RESEARCH · PERCEPTION & LEARNING

Structure Before Meaning

Giving a cognitive substrate sight — a single visual faculty that perceives structure by algorithm and learns meaning by induction

Stefan Ragland

Dominion Labs Research & Development

Dominion Labs

research@dmnlabs.org

May 20, 2025

Abstract

Vision is usually treated as one capability, and in the current practice it is bought whole: an image is passed to a network trained on millions of labelled pictures, which returns names. We take a different view. Seeing decomposes into two quite different acts — *perceiving structure* (there is a large rounded region here, moving, brighter than its surround) and *recognising a concept* (that region is a face). The first act has been solved by algorithms for two decades, with no learning at all[2][3][4]: segmentation, region and object proposals, saliency, local-feature description and instance matching, optical flow. The second act — naming a novel category — is the only place a learned model is genuinely required, and, tellingly, every recent "training-free" recogniser supplies it not by any new algorithm but by leaning on a pretrained perceptual backbone[11][12][13][14]. Our thesis is that a cognitive substrate which *already* learns concepts by induction — from a handful of examples, over a knowledge graph it reasons on, holding graded beliefs and abstaining when it cannot ground an answer — can supply the second half of seeing itself, occupying the role the backbone plays, with no neural network anywhere in the loop. We describe the architecture and, as its centre, a methodological commitment: vision is built into the substrate as a *faculty* — one entry point for all sight, the way the substrate already has one reader that turns a sentence into propositions and one voice that renders its state into words. The faculty perceives structure with classical vision, admits what it sees through the substrate's single knowledge-admission door as typed, provenance-bearing observations, and lets the substrate learn what those structures *mean* exactly as it learns everything else. Structure is perceived; meaning is learned; the two are kept separate, and that separation is what makes machine sight legible.

1 Two acts, not one

Ask what it takes to see a person walk through a doorway and you are really asking two questions with different answers. The first is geometric and photometric: *where* in the frame is there a coherent thing, how big is it, what is its shape, what are its colours, is it moving and how. The second is semantic: *what* is that thing — a person, a door, a shadow. David Marr made the case forty years ago that vision is layered in exactly this way, that a system recovers structure before it assigns meaning, and that conflating the layers is a mistake of method, not just of engineering[1].

The modern practice conflates them on purpose. A convolutional network[10] or a vision transformer is handed raw pixels and returns names, having folded perception and recognition into one trained mapping. This works, and it hides the seam. But the seam is real, and where it lies decides what actually needs a model. Almost everything on the *structure* side of the seam can be computed directly from an image, deterministically, with no training. Only the *meaning* side — putting a name to a category the system has never been told about — genuinely requires a learner. The interesting question, for a system that is not built on a trained perceptual backbone, is whether it can be the learner for that second half itself.

> *Vision is not one capability to be acquired whole. It is the composition of a perception that algorithms already perform and a recognition that must be learned — and the recognition can be learned by the substrate that will reason over it, rather than bought from a network trained elsewhere.*

2 What algorithms already see

The structure side of vision is not a frontier; it is a mature body of work, and it is worth being concrete about how much of "seeing" it delivers with no learning in the loop.

- **Regions and segments.** Efficient graph-based segmentation groups pixels into coherent regions from colour and contrast alone[2]; superpixel methods do the same at controllable granularity. The image becomes a small set of parts, each with a boundary.
- **Object-like blobs.** This is the capability easiest to underestimate. Maximally stable extremal regions detect blobs that persist across thresholds[5]; selective search hierarchically groups segments into *object proposals* — "here are the thing-shaped regions worth attending to" — explicitly *without any learning procedure*, on colour, texture, size and shape cues[3]. Objectness measures rank windows by how object-like they are using edge density, saliency, superpixel straddling and colour contrast[6]. Together these answer *how many* distinct things are present, *where*, and *how large* — the very "blob capabilities" one might assume needed a network.

- **Salience.** Center–surround contrast picks out the region a viewer's eye would go to first, again from low-level cues [6].
- **Shape and geometry.** Contours, convexity, aspect ratio and moment invariants [7] describe a region's form; the Hough transform recovers lines and circles. A region can be called round, elongated, quadrilateral — described, not merely located.
- **Instance identity.** Scale-invariant local features [8] and their fast binary successors [9] describe keypoints so distinctively that matching them against a reference image recognises a *specific* object, logo or face across viewpoint and lighting. This is genuine recognition — of known instances — with no training, only matching.
- **Motion.** Optical flow [15], frame differencing and background subtraction give per-pixel and per-region motion; tracking follows a region through a clip. For video, "something moved, here, this fast, and this is the same thing as the previous frame" is algorithmic.
- **Encoded content.** Where a scene literally contains encoded symbols — a QR code, a barcode, a fiducial marker — the exact payload is decoded deterministically. This is semantic content read from pixels with perfect fidelity and no model at all.

The one classical recogniser that crosses into semantics — the boosted cascade that detects faces [16] — does so precisely by *training* a classifier on labelled examples, which is the exception that proves where the seam lies. Everything above it in this list needs no examples; the moment a novel category must be named, learning enters.

3 The wall, and how the field gets past it

So there is a wall, and it is worth stating precisely rather than vaguely. Classical vision will tell you there is a coherent, salient, roughly round region in the upper-centre of the frame, skin-toned, moving leftward, textured like the reference image of a particular person. What it will not do, unprompted, is emit the word *face* for a category it has never been shown. Assigning an open-ended name to a novel visual category is the part that requires a learner.

The instructive fact is *how* the recent literature clears this wall. A wave of "training-free" open-vocabulary recognisers has appeared, and the phrase is easy to misread. They are training-free only in the sense that they add *no new fine-tuning*. The semantics they emit comes entirely from a *pretrained perceptual backbone*: methods that assign names to regions do so by harnessing vision foundation models — CLIP for image–text alignment [11], DINO for emergent object structure [12], SAM for class-agnostic masks [13] — stitched together at inference [14]. Strip the backbones out and the meaning goes with them; what remains is exactly the structure-level machinery of §2. The honest reading of the state of the art is that there is no algorithm that assigns open-category names without a model that learned those names from data somewhere. The meaning is always learned. The only question is *by what*, and *where*.

> *Every current recogniser fills the semantic half of vision with a network trained on labelled images. The substrate can fill it instead — with the same faculty it already uses to learn concepts from a few examples — and thereby own the whole of sight without a perceptual model in the loop.*

4 The substrate already supplies the missing half

The reason this is not wishful is that the components the semantic half of vision needs are not new capabilities to be invented; they are faculties the substrate already runs, for language and for action. Naming a perceived structure is a learning-and-reasoning problem, and the substrate is a learning-and-reasoning system.

- **It induces general concepts from few examples.** The substrate learns relational rules from a handful of demonstrations, and its learned competence is localised to the rules themselves — established, not asserted. Learning "a region with *this* signature is a *door*" from a few labelled regions is the same act as learning a relational action rule from a few transitions.
- **It holds a structured concept graph and reasons over it.** Concepts sit in an is-a taxonomy the substrate walks to derive consequences. A recognised *poodle* inherits *dog*, *mammal*, *animal* with no extra teaching — recognition immediately becomes inference.
- **It holds graded, revisable beliefs and abstains when ungrounded.** Every claim the substrate holds is a posterior it can revise, and it treats *"I do not know"* as a first-class answer rather than guessing. A recognition can therefore be held at a confidence, strengthened by corroboration, and withheld when the evidence is thin — the discipline that separates a reliable knower from a fluent one.
- **It admits knowledge through one provenance-gated door.** New knowledge — taught, researched, or perceived — enters through a single admission point that records where each fact came from and types the values it carries (a number is held as a quantity, a date as a point in time). A perceived observation is admissible evidence like any other, and carries the fact that it came from sight.

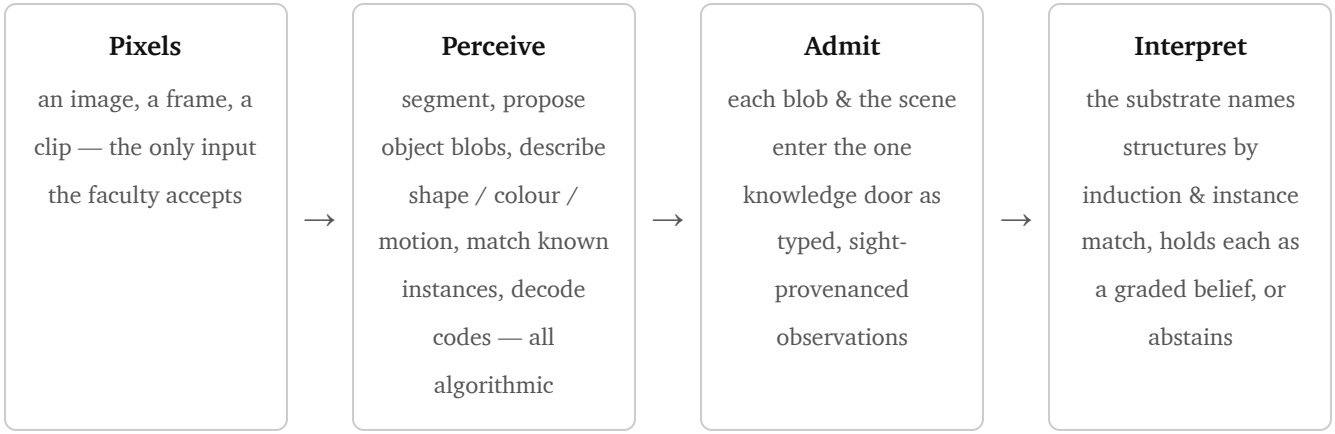
These are the exact ingredients a recogniser needs: a learner to map signatures to categories, a place to put the result so it participates in inference, a way to hold it as revisable belief, and a record of where it came from. The substrate has them already. What it lacks is a way for *pixels* to reach them.

5 Methodology: vision as a faculty of the substrate

The way pixels should reach the substrate is not as one more caller that writes observations, but as a *faculty* — a first-class organ of the substrate with a single entry point, the sole route by which visual input becomes knowledge. This is the substrate's governing pattern, applied to sight. The substrate does not have many ad-hoc ways to read text; it has *one reader*, which turns a sentence into propositions, and every path that consumes language goes through it. It does not have scattered ways to speak; it has *one voice*, which renders its

internal state into words. Vision is built the same way: **one faculty, the single entry point for all sight**, so that "what the substrate has seen" has one authoritative answer and one place it is produced — never a second, private path that admits pixels behind the faculty's back.

The parallel to the reader is exact and worth drawing out, because it is the whole method. The reader does not *understand* a sentence by an inscrutable mapping; it *derives* a proposition from a token sequence by a procedure the substrate can exhibit, and the meanings of the words are *learned*, held in the same store as everything else. The vision faculty does not *understand* an image by an inscrutable mapping; it *derives* observations from a pixel array by algorithm, and the meanings of the structures are learned, held in the same store. Same shape, different modality. Sight is to pixels what reading is to text.



The vision faculty: one entry point for sight. Structure is perceived on the left; meaning is learned on the right; the substrate's existing admission, belief and induction faculties do the second half.

The faculty has three internal stages, and the discipline is in which stage is allowed to do what.

1. **Perceive — structure, by algorithm.** The faculty runs the classical pipeline of §2 over the real pixels: it segments the image, proposes object-like blobs, and for each computes a *signature* — position in the frame, size, shape class, dominant colours, texture and keypoint density, and, for video, motion and track identity. It matches local features against a library of known instances, and decodes any encoded symbols. This stage invents nothing; every value is measured from the image. It does not assign a category.
2. **Admit — observations, through the one door.** Each blob becomes an observation and enters through the substrate's single admission point, tagged with sight as its source and with its measured values typed (a width is a quantity, a capture time a point in time). Because admission is the same door taught facts use, a perceived observation fans out to belief and to the relevant domain exactly as a taught fact does, and it is corroborable, revisable, and traceable to the frame it came from. The faculty never writes to the concept store directly; it observes, and admission decides.

3. **Interpret — meaning, by the substrate.** Naming is not the faculty's to do; it is the substrate's. An instance match yields a name immediately — this is the reference object it was shown. A novel category is learned: shown a few labelled regions, the substrate induces a rule from signature to concept, and thereafter names a matching blob by derivation, inheriting the concept's place in the taxonomy. Each naming is held as a graded belief; where neither an instance match nor an induced rule grounds a name, the faculty reports the structure and the substrate says *unknown* — it does not guess a label. The honesty discipline that governs the substrate's answers governs its sight.

The consequence of building sight as a faculty rather than a pipeline is that vision inherits the substrate's whole constitution for free. It is grounded (a recognition traces to a signature plus a learned rule or an instance match), revisable (a belief, not a verdict), legible (one can ask *why* and get a chain, not an activation), and singular (one authority for what has been seen). And because the naming lives in the substrate, a thing seen and a thing read and a thing told are the same kind of knowledge: the substrate can be *told* what a new object is in words, and recognise it by sight afterwards, because both routes end at the same concept.

6 Why the integration is possible now

None of the second half of the faculty is speculative infrastructure. The substrate's admission door already accepts perception as a first-class, root-level kind of evidence — a named thing observed in a named condition — distinct from a taught fact only in its provenance. It already types the values an observation carries, so a measured width or a capture date is held as a quantity or a time rather than as a word. Admitted observations already fan out to graded belief and to domain formation through the one learning path. And the induction that would learn a signature-to-concept rule is the same induction the substrate already uses to learn action rules and to learn readings from example sentences. The faculty's first stage — the classical perception — is the genuinely new code, and it is a composition of algorithms that have existed for decades and ship in mature libraries. The second stage plugs into a door that is already open; the third stage is a faculty the substrate already runs. Integration is a matter of building the eye and wiring it to organs that already exist, not of inventing the organs.

7 Discussion

The claim of this paper is deliberately narrow and, we think, therefore strong. We do not claim that pixels can be named by algorithm alone — the literature is clear that they cannot, and that every system which appears to do so has a trained model inside it. We claim that *naming is the only part that must be learned*, that the rest of seeing is algorithmic, and that a substrate which already learns concepts is the right thing to do the naming — not a network trained on someone else's labels and frozen, but a learner that improves with what it is shown and reasons with what it recognises.

Built this way, sight is not a bolted-on classifier but an organ continuous with the rest of the mind. A recognised object is a concept, so it inherits, participates in inference, and can be reasoned about the instant it is seen. It is a belief, so it can be doubted, corroborated, and revised. It has a provenance, so the substrate can always say it knows this because it saw it, in this frame. And it is produced at one entry point, so there is a single, auditable answer to what the system has perceived. The separation of structure from meaning — perceiving first, naming second, by different faculties — is not merely a tidy decomposition. It is what lets a seeing machine tell you *why* it thinks it saw what it saw, and lets it say, without embarrassment, when it has seen a shape it cannot yet name. Structure before meaning is how sight becomes legible.

References

[1] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, 1982.

[2] P. F. Felzenszwalb, D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[3] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[4] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer, 2010.

[5] J. Matas, O. Chum, M. Urban, T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004.

[6] B. Alexe, T. Deselaers, V. Ferrari. Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2189–2202, 2012.

[7] M.-K. Hu. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, 8(2):179–187, 1962.

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[9] E. Rublee, V. Rabaud, K. Konolige, G. Bradski. ORB: an efficient alternative to SIFT or SURF. In *ICCV*, 2011.

[10] A. Krizhevsky, I. Sutskever, G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[11] A. Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[12] M. Caron et al. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

[13] A. Kirillov et al. Segment Anything. In *ICCV*, 2023.

[14] Harnessing vision foundation models for high-performance, training-free open-vocabulary segmentation. *arXiv:2411.09219*, 2024.

[15] B. D. Lucas, T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 1981.

[16] P. Viola, M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[17] G. A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[18] A. d'Avila Garcez, L. C. Lamb. Neurosymbolic AI: the 3rd wave. *arXiv:2012.05876*, 2020.

Dominion Labs Research & Development · May 20, 2025. A thesis on giving the cognitive substrate sight: structure is perceived by algorithm, meaning is learned by the substrate, and vision is built as a single faculty — one entry point for all sight.