, the Neocognitron suffered from limitations such as high computational complexity, low training efficiency, and poor adaptability to complex tasks.

In 1998, Yann LeCun refined the Neocognitron to develop LeNet-5, marking the first implementation of the canonical architecture integrating convolutional, pooling, and fully connected layers. By introducing the backpropagation algorithm for network training, LeCun enabled trainable hierarchical feature extraction, transforming the prior static architectural a priori into a dynamic architectural a priori optimizable via data. This breakthrough signaled the transition of CNNs from theoretical constructs to practical applications (e.g., handwritten digit recognition). LeNet-5 achieved a test accuracy of 99.2% on the MNIST handwritten digit dataset and was deployed in check recognition systems for American banks, cementing its status as the first commercially applied convolutional neural network system.

In 2012, AlexNet introduced a brand-new deep architecture, with the number of model layers expanded to include 5 convolutional layers and 3 fully connected layers. In terms of algorithm optimization, it adopted the ReLU activation function to replace the Sigmoid function, addressing the problem of gradient vanishing. For regularization, it incorporated the Dropout strategy to suppress overfitting and employed GPU parallel computing to accelerate training. Such objective optimization and regularization a priori, through the prior constraints of "suppressing gradient vanishing" and "preventing overfitting", significantly enhanced model performance. As a result, AlexNet won the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) with a performance far superior to that of the second-place finisher, demonstrating the potential of deep convolutional neural networks in complex visual tasks.

The success of CNNs is largely attributed to the explicit encoding of human prior knowledge about the visual world into the network architecture. Every architectural advancement in the evolution of CNNs embodies an understanding

and cognition of the target domain—such as locality, translational invariance, and residual mapping. These pieces of prior knowledge reduce the hypothesis space and make learning feasible. This is highly consistent with the cognitive science principle that "perception is guided by prior knowledge": the human visual system does not interpret images from scratch, but rather relies on pre-existing cognition of spatial structures, local patterns, and transformation invariance. Through its architectural design, the CNN formalizes such prior knowledge into structural design and mathematical computation, thereby substantially boosting the model's learning efficiency and generalization ability (Table 1).

Table 1. Priors associated with the evolution of CNN.

Era	Key Contribution	Relevance to Prior Knowledge
1958	Hubel & Wiesel—simple/complex cells in the visual cortex	Biological inspiration for hierarchical processing
1970s	Marr's computational theory of vision	Formalized hierarchical stages (edge → shape)
1980	Fukushima—Neocognitron (convolution + pooling)	First CNN-like architecture; it lacked back-propagation
1998	LeCun—LeNet-5 (conv-pool-FC) + back-propagation	Demonstrated trainable hierarchical feature extraction
2012	Krizhevsky et al.—AlexNet (ReLU, dropout, GPU)	Introduced modern regularization and optimization priors

4. Prior Knowledge in Models and Their Functional Mechanisms

4.1. Approaches to Encoding Prior Knowledge in Models

In visual models, prior knowledge is not an abstract notion; it is concretely realized through model-architecture design, data-organization strategies, and objective-function constraints.

At the architectural level, CNNs hardcode locality and translation equivariance into their network backbone. A convolutional kernel performs multiply-accumulate operations only within a fixed-size receptive field, which inherently assumes that the basic features of an image are jointly determined by adjacent pixels. The sharing of weights of the same convolutional kernel across the entire feature map hardcodes the translation invariance of images into the network, eliminating the need for the model to re-learn identical edges or textures at every spatial location. Pooling or strided convolution further aggregates local features into larger-scale representations—a process that implicitly incorporates the a priori that visual information exhibits a spatial hierarchical structure, where lower layers capture fine-grained details and higher layers encode global shapes. Architectural designs such as residual connections, pyramid feature fusion, and dilated convolution expand the receptive field while preserving local details, enabling the network to balance local precision and global contextual information at the same hierarchical level. For Transformer-based visual models, the a priori is implemented by integrating convolution or local window attention into the self-attention module, injecting relative positional biases into attention scores, or downsampling patches

at each stage. These modifications endow the originally global and position-agnostic attention mechanism with constraints related to translation, scale, and hierarchical representation. The role of architectural a priori is to significantly reduce the degrees of freedom that need to be learned, allowing the model to converge rapidly even with limited data and maintain robustness against common transformations such as translation, scaling, and rotation.

At the data level, priors are manifested through the processing methods applied to training samples. The most straightforward approach is data augmentation: performing operations such as random cropping, horizontal flipping, rotation, scaling, color jittering, random masking, CutMix, and MixUp on images. These transformations directly encode assumptions about symmetry, illumination invariance, and local occlusion into the input of each training iteration, forcing the model to learn features that remain consistent despite such perturbations. Higher-level priors are embodied in large-scale pre-training and self-supervised learning. By conducting supervised learning, masked reconstruction, or contrastive learning on datasets like ImageNet or massive unlabeled image corpora, the model absorbs universal visual concepts (e.g., edges, textures, and shapes) during the pre-training phase. These concepts are shared across most downstream tasks, essentially injecting the statistical regularities common to natural images into the model in advance. Domain randomization and style transfer encode the prior of cross-domain invariance into the data generation pipeline, exposing the model to diverse shooting conditions, backgrounds, and textures during training, thereby enhancing its robustness to domain shift in real-world deployment. Priors at the data level significantly improve model generalization and reduce reliance on labeled data by enriching input diversity and exposing the model to universal visual structures upfront.

Regularization at the objective function level encodes priors into the optimization target of learning. Common techniques include L2 weight decay and Dropout, which respectively assume that network parameters should remain small in magnitude and avoid over-reliance on individual neurons, thereby suppressing overfitting. Label smoothing treats the label of each class as a soft distribution, implicitly embedding the prior that similarities exist between different categories. More specialized constraints include the consistency loss in contrastive learning, which requires that different augmented views to yield similar feature representations—essentially encoding viewpoint invariance and semantic consistency into the loss function. Perceptual loss or feature reconstruction loss ensures that generated images are close to the original ones in the high-level feature space, reflecting the human visual system's sensitivity to global structure and texture. For segmentation and detection tasks, regularization terms such as edge smoothness, shape priors, and IoU (Intersection over Union) regularization directly constrain the geometric properties of outputs, ensuring that predictions conform to the continuity and reasonable shapes of objects. Self-supervised tasks like masked reconstruction and jigsaw puzzle prediction take spatial integrity and local consistency

as learning objectives, enabling the model to capture structural regularities of images even without labeled data. By incorporating these priors as regularization terms into the objective function, the model is forced to satisfy additional constraints during optimization, thus maintaining strong performance with limited training data, converging faster, and exhibiting greater resilience to noise and perturbations.

Structural designs (e.g., convolutional or attention mechanisms) embed priors such as local receptive fields, translation equivariance, and hierarchical pyramids into the model's computation graph. Data-level strategies (e.g., augmentation, pre-training, and cross-domain synthesis) embody empirical priors—including symmetry, illumination invariance, and universal visual statistics—in the input distribution. Objective function regularization imposes priors on parameter spaces, feature spaces, and output spaces onto the learning process through weight constraints, consistency losses, and geometric constraints. These three components work in tandem, enabling visual models to maintain efficient, robust, and transferable performance in real-world scenarios characterized by scarce data, environmental variations, and diverse tasks.

4.2. Prior Knowledge in CNNs

In practical applications, when constructing a simple convolutional neural network for handwritten character recognition, the first step is to determine the model architecture, including the number of convolutional layers, pooling layers, fully connected layers, and other components, as well as the specific parameter settings for each layer. The second step involves defining the training optimization strategies, along with the settings for training batches and training steps.

These parameter settings directly affect the model's structural complexity, learning capacity, and generalization performance. They play a crucial regulatory role in the model's training process and performance outcomes, and such parameters are referred to as hyperparameters, set by developers prior to model training. In contrast, model parameters (such as the weights and biases of a neural network) are generally initialized as random values and then automatically learned through training data and optimization algorithms.

Hyperparameters related to the model architecture determine the network's receptive field of the input space, the scale of translation invariance, and the capacity of feature representation; hyperparameters related to model training regulate the speed of model parameter updates, the strength of regularization, and the convergence behavior (Goodfellow, Bengio, & Haffner, 2016).

Hyperparameters are essentially a formal expression of task-specific or data-driven priors: in the task of handwritten character recognition, the size of convolution kernels is often set to 5×5, a choice that implicitly encodes the prior that strokes vary smoothly at the pixel scale and that local structures are sufficient to capture key features; setting the stride to 1 indicates that the model should maintain invariance to slight translations of characters in images; max-pooling per-

forms dimensionality reduction while preserving local extremal values, which is equivalent to imposing a smoothing prior on the overall contour of characters to suppress fine-grained noise (LeCun et al., 1998).

Convolution and pooling represent a powerful prior probability distribution, which reflects the assumptions we hold about the target and the set of models identified as reasonable through such assumptions. The prior probability distribution serves as a bridge connecting existing knowledge and observed data, and its core value lies in converting subjective cognition into quantifiable probabilistic forms to assist inference in uncertain environments. The selection of optimization algorithms, the setting of training batches, and the configuration of learning rates during the training process represent another type of prior, reflecting our pre-existing understanding of the model parameter space, the shape of the loss function, and the statistical characteristics of the training data. Such understanding is directly encoded into the network's learning process through the form of algorithms and settings.

The ReLU activation function shifts the prior toward sparsity and positive activation: only positive values are retained, while negative values are directly truncated to zero. This assumption aligns with the feature distribution of natural images—most meaningful features manifest as activation peaks after convolution, whereas responses at most positions can remain silent, thereby forming sparse activation maps in high-dimensional space. Sparsity not only reduces computational costs but also ensures that the gradient remains at 1 during forward propagation, preventing gradient vanishing. This embodies a prior that "gradients should maintain favorable flow" throughout the optimization process.

Regularization techniques also constitute an implementation of prior knowledge. L2 weight decay encodes the smoothness prior that "network parameters should remain small in magnitude to avoid over-reliance on individual features" into the loss function, thereby suppressing overfitting. Dropout, by randomly deactivating some neurons during training, is equivalent to averaging sub-models of the network and implicitly incorporates a redundancy prior that "different feature subsets should be capable of completing tasks independently". Batch Normalization standardizes features at each layer, enforcing consistency of feature distributions across different mini-batches. This is equivalent to imposing a constraint based on the prior that "the statistical characteristics of features should remain stable throughout the training process".

4.3. Prior Knowledge in ViT

ViT was originally designed to directly split an image into fixed-size patches and feed each patch into a global self-attention mechanism for modeling. Without an explicit convolutional structure, this approach makes the model prone to overfitting on small datasets and unable to effectively capture local spatial patterns. Subsequent studies have therefore incorporated various visual priors into ViT at different levels, enabling it to retain the expressive power of global self-attention

while also inheriting the advantages of convolutional neural networks, such as local receptive fields and hierarchical architectures.

For instance, the InPK (Infusing Prior Knowledge) network structurally embeds category-specific prior knowledge into learnable prompts, thereby imposing explicit constraints on category-specific knowledge and enabling multi-level enhanced interactions. This mechanism can guide attention to focus on semantically relevant regions and suppress irrelevant noise; meanwhile, by dynamically strengthening the coupling between tokens and prior knowledge, it effectively alleviates the dilution of category information during multi-round encoding, accelerates the model's optimization convergence, and further overcomes the problems of overfitting and insufficient generalization ability that are prone to occur in traditional methods (Zhou et al., 2025).

The SPAN (Spatial Prior Attention) framework explicitly injects spatial priors and semantic priors into the attention mechanism of Vision Transformers. It imposes prior regularization on a subset of attention heads, biasing their attention distributions toward task-relevant spatial regions; meanwhile, it retains a certain proportion of unconstrained heads to maintain the model's versatility and generalization ability. Through the synergistic effect of specificity and generalization, SPAN can significantly improve the model's robustness and enhance the interpretability of attention maps in data-scarce scenarios. In contrast, traditional self-attention relies entirely on data-driven similarity calculations and lacks any explicit domain knowledge constraints. Therefore, when the target is off-center or the background noise is substantial, the attention tends to exhibit scattered or even meaningless distributions. Furthermore, all heads in conventional multi-head attention learn general features in parallel without clear functional division, resulting in difficulties in accurately capturing specific semantic components (Miao et al., 2022).

The patch embedding in ViT performs the function of convolutional kernels in CNNs and conducts downsampling with a specific stride. In this way, each patch integrates local patterns of adjacent pixels, yielding a receptive field similar to that of CNNs. When computing attention, a local window constraint is imposed—the attention scope is restricted to a fixed spatial window where each patch is located, and then information flow across windows is achieved through window shifting or cross-window interaction. This window-based attention not only improves the computational efficiency of the model but also makes the attention weights more consistent with the human visual system's focus on adjacent regions.

The hierarchical architecture of ViT is also an important component of visual priors. Natural images tend to exhibit a feature hierarchy from fine-grained to coarse-grained details, and CNNs achieve this pyramidal feature abstraction through layer-wise pooling or strided convolution. To obtain similar hierarchical representations, ViT is designed with a multi-stage patch sampling and aggregation mechanism. A typical practice is to perform downsampling or pooling on patches at the end of each stage, merging several adjacent patches into a coarser

patch, which is then fed into the self-attention module for computation at a larger spatial scale. This hierarchical patch structure enables the model to capture fine-grained details at lower layers and global semantic information at higher layers, with computational complexity decreasing progressively across layers, thus significantly improving efficiency.

ViT also incorporates prior knowledge through training strategies. Large-scale supervised pre-training itself constitutes a form of empirical prior: it allows the model to learn universal visual concepts on massive labeled datasets, which can then be fine-tuned for downstream tasks, thereby overcoming the sample scarcity issue inherent in training from scratch. Self-supervised methods (e.g., MAE, DINO) design masking or contrastive learning objectives to force the model to recover missing information or align features in the absence of explicit labels. This is equivalent to imposing the priors of intra-image consistency and cross-view invariance onto the learning process. Furthermore, knowledge distillation leverages soft labels from large teacher networks as additional supervision signals, introducing high-level semantic priors. This enables lightweight ViT variants to retain computational efficiency while acquiring richer category-aware information.

By integrating empirical visual priors into multiple components—including input embedding, attention mechanisms, positional encoding, hierarchical architecture, and training objectives—ViT equips the pure Transformer, which originally lacks inductive bias, with enhanced adaptability to the statistical properties of natural images, achieving a favorable balance between expressive power and data efficiency. For example, the ViTAE/ViTAEv2 framework injects the inherent inductive bias of CNNs into Vision Transformers as prior knowledge, addressing the limitations of traditional ViT models such as insufficient locality, lack of scale invariance, and poor data efficiency (Zhang, Xu, Zhang, & Tao, 2022).

5. Limitations and Potential Risks of Priors

5.1. Risks of Subjectivity and Challenges in Formalization

Priors encode human experience of the visual world into models, enabling them to rapidly capture fundamental regularities such as edges, textures, and translation invariance when data is scarce. However, these experiences themselves tend to be subjective, and this subjectivity differs from architectural inductive bias and dataset bias. Architectural inductive bias is an implicit, structural tendency that reflects the model's characteristics at the perceptual level; dataset bias is an external, data-driven deviation that reflects the limitations of the empirical distribution.

Subjective risks arise from the very source of the priors we embed in a model. When researchers decide, for instance, on the size of convolutional kernels, the geometry of attention windows, or the set of data-augmentation operations, they are implicitly assuming that natural images obey certain regularities—such as local smoothness, translation equivariance, and approximately Gaussian color distributions—based on their intuition or the statistical characteristics of existing datasets (LeCun, Bengio, & Hinton, 2015). These assumptions are indeed effective

for common data types like natural photography and urban street scenes, but they may not hold for non-standard visual domains such as medical imaging, remote sensing imagery, or cultural symbols. If a model is rigidly constrained by priors that do not align with reality, anomalous patterns will be misclassified as noise, leading to systematic errors. More importantly, priors often embody the cultural, aesthetic, and ethical biases of annotators or data collectors—for instance, the binary classification of gender and skin tone in facial attribute annotation (Torralba & Efros, 2011). When such subjective priors are hardcoded into a model's architecture or loss function, the model's judgments will replicate and amplify these inherent biases, resulting in fairness issues and diminished generalization capabilities across cross-domain and cross-cultural scenarios.

The challenge of formalization manifests in the process of translating abstract priors into differentiable, optimizable mathematical constraints. Many priors are inherently high-level semantic descriptions—for example, "objects should maintain complete geometric shapes" and "illumination in a scene should adhere to physical consistency"—concepts that cannot be fully captured by a single regularization term or explicit functional form. Using smoothness constraints or local consistency losses to approximate these priors often only captures their coarse-grained features, while fine-grained structural information is weakened or lost during optimization. Excessively strong regularization severely restricts the model's degrees of freedom, leading to underfitting; conversely, overly weak regularization renders the prior constraints ineffective, leaving the model vulnerable to overfitting on noisy or biased data. The approximation process inevitably dilutes the original intent of priors, reducing them to crude information that the model can only partially perceive. Overly strict constraints overly compress the model's degrees of freedom, potentially causing underfitting; conversely, insufficiently strict constraints become almost irrelevant, defeating the purpose of incorporating priors in the first place. Furthermore, different priors may conflict with one another: local smoothness tends to eliminate fine details, whereas edge preservation requires retaining high-frequency variations. Balancing such conflicts within a unified loss function often relies on empirical coefficient tuning, with no universally accepted theoretical guidelines to follow.

The risk of subjectivity causes priors to hardcode human biases, domain-specific assumptions, and cultural limitations into models, impairing their cognitive performance in atypical scenarios or those with stringent fairness requirements. The challenge of formalization, on the other hand, limits the ability to fully and accurately translate complex, abstract priors into trainable constraints, leading priors to either be diluted or to conflict with learning objectives. This in turn restricts the model's capacity to independently discover new regularities.

5.2. Boundary Constraints on Cognitive Abilities

As the inductive bias of a model, the a priori not only helps neural networks rapidly capture statistical regularities but also delineates the boundaries of the model's

cognitive capabilities. Because priors are often pre-hardcoded into network architectures, loss functions, or initialization distributions based on existing datasets or theoretical assumptions, they act as strong constraints during the training phase (LeCun, Bengio, & Hinton, 2015). However, such constraints themselves lack the ability to self-adjust in response to environmental changes. In other words, priors compress the model's search space into a relatively closed subset, leading to cognitive rigidity when faced with new scenarios. When the input distribution of a task drifts continuously over time—for example, the rapid changes in illumination, viewpoint, or object morphology encountered in robotic vision—the model can often only perform fine-tuning within the scope of local translation or local smoothness covered by the priors, rather than reconstructing its internal representations at a higher level (Lake et al., 2017). The static nature of priors fundamentally limits the model's ability to respond instantaneously to environmental changes, making it difficult to maintain continuity and robustness in the cognitive process.

Another layer of boundary constraints imposed by priors is reflected in the capacity for creative generation. Innovative attempts by generative models in fields such as art, text, or design are often described as "combinations within the prior space", meaning that models can only explore combinations of patterns they have already learned (Elgammal et al., 2017). When priors are overly dominant, the network tends to replicate existing textures, structures, or concepts, resulting in outputs that lack genuine discreteness and unexpectedness—a stark contrast to human divergent thinking (Marcus, 2018). This limitation in creativity not only undermines the value of models in high-level tasks such as art and design, but also restricts their potential to propose groundbreaking hypotheses in scenarios involving scientific exploration or technological innovation.

The boundary constraints of priors on the model's cognitive capabilities are reflected in two aspects: first, the lack of self-regulatory ability in response to dynamic environmental changes leads to unstable performance in situations involving distribution shifts or continuous learning; second, the strong constraints on the generation space suppress true innovative potential, rendering the model more of a statistical imitator than a creative thinker.

6. Conclusion

Fundamentally, a machine-learning model extracts regularities from data. When the available data are scarce, noisy, or the task is intrinsically complex, relying solely on data often leads to over-fitting and slow convergence. In such regimes, prior knowledge provides a principled remedy. Prior knowledge can be incorporated into models in three canonical ways.

Structural priors encode assumptions about the geometry of the input space. For example, the observation that object boundaries correspond to abrupt luminance changes motivates the design of convolutional layers in a Convolutional Neural Network (CNN). By embedding this edge-detecting prior, the network is

encouraged to focus on edge features, which are later combined to form higher-level shape representations.

Parameter priors express beliefs about the distribution of model parameters. Regularization techniques instantiate this type of prior: penalizing large weights encodes the assumption that model parameters should remain modest, thereby discouraging overly complex solutions that would otherwise fit noise.

Pre-trained priors capture knowledge acquired during large-scale pre-training and transferred to downstream tasks. Large language models (LLMs) exemplify this category: pre-training on massive text corpora endows the model with linguistic priors—grammar, semantic relations, and world knowledge—so that fine-tuning on a modest amount of task-specific data yields high accuracy with minimal training effort.

By explicitly embedding these structural, parameter, and pre-trained priors, machine-learning systems become more data-efficient, robust, and capable of tackling challenging problems that would be intractable with data alone. The logical necessity of prior knowledge is reflected in the following: once we treat certain input symbols as a means of specifying or identifying a function, we encounter a significant challenge—the set of potential input-output functions is so vast that we cannot identify them using a finite set of instructions or any other finite method. The collection of functions is not only infinite but also uncountable (Pylyshyn, 1984). The practical significance of prior knowledge lies in how to solve complex problems under various resource constraints. Solutions that disregard resource limitations—such as setting no time targets or providing unlimited computing power—are impractical in real-world scenarios, even if they are theoretically feasible. The role of prior knowledge in machine cognition is analogous to that of innate structures in philosophy: it does not provide specific sensory content, but determines how the system organizes, interprets, and utilizes such content. Without prior knowledge, machines can only perform blind fitting relying on massive datasets; with prior knowledge, machines can demonstrate higher efficiency, robustness, and intelligence in tasks involving scarce data, complex environments, and the need for causal explanations.

Prior knowledge is both a boon and a constraint to the cognitive capabilities of models. The risk of subjectivity causes models to hardcode human biases and limitations at the perceptual level, making them prone to failure in cross-domain, cross-cultural, or non-natural data scenarios. The challenge of formalization makes it difficult to fully and accurately translate complex, abstract priors into trainable constraints, leading to priors that are either weakened or in conflict with learning objectives.

The boundary constraints of priors on the cognitive capabilities of models and the resulting problem of cognitive rigidity can be significantly alleviated by introducing new research methods. Neuro-symbolic artificial intelligence provides models with means for cross-domain transfer, knowledge representation, and reasoning through explicit logical constraints and editable rule systems (Colelough

& Regli, 2025). Causal reasoning, by capturing underlying generative mechanisms and supporting intervention and counterfactual reasoning, lays a theoretical foundation for the cross-environment robustness and mechanistic interpretability of models (Pearl, 2009). The integrated neuro-symbolic causal system enables synergistic evolution across three dimensions—perception, symbolic reasoning, and causal inference. This breakthrough can transcend the cognitive boundaries of current artificial intelligence in open-world scenarios, serving as a pivotal pathway to overcoming boundary constraints and cognitive rigidity.

The epistemological duality of prior knowledge is manifested in two aspects: it is both a necessary condition for knowledge generation—without prior knowledge, experience cannot be elevated to universal knowledge—and a source of cognitive limitations, as the presuppositional nature of prior knowledge defines the boundaries of knowledge. This reflects the initiative of cognition: the subject reconstructs experience through prior knowledge. It also reveals the historicity of cognition: prior knowledge evolves and is revised as experience accumulates. Prior knowledge provides the starting point and constraints for the subject's cognition, while learning maps new information into existing prior knowledge through assimilation and adjusts prior knowledge to the extent that it can explain new information. These two processes are perpetually intertwined and mutually reinforcing, forming the fundamental driving force behind cognitive development.

Acknowledgements

The author would like to express gratitude to all the reviewers for their meticulous scrutiny and valuable suggestions and also extends sincere thanks to the editorial staff of this paper.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Anderson, J. R. (1983). *The Architecture of Cognition*. Harvard University Press.

Cameron, J. B. (2023). *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence*. Oxford University Press.

Clark, A. (2013). Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science. *Behavioral and Brain Sciences, 36,* 181-204. https://doi.org/10.1017/s0140525x12000477

Colelough, B. C., & Regli, W. (2025). *Neuro-Symbolic AI in 2024: A Systematic Review*. https://arxiv.org/abs/2501.05435

Dong, C. Y. (2023). The Nature of AI and Its Limits in the Light of the Opacity of Machine Cognition. *Social Sciences in China, 5,* 148.

Elgammal, A., Liu, B. C., Elhoseiny, M., & Mazzone, M. (2017). *CAN: Creative Adversarial Networks, Generating "Art" by Learning about Styles and Deviating from Style Norms*. https://arxiv.org/abs/1706.07068

Friston, K. (2010). The Free-Energy Principle: A Unified Brain Theory? *Nature Reviews Neuroscience, 11,* 127-138. https://doi.org/10.1038/nrn2787

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Kant, I. (1998). *Critique of Pure Reason*. Cambridge University Press.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building Machines That Learn and Think Like People. *Behavioral and Brain Sciences, 40,* e253. https://doi.org/10.1017/s0140525x16001837

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature, 521,* 436-444. https://doi.org/10.1038/nature14539

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based Learning Applied to Document Recognition. *Proceedings of the IEEE, 86,* 2278-2324. https://doi.org/10.1109/5.726791

Marcus, G. (2018). *Deep Learning: A Critical Appraisal*. https://arxiv.org/abs/1801.00631

Miao, K., Gokul, A., Singh, R., Petryk, S. et al. (2022). *Prior Knowledge-Guided Attention in Self-Supervised Vision Transformers*. https://arxiv.org/abs/2209.03745

Pearl, J. (2009). *Causality*. Cambridge University Press. https://doi.org/10.1017/cbo9780511803161

Piaget, J. (1970). *Structuralism*. Basic Books.

Pylyshyn, Z. W. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. MIT Press.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects. *Nature Neuroscience, 2,* 79-87. https://doi.org/10.1038/4580

Schlicht, T. (2022). Minds, Brains, and Deep Learning: The Development of Cognitive Science through the Lens of Kant's Approach to Cognition. In *Kant and Artificial Intelligence* (pp. 1-38). De Gruyter. https://doi.org/10.1515/9783110706611-001

Torralba, A., & Efros, A. (2011). Unbiased Look at Dataset Bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1521-1528). IEEE. https://doi.org/10.1109/CVPR.2011.5995347

Zhang, Q. M., Xu, Y. F., Zhang, J., & Tao, D. C. (2022). *ViTAEv2: Vision Transformer Advanced by Exploring Inductive Bias for Image Recognition and beyond*. https://arxiv.org/abs/2202.10108

Zhou, S. C., Wei, J. W., He, S. Y., Zhou, Y. Y. et al. (2025). *InPK: Infusing Prior Knowledge into Prompt for Vision-Language Models*. https://arxiv.org/abs/2502.19777