

RESEARCH ARTICLE

Representation Learning in Sensory Cortex: A Theory

FABIO ANSELMI^{1,2,3} AND TOMASO POGGIO¹¹Center for Brains, Minds and Machines, McGovern Institute, Massachusetts Institute of Technology, Cambridge, MA 02139, USA²Baylor College of Medicine, Houston, TX 77030, USA³University of Trieste, 34127 Trieste, Italy

Corresponding author: Fabio Anselmi (anselmi@mit.edu)

This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF Science and Technology Centers (STC) under Award CCF-1231216.

ABSTRACT We review and apply a computational theory based on the hypothesis that the feedforward path of the ventral stream in visual cortex's main function is the encoding of invariant representations of images. A key justification of the theory is provided by a result linking invariant representations to small sample complexity for image recognition - that is, invariant representations allow learning from very few labeled examples. The theory characterizes how an algorithm that can be implemented by a set of "simple" and "complex" cells - a "Hubel Wiesel module" - provides invariant and selective representations. The invariance can be learned in an unsupervised way from observed transformations. Our results show that an invariant representation implies several properties of the ventral stream organization, including the emergence of Gabor receptive fields and specialized areas. The theory requires two stages of processing: the first, consisting of retinotopic visual areas such as V1, V2 and V4 with generic neuronal tuning, leads to representations that are invariant to translation and scaling; the second, consisting of modules in IT (Inferior Temporal cortex), with class- and object-specific tuning, provides a representation for recognition with approximate invariance to class specific transformations, such as pose (of a body, of a face) and expression. In summary, our theory is that the ventral stream's main function is to implement the unsupervised learning of "good" representations that reduce the sample complexity of the final supervised learning stage.


INDEX TERMS Visual cortex, Hubel Wiesel model, simple and complex cells, artificial neural networks, invariance, sample complexity.

I. INTRO AND BACKGROUND

The ventral visual stream is believed to underlie object recognition abilities in primates. Fifty years of modeling efforts, beginning with the original Hubel and Wiesel proposal (HW in the rest of the paper) of a hierarchical architecture iterating in different layers the motif of simple and complex cells in V1, have led to a series of quantitative models from Fukushima [36] and Riesenhuber and Poggio [106] HMAX (Hierarchical architecture with MAX pooling) to more recent architectures based on contrastive [20], [139] or slow features [109] learning. These models are increasingly faithful to biological architecture constraints and are able to mimic

properties of cells in different visual areas while achieving human-like recognition performance under restricted conditions. Starting from the architectures in [45], [112], and [138], deep learning convolutional networks, which are hierarchical but otherwise do not respect the ventral stream architecture and physiology, have been trained with very large labeled datasets. The resulting model's neuron population accurately mimic the object recognition performance of the macaque visual cortex (e.g., [17], [59], [132], [133], and [139]). However, the nature of the computations carried out in the ventral stream is not fully explained by such models that, despite being simulated on a computer, remain rather opaque.

In other papers (in particular [6], [7], [10], and [102]) we have developed a mathematics of invariance that can be applied to the ventral stream. Invariance and equivariance

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Xiang .

are key features of modern neural networks architectures (e.g., [4], [9], [16], [20], [23], [44], [79], and [80]). In this work we outline a comprehensive theory of the feedforward computation of invariant representations in the ventral stream - a theory of the first 100 milliseconds of visual perception, from the onset of an image to activation of IT (Inferior Temporal) neurons. In particular, such representations are likely to underlie rapid categorization – immediate object recognition from flashed images [104], [126]. We emphasize that this theory is not a full theory of vision that will explain top down effects and the role of backprojections, but only a precursor to it.

Our theory is based on the hypothesis that the main computational goal of the ventral stream is to compute neural representations of images that are invariant to transformations commonly encountered in the visual environment, and learned from unsupervised experience. The main novelty of our theory consists in explaining various aspects of the ventral stream architecture and its neurons in the light of this hypothesis and linking it with low sample complexity learning. Since invariant representations turn out to be “good” representation for supervised learning, characterized by small sample complexity, the architecture of the ventral stream may ultimately be dictated by the need to learn from very few labeled examples, similar to human learning but quite different from typical supervised machine learning algorithms trained on large sets of labeled examples.

We use our theory to compactly summarize and explain several key aspects of the neuroscience of visual recognition, while predict others. Our main contributions are:

- We introduce a novel general theoretical framework for a computational theory of invariance (section II), and a theory of the basic biophysical mechanisms and circuits in (section III). In particular, we compactly describe relevant recent mathematical results on invariant representations in vision whose details and proofs can be found in [6], [7], [8], and [10]. The starting point is a result proving that image representations (a feature vector that we call a signature) which are invariant to translation and scaling and approximately invariant to some other transformations (e.g., face expressions) *can considerably reduce the sample complexity of learning* (section IIA). We then describe how an invariant and unique (selective) image representation can be computed for each image or image patch; this invariance can be exact in the case of group transformations (we focus on groups such as the affine group in 2D and one of its subgroups, the similitude group consisting of translation and uniform scaling) and approximate under non-group transformations (sections IIB, IID). A module performing filtering and pooling, like the simple and complex cells described by Hubel and Wiesel (HW module), can compute such estimates. Each HW module provides a signature, for the part of the visual field that is inside its receptive field.

- We prove that Gabor functions are the optimal templates for maximizing simultaneous invariance to translation and scale (IIC). Hierarchies of HW modules retain their properties, while alleviating the problem of clutter in the recognition of wholes and parts, (sections IIE,IIF).
- We show that the same HW modules at high levels in the hierarchy are able to compute representations which are approximately invariant to a much broader range of transformations (e.g., 3D expression of a face, pose of a body, and viewpoint). They do so by using templates, reflected in neuron’s tuning, that are highly specific for each object class (sections IID, IIE).
- We describe (section III) how neuronal circuits may implement the operation required by the HW algorithm. We specifically discuss new models of simple and complex cells in V1 (sections IIIA, IIIB). We also introduce plausible biophysical mechanisms for tuning, pooling, and learning the wiring based on Hebbian-like unsupervised learning (sections IIIC, IIID, IIIE).
- We explain (section IV) how the final IT stage computes class-specific representations that are quasi-invariant to non-generic transformations. We also discuss the modular organization of anterior IT in terms of the theory; in particular, proposing an explanation of the architecture and of some puzzling properties of the face patches system.

We conclude the paper with a discussion of predictions to be tested and open problems (section V).

II. COMPUTATIONAL LEVEL: MATHEMATICS OF INVARIANCE

For this paper, we will use the following conceptual framework for primate vision:

- The first 100ms of vision in the ventral stream are mostly feed forward. The main computational goal is to generate a number of image representations, each one invariant or approximately invariant to some transformations experienced during development and at maturity, such as scaling, translation, and pose changes. The representations are used to answer basic questions about image type and what may be in it.
- The answers will often have low confidence, requiring an additional “verification/prediction step”, which may require a sequence of shifts of gaze and attentional changes. This step may rely on generative models and probabilistic inference and/or on top-down visual routines following memory access. Routines that can be synthesized on demand as a function of the visual task are needed to go beyond object classification.

We consider only the feedforward architecture of the ventral stream and its computational function. To help the reader to more easily understand the mathematics of this section, we give here an overview of the network of visual areas that we propose for computing invariant representations in feedforward visual recognition. There are two main stages: the first one computes a representation that is invariant

to affine transformations, followed by a second stage that computes approximate invariance to object specific, non-group transformations. The second stage consists of parallel pathways, each one for a different object class (see Figure 4 stage 2). The results of this section do not strictly require these two stages: the second one may not be present, in which case the output of the first stage directly accesses memory for classification. If both are present, as seems to be the case for the primate ventral stream, the mathematics of the theory requires that the object specific stage follows the one dealing with affine transformations. According to our theory, the HW module mentioned earlier is the basic module for both stages. The first and second stage pathways may consist of a single layer of HW modules. However, mitigation of interference by clutter requires a hierarchy of layers (possibly corresponding to visual areas such as V1, V2, V4, PIT (Visual 1,2,3 and Posterior Infero-Temporal area)) within the first stage. This may not be required in visual systems with lower resolution such as the mouse. The final architecture we use is shown in Figure 4: in the first stage computes representations that are increasingly invariant to translation and scale, while in the second stage a large number of class-specific parallel pathways induce approximate invariance to transformations that are specific for objects and classes. Notice that for any representation which is invariant to feature X and selective for feature Y, there may be a dual representation which is invariant to Y but selective for X. In general, they may both be needed for different tasks, and both can be computed by a HW module and the machinery computing them possibly shares a good deal of overlap. As an example, we would expect that different face patches in cortex are used to represent different combinations of invariance and selectivity.

A. INVARIANCE REDUCES SAMPLE COMPLEXITY OF LEARNING

Images of the same object usually differ from each other because of generic transformations such as translation or scale (distance), or more complex transformations such as viewpoint (rotation in depth) or change in pose (of a body) or expression (of a face) (see also [5], par 3.1.2 for a back of envelope estimation of the number of possible transformations of an image). In a machine learning context, invariance to image translations can be built up trivially by memorizing examples of the specific object in different positions. On the other hand, human vision is clearly invariant to novel objects seen only once: people do not have any problem recognizing a human face seen only once at different distances, e.g., in a distance-invariant way. It is rather intuitive that representations of images that are invariant to transformations such as scaling, illumination, and pose should allow supervised learning from far fewer examples.

This conjecture is supported by previous theoretical work showing that a significant portion of the complexity in recognition tasks is often due to the viewpoint and illumination nuisances that swamp the intrinsic characteristics of the

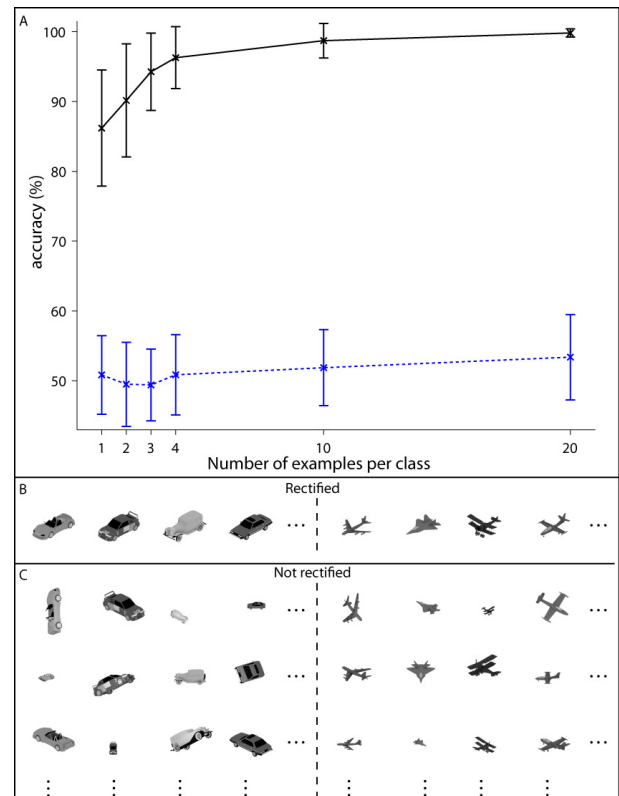


FIGURE 1. If an “oracle” factors out all transformations in images of many different cars and airplanes (C), providing “rectified” images (B) with respect to viewpoint, illumination, position and scale, the problem of categorizing cars vs airplanes becomes easy (A): it can be done accurately with very few labeled examples. In the figure (C), good performance (black line) was obtained from a single training image from each rectified class, using a linear classifier operating on pixels, whereas training from the unrectified training set yields chance performance. In other words, the sample complexity of the problem becomes much lower with a rectified (and therefore invariant) representation ([7]).

object [4], [74], [110]. This implies that in many cases, recognition (both identification, e.g., of a specific car relative to other cars, as well as categorization, e.g., distinguishing between cars and airplanes) would be much easier (only a small number of training examples would be needed to achieve a given level of performance) if the object images were rectified with respect to all transformations, or equivalently, if the image representations were invariant. The case of identification is obvious since the difficulty in recognizing exactly the same object, e.g., an individual face, is due solely to transformations. Figure 1 provides suggestive evidence from a classification task, showing that if an oracle factors out all transformations in images of many different cars and airplanes, providing “rectified” images with respect to viewpoint, illumination, position and scale, the problem of categorizing cars vs airplanes becomes easy: it can be done accurately with very few labeled examples. In this case, good performance was obtained from a single training image of each class, using a simple classifier. In other words, the sample complexity of the problem seems to be very low. A proof of the conjecture for the special case of translation

is provided in [7] for images defined on a grid of pixels and, with the main results restated below.

1) SAMPLE COMPLEXITY FOR TRANSLATION INVARIANCE

Consider a space of images of dimensions $p \times p$ which may appear in any position within a window of size $rp \times rp$. The natural image representation yields a sample complexity (for a linear classifier) of order $m_{image} = O(r^2 p^2)$; the invariant representation yields a sample complexity of order:

$$m_{inv} = O(p^2).$$

The result says that an invariant representation can considerably decrease the sample complexity – that is, the number of supervised examples necessary for a certain level of accuracy in classification. A heuristic rule corresponding to the result is that the sample complexity gain is on the order of the number of virtual examples generated by the action of the group on a single image (see also [2], [33], [94], and [119]). The result does not provide an algorithm but it supports the hypothesis that the ventral stream in visual cortex tries to approximate such an oracle. The next section describes a biologically plausible algorithm that the ventral stream may use to achieve this goal.

B. UNSUPERVISED LEARNING AND COMPUTATION OF AN INVARIANT SIGNATURE (ONE LAYER ARCHITECTURE)

The following HW algorithm is biologically plausible, as we will discuss in further detail in section II and III, where we argue that it may be implemented in cortex by a HW module. The module consists of a set of KH complex cells with the same receptive field, each pooling the output of a set of simple cells whose sets of synaptic weights correspond to one of the K “templates” of the algorithm and its transformations (which are also called templates) and whose output is filtered by a sigmoid function with a Δh threshold, $h = 1, \dots, H$.

HW algorithm for group transformations (see Figure 2)

- “Developmental” stage:
 - 1) Given one of the K isolated (on an empty background) objects (the “training set”), e.g., “templates”, memorize a sequence Γ of G frames corresponding to its transformations ($g_i, i = 1, \dots, |G|$) observed over a time interval (thus $\Gamma = g_0 t, g_1 t, \dots, g_{|G|} t$ for template t ; for template t^k the corresponding sequence of transformations is denoted Γ_k).
 - 2) Repeat for each of the K templates.
- “Run-time” computation of invariant signature for a single image (of any new object):
 - 1) For each Γ_k compute the dot product of the image with each of the $|G|$ transformations in Γ_k .
 - 2) For each k compute cumulative histogram of the resulting values.

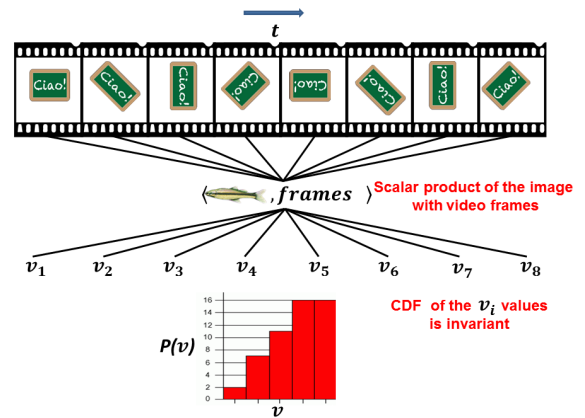


FIGURE 2. A graphical summary of the HW algorithm. The set of $\mu_h^k(I) = 1/|G| \sum_{i=1}^{|G|} \sigma(I, g_i t^k) + h\Delta$ values (eq. 1) in the main text correspond to the the histogram where $k=1$ denotes the template” green blackboard”, h the bins of the histogram, and the transformations are from the rotation group. Crucially, mechanisms capable of computing invariant representations under affine transformations can be learned (and maintained) in an unsupervised, automatic way just by storing sets of transformed templates which are unrelated to the object to be represented. In particular the templates could be random patterns.

- 3) The signature is the set of K cumulative histograms that is the set of:

$$\mu_h^k(I) = \frac{1}{|G|} \sum_{i=1}^{|G|} \sigma(I, g_i t^k) + h\Delta \quad (1)$$

where I is an image, σ is a threshold function, $\Delta > 0$ is the width of bin in the histogram, and $h = 1, \dots, H$ is the index of the bins of the histogram.

The algorithm consists of two parts: the first is unsupervised learning of transformations by storing transformed templates, which are “images”. This can be thought of as a “only once” stage, possibly done during development of the visual system. The second part is the actual computation of invariant signatures during visual perception.

This is the algorithm used throughout the paper. The guarantees we can provide depend on the type of transformations. The main questions are a) whether the signature is invariant under the same types of transformations that were observed in the first stage and b) whether it is selective, e.g., can it distinguish between N different objects. A summary of the main results of [5], [6], [7], [8], and [10] is that the HW algorithm is invariant and selective (for K in the order of $\log N$) if the transformations form a group. In this case, any set of randomly chosen templates will work for the first stage. Given that we are interested in transformations from a 2D image to a 2D image, the natural choice is the affine group consisting of translations, rotations in the image plane, scaling (possibly non-isotropic), and their compositions. The HW algorithm can learn with exact invariance and a desired selectivity level in the case of the affine group or its subgroups. In the case of 3D “images” consisting of voxels with x, y, z coordinates, rotations in 3D are also a group that

in principle can be dealt with, achieving exact invariance from generic templates by the HW algorithm (in practice this is rarely possible because of correspondence problems and self-occlusions). Later in section II.E we will show that the same HW algorithm provides approximate invariance (under some conditions) for non-group transformations such as the transformations from R^3 to R^2 induced by views of 3D rotations of an object.

In the case of compact groups the guarantees of invariance and selectivity are provided by the following two results (given informally here; detailed formulation in [5], [6], [7], [8], and [10]).

Result 1: Invariance

The distributions represented by equation 1 are invariant, that is each bin is invariant, e.g.,

$$\mu_h^k(I) = \mu_h^k(gI) \quad (2)$$

for any g in G , where G is the (locally compact¹) group of transformations labeled g_i in equation 1.

Result 2: Selectivity

For groups of transformations (such as the affine group), the distributions represented by equations 1) can achieve any desired selectivity for an image among N images in the sense that they can ϵ -approximate the true distance between each pair of the images (and any transform of them) with probability $1 - \delta$ provided that

$$K > \frac{c}{\epsilon^2} \ln \frac{N}{\delta} \quad (3)$$

where c is a universal constant.

The signature provided by the K cumulative histograms is a feature vector corresponding to the activity of the (HK) complex cells associated with the HW module. It is selective in the sense that it corresponds uniquely to an image of a specific object independently from its transformation. It should be noted that the robustness or stability of the signature under noisy measurements remains an interesting open problem in the theory. Because of the restricted dynamic range of cortical cells the number of bins H is likely to be small. It is important to remark that other, related representations are possible (see also [7], [62], and [65]). A cumulative distribution function (CDF) is fully represented by all its moments; often a few moments

$$\begin{aligned} \mu_{av}^k(I) &= \frac{1}{|G|} \sum_{i=1}^{|G|} \langle I, g_i t^k \rangle \\ \mu_{energy}^k(I) &= \frac{1}{|G|} \sum_{i=1}^{|G|} \langle I, g_i t^k \rangle^2 \\ \mu_{max}^k(I) &= \max_{g_i \in G} \langle I, g_i t^k \rangle \end{aligned} \quad (4)$$

¹A group is called compact if it is supported on a compact set. For example the rotation group is a compact group on the set of angles in $[0, 2\pi]$. Its is a locally compact group if it is supported on a locally compact set. For example the locally compact group of translations supported on the set of translations $[0, +\infty)$.

such as the average or the variance (energy model of complex cells, see [3]) or the max, can effectively replace the cumulative distribution function. Notice that any linear combination of the moments is also invariant and a small number of linear combinations is likely to be sufficiently selective. We will discuss implications of this remark for models of complex cells in the last section.

C. OPTIMAL TEMPLATES FOR SCALE AND POSITION INVARIANCE ARE GABOR FUNCTIONS

The previous results apply to all groups, in particular to those which are not compact but only locally compact such as translation and scaling. In this case it can be proved that invariance holds within an observable window of transformations [6], [7]. For the HW module the observable window corresponds to the receptive field of the complex cell (in space and scale). In order to maximize the range of invariance within the observable window, [6], [7] proves that the templates must be maximally sparse relative to generic input images (see below for definition of sparseness). In the case of translation and scale invariance, this sparsity requirement is equivalent to localization in space and spatial frequency, respectively: templates must be maximally localized for maximum range of invariance in order to minimize boundary effects due to the finite window. Assuming therefore that the templates are required to have simultaneously a minimum size in space and spatial frequency, it follows from the results of Gabor ([40], see also [29]) that they must be Gabor functions. The following surprising property holds:

Optimal invariance result

Gabor functions of the form (here in 1D) $t(x) = e^{-\frac{x^2}{2\sigma^2}} e^{i\omega x}$ are the templates that are simultaneously maximally invariant for translation and scale (at each x and ω .)

In general, templates chosen at random from the space of images can provide scale and position invariance. However, for optimal invariance under scaling and translation, templates of the Gabor form are optimal. This is the only computational justification we know of the Gabor shape of simple cells in V1 which seems to be remarkably universal: it holds in primates (Optimal invariance result [107]), cats (Optimal invariance result [61]) and mice (Optimal invariance result [93]) (see also Figure 3 for results of simulations and [92]).

D. QUASI-INVARIANCE TO NON-GROUP TRANSFORMATIONS REQUIRES CLASS-SPECIFIC TEMPLATES

All the results so far require a group structure and provide exact invariance for a single new image. In 2D this induces all combinations of translation, scaling, and rotation in the image plane but does not include the transformations induced on the image plane by 3D transformations such as view-point changes and rotation in depth of an object. The latter forms a group in 3D, as if images and templates were 3D views; in principle motion or stereopsis can provide the third

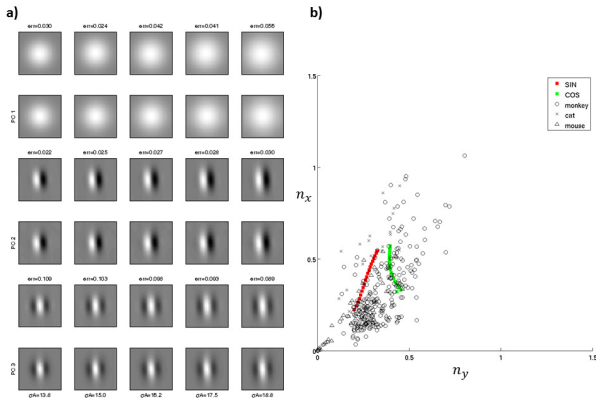


FIGURE 3. a) Simulation results for V1 simple cells learning via PCA (Principal Component Analysis). Each “cell” sees a set of images through a Gaussian window (its dendritic tree), shown in the top row. Each cell then “learns” the same weight vector, extracting the principal components of its input. b) This figure shows $n_y = \sigma_y/\lambda$ vs $n_x = \sigma_x/\lambda$ for the modulated (x) and unmodulated (y) direction of the Gabor wavelet. Notice that the value of the slope σ_y/σ_x is a robust finding in the theory and apparently also in the physiology data. Neurophysiology data from monkeys, cats and mice are reported together with our simulations. Figure from [92].

dimension, though available psychophysical evidence [32], [124] suggests that human vision does not use it for recognition. Notice that transformations in the image plane are affected not only by orthographic projection of the 3D geometry but also by the process of image formation which depends on the 3D geometry of the object, its reflectance properties, and the relative location of the light source and viewer.

It turns out that the HW algorithm can still be applied to non-group transformations - such as transformations of the expression of a face, or of the pose of a body - to provide, under certain conditions, approximate invariance. In this case bounds on the invariance depend on specific details of the object and the transformation: we do not have general results and suspect they may not exist. The key technical requirement is that a new type of sparsity condition holds: sparsity for the class of images I_C with respect to the dictionary t^k under the transformations T_r (we consider here a one parameter (r) transformation)

$$\langle I_C, T_r t^k \rangle \approx 0 \quad |r| > a \quad a \gtrsim 0. \quad (5)$$

This property, which is an extension of the compressive sensing notion of “incoherence”, requires that images in the class and the templates have a representation with sharply peaked correlation and autocorrelation (the constant a above is related to the support of the peak of the correlation). This condition can be satisfied by templates that are similar to images in the set and are sufficiently “rich” to be incoherent for “small” transformations. This relative sparsity condition is usually satisfied by the neural representation of images and templates at some high level of the hierarchy of HW modules that we describe next. Like standard sparsity [29] our new sparsity condition is generic: most neural patterns - templates and images from the same class - chosen

at random will satisfy it. The result [7] takes the following form:

Class-specific property

$\mu_h^k(I)$ is approximately invariant around a view if

- I is sparse in the dictionary of templates relative to the transformations
- I transforms “in the same way” as the templates
- the transformation is smooth

The main implication is that approximate invariance can be obtained for non-group transformation by using templates specific to the class of objects. This means that class specific modules are needed, one for each class; each module requires highly specific templates, that is cell tunings. The obvious example is face-tuned cells in the face patches. Unlike exact invariance for affine transformations where tuning of the “simple cells” is non-specific in the sense that does not depend on the type of image, non-group transformations require highly tuned neurons and yield at best only approximate invariance (see, e.g. [46] and [137]).

E. TWO STAGES IN THE COMPUTATION OF AN INVARIANT SIGNATURE

Hierarchical architectures are advantageous for several reasons which are formalized mathematically in [6], [7], [31], [87], and [105]. It is illuminating to consider two extreme “cartoon” architectures for the first of the two stages described at the beginning of section II:

- one layer comprising one HW module and its KH complex cells, each one with a receptive field covering the whole visual field
- a hierarchy comprising several layers of HW modules with receptive fields of increasing size, followed by parallel modules, each devoted to invariances for a specific object class.

In the first architecture invariance to affine transformations is obtained by pooling over KH templates, with each one transformed in all possible ways: each of the associated simple cells corresponds to the transformation of a template. Invariance over affine transformation is obtained by pooling over the whole visual field. In this case, it is not obvious how to incorporate invariance to non-group transformations directly in this one-hidden layer architecture.

Notice however that a HW module dealing with non-group transformations can be added on top of the affine module. The results in [5] and [7] allow for this factorization. Interestingly, they do not allow in general for factorization of translation and scaling (e.g., one layer computing translation invariance and the next computing scale invariance). Instead, what the mathematics allows is factorization of the range of invariance for the same group of transformations (see also [5] par 3.6-7-8-9). This justifies the first layers of the second architecture, corresponding to Figure 4 stage 1, where the size of the receptive field of each HW module and the range of its invariance increases from lower to higher layers.

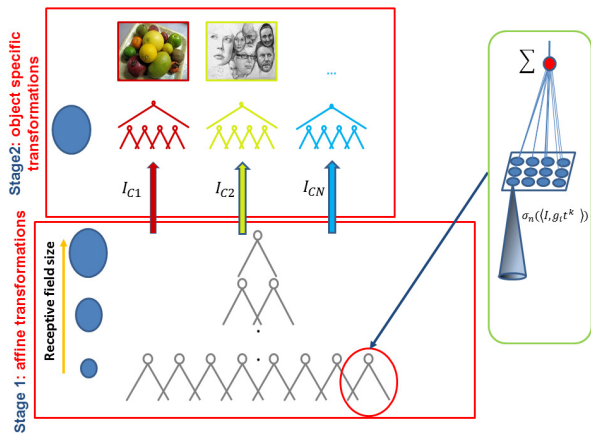


FIGURE 4. A hierarchical architecture of HW modules. The signature provided by each of the nodes at each layer may be used by a supervised classifier. Stage 1: a hierarchy of HW modules (green inset) with growing receptive fields provide a final signature (top of the hierarchy) which is globally invariant to affine transformations by pooling over a cascade of locally invariant signatures at each layer. Stage 2: transformation specific modules provide invariance for non group transformations (e.g., rotation in depth).

F. INVARIANCE TO TRANSLATION AND SCALE (STAGE 1) WITH CLUTTER TOLERANCE REQUIRES A HIERARCHICAL ARCHITECTURE

The main problem with the one-layer architecture is that it can recognize isolated objects in the visual field in an invariant way but cannot recognize objects in clutter: *the key result about invariance assumes that image and templates portray isolated objects*. Otherwise the signature may change because of different clutter at different times.² The problem of clutter - of recognizing an object independently of the presence of another one nearby - is closely related to the problem of recognizing “wholes” and “parts”. Recognizing an eye in a face has the problem that the rest of the face is clutter. This is the old conundrum of recognizing a tree in a forest while still recognizing the forest.

A partial solution to this problem is to use a hierarchical architecture for stage 1 in which lower layers provide signatures with a small range of invariance for “small” parts of the image and higher layers provide signatures with greater invariance for larger parts of the image. This signature could then be used by class specific modules. Two points are of interest here.

Factorization of a range of invariances is possible if a certain property of the hierarchical architecture, called equivariance, holds. Assume a group transformation of the image, e.g., a translation or scaling of it. The first layer in a hierarchical architecture is called equivariant if the pattern of neural activity at the output of the complex cells transforms accordingly to the same group of transformations. The equivariance property is also very important in modern neural networks

²Notice that because images are filtered by the retina with spatial bandpass filters (ganglion cells), the input to visual cortex is a rather sparse pattern of activities, somewhat similar to a sparse edge map.

(see, e.g., [23] and [24]). It turns out that the architectures we describe have this property (see [5] and [7] par 3.5.3 for the translations case): isotropic architectures, like the ones considered in this paper, with point-wise nonlinearities are equivariant. The key difference from the architecture described above is that equivariance can be achieved when the complex cells pool over single cells responses coming from templates transforming w.r.t. a subset of all group transformations. In this way the complex first layer representation will be invariant to “small” transformations (e.g., small translations) but still carry information about “large” transformations (equivariance). Since each module in the architecture gives an invariant output if the transformed object is contained in the pooling range, and since the pooling range increase from one layer to the next, there is an invariance over larger and larger transformations. The second point is that in order to make recognition possible for both parts and wholes of an image, the supervised classifier should receive signatures not only from the top layer (as in most modern neural architectures) but also from the other levels as well (directly or indirectly).

III. BIOPHYSICAL MECHANISMS OF INVARIANCE: UNSUPERVISED LEARNING, TUNING AND POOLING

A. A SINGLE CELL MODEL OF SIMPLE AND COMPLEX CELLS

There are at least two possible biophysical models for the HW module implied by our theory. The first is the original Hubel and Wiesel model of simple cells feeding into a complex cell. Our theory proposes the “ideal” computation of a CDF, in which case the nonlinearity at the output of the simple cells is a threshold function. A complex cell, summing the outputs of a set of simple cells, would then represent a bin of the histogram; a different complex cell in the same position pooling a set of similar simple cells with a different threshold would represent another bin of the histogram. Another possibility is that the nonlinearity at the output of the simple cells is a square or any power or combination of powers. In this case the complex cell pooling simple cells with the same nonlinearity would represent a moment of the distribution, including the linear average. The nonlinear transformation at the output of the simple cells would correspond to the spiking mechanism in populations of cells (see, e.g., references in [65]).

The second biophysical model for the HW module that implements the computation required by our theory consists of a single cell where dendritic branches play the role of simple cells (each branch containing a set of synapses with weights providing, for instance, Gabor-like tuning of the dendritic branch) with inputs from the LGN (lateral geniculate nucleus); active properties of the dendritic membrane distal to the soma provide separate threshold-like nonlinearities for each branch separately, while the soma sums the contributions for all the branches. This model would solve the puzzle that there seems to be no morphological difference between pyramidal cells classified as simple vs complex by physiologists.

It is interesting that our theory is robust with respect to the nonlinearity from simple to the complex “cells”. We conjecture that almost any set of non trivial nonlinearities will work. This argument rests on the fact that a set of different complex cells pooling from the same simple cells should compute the cumulative distribution or equivalently its moments or combinations of moments (each combination is a specific nonlinearity). Any nonlinearity will provide invariance, if the nonlinearity does not change with time and is the same for all the simple cells pooled by the same complex cells. A sufficient number of different nonlinearities, each corresponding to a complex cell, can provide appropriate selectivity.

B. LEARNING THE WIRING

A simple possibility for how the wiring between a group of simple cells with the same tuning (for instance representing the same eigenvector, with the same orientation etc.) and a complex cell may develop is to invoke a Hebbian trace rule ([35], see also [10] and [89]). In a first phase complex cells may have subunits with different selectivities (e.g. orientations), for instance because natural images are rotation invariant and thus eigenvectors with different orientations are degenerate. In a second plastic phase, subunits which are inactive when the majority of the subunits are active will be pruned out according to a Foldiak-like rule.

C. HEBB SYNAPSES AND PCAs

Our theory provides the following algorithm for learning the relevant invariances during unsupervised visual experience: store a sequences of images for each of a few objects (called “templates”) with transformations - for instance translating, rotating, and looming. Section II shows that in this way invariant hierarchical architectures can be learned from unsupervised visual experience. Such architectures represent a significant extension beyond simple translation invariance, and beyond hardwired connectivity, of models of the ventral stream such as Fukushima’s Neocognitron [36] and HMAX [106], [118] – as well as deep neural network of convolutional type ([70] and related models, e.g., [1], [73], [88], [97], [98], [115], [122]) or models where the symmetry is not learned but hardwired, see, e.g., [25], [108]. Note that other algorithms to learn symmetries has been recently proposed for artificial neural networks, e.g., [13], [26], [134]. However their biological plausibility is not clear (see also [47]).

In biological terms, the sequence of transformations of one template would correspond to a set of simple cells, each one storing in its tuning a frame of the sequence. In the second learning step a complex cell would be wired to those “simple” cells. However, the idea of a direct storage of sequences of images or image patches in the tuning of a set of V1 cells by exposure to a single object transformation is biologically rather implausible. Since Hebbian-like synapses are known to exist in visual cortex a more natural hypothesis is that synapses would incrementally change over time as an effect of the visual inputs - that is over many sequences of images resulting from transformations of objects, e.g., templates.

The question is whether or not such a mechanism is compatible with our theory and how to implement it if so.

We explore this question for V1 in a simplified setup that can be extended to other areas. We assume:

- a) that the synapses between LGN inputs and (immature) simple cells are Hebbian and in particular that their dynamics follows Oja’s flow [64], [95]. In this case, the synaptic weights will converge to the eigenvector with the largest eigenvalue of the covariance of the input images.
- b) that the position and size of the untuned simple cells is set during development according to an inverted pyramidal lattice (see Figure 3 in [102]). The key point here is that the size of the Gaussian spread of the synaptic inputs and the positions of the ensemble of simple cells are assumed to be set independently of visual experience.

In summary we assume that the neural equivalent of the memorization of frames (of transforming objects) is performed online via Hebbian synapses that change as an effect of visual experience. Specifically, we assume that the distribution of signals “seen” by a maturing simple cell is Gaussian in x, y reflecting the distribution on the dendritic tree of synapses from the lateral geniculate nucleus. We also assume that there is a range of Gaussian distributions with different σ variances which increase with retinal eccentricity. As an effect of visual experience the weights of the synapses are modified by a Hebb rule [50]. Hebb’s original rule can be written as

$$w_n = \alpha y(x_n)x_n \quad (6)$$

where α is the “learning rate”, x_n is the input vector, w is the presynaptic weights vector, and y is the postsynaptic response. In order for this dynamical system to actually converge, the weights have to be normalized. In fact, there is considerable experimental evidence that the cortex employs normalization [130] and references therein). Hebb’s rule, appropriately modified with a normalization factor, turns out to be an online algorithm to compute PCA from a set of input vectors. In this case it is called Oja’s flow. Oja’s rule [64], [95] defines the change in presynaptic weights w given the output response y of a neuron given its inputs x to be

$$w_n = w_{n+1} - w_n = \alpha y_n(x_n - y_n w_n) \quad (7)$$

where $y_n = w_n^T x_n$. The equation follows from expanding to the first order Hebb’s rule, normalized to avoid divergence of the weights.

Since the Oja flow converges to the eigenvector of the covariance matrix of the x_n which has the largest eigenvalue, we analyze the spectral properties of the inputs to “simple” cells and study whether a PCA computation can be used by the HW algorithm and in particular whether it satisfies the selectivity and invariance results.

Alternatives to the Oja’s rule that still converge to PCAs can also be considered (see [113] and [96]). Also notice that a relatively small change in the Oja equation gives an

online algorithm for computing ICAs (Independent Component Analysis) instead of PCAs (see [56]). Which kind of plasticity is more biologically plausible remains an open question.

D. SPECTRAL THEORY AND POOLING

Consider stage 1, which is retinotopic, and particularly the case of simple cells in V1. From our assumptions in section V, the lattice in x, y, s of immature simple cell is set during the development of the organism (s is the size of the Gaussian envelope of the immature cell). Assume that all of the simple cells are exposed, while in a plastic state, to a possibly large set of images $T = (t_1, \dots, t_K)$. A specific cell at a certain position in x, y, s is exposed to the set of transformed templates g_*T (where g_* corresponds to the translation and scale that transforms the “zero” cells to the chosen neuron in the lattice) and therefore the associated covariance matrix $g_*TT^Tg_*^T$. Thus it is possible to choose PCA as new templates, and pooling over corresponding PCAs across different cells is equivalent to pooling over a template and its transformations. Both the invariance and selectivity result are valid. Empirically, we find ([77]) that PCA of natural images provides eigenvectors that are Gabor-like wavelets with a random orientation for each sized receptive field. The random orientation is because of the argument above, together with the fact that the covariance of natural images is approximately rotation invariant. The Gabor-like shape can be qualitatively explained in terms of translation invariance of the correlation matrix associated with a set of natural images (and their approximate scale invariance which corresponds to a $\approx 1/f$ spectrum, see also [111], [114], and [127]).³ Thus the Oja rule acting on natural images provides “equivalent templates” that are Gabor-like: the optimal templates, according to the theory of section IIC.

Consider now non-retinotopic stage 2 in which transformations are not in scale or position, such as the transformation induced by a rotation of a face. Assume that a “simple” cell is exposed to “all” transformations g_i (g_i is a group element of the finite group G) of each of a set $T = (t_1, \dots, t_K)$ of K templates. The cell is thus exposed to a set of images (columns of X) $X = (g_1T, \dots, g_{|G|}T)$. For the sake of this example, assume that G is the discrete equivalent of a group. Then the covariance matrix determining the Oja’s flow is

$$C = XX^T = \sum_{i=1}^{|G|} g_iTT^Tg_i^T. \quad (8)$$

It is immediately clear that if ϕ is an eigenvector of C then $g_i\phi$ is also an eigenvector with the same eigenvalue (for more details on how receptive fields look like in V1 and higher

³Suppose that the simple cells are exposed to patterns and their scaled and translated versions. Suppose further that images are defined on a lattice and translations and scaling (a discrete similitude group) are carefully defined on the same lattice. Then a set of discrete orthogonal wavelets - defined in terms of discrete dilation and shifts - exist and is invariant under the group. The Oja rule (extended beyond the top eigenvector) could converge to specific wavelets.

layers see also [5], [101] par 4.3.1 and 4.7.3, [10], [41], [42], [51]). Consider an example G to be the discrete rotation group in the plane: then all the (discrete) rotations of an eigenvector are also eigenvectors. The Oja rule will converge to the eigenvectors with the top eigenvalue and thus to the subspace spanned by them. It can be shown that L^2 pooling over the PCA with the same eigenvalues represented by different simple cells is then equivalent to L^2 pooling over transformations, as the theory of section II.B dictates, in order to achieve selectivity and invariance ([5] par 4.6.1 and [10]). This argument can be formalized in the following variation of the pooling step in the HW algorithm:

Spectral pooling. Suppose that M is the matrix corresponding to the group transformations of template t (each column is a transformation of the template). Consider the set of eigenvectors $\{\phi_k^*\}_{k=1}^K$ of covariance matrix $C = MM^T$ with eigenvalue λ^* . Because of the above argument $\langle g_mI, \phi_k \rangle = \langle I, \phi_p^* \rangle$ where $g_m^{-1}\phi_k^* = \phi_p^*$. Therefore to achieve invariance a complex cell can pool with a quadratic nonlinearity over the eigenvectors of C instead of over the transformations of the template. Thus, components of an invariant signature can be computed as

$$\mu_*(I) = \sum_i \|\langle I, \phi_i^* \rangle\|^2. \quad (9)$$

E. TUNING OF “simple” CELLS

The results of section II on the HW module imply that the templates, and therefore the tuning of the simple cells, can be the image of any object. At higher levels in the hierarchy, the templates are neuroimages - patterns of neural activity - induced by actual images in the visual field. The previous section, however, offers a more biologically plausible way to learn the templates from unsupervised visual experience via Hebbian plasticity. In the next sections we will discuss predictions following from this assumptions for the tuning of neurons in the various areas of the ventral stream.

IV. STAGE 2 IN IT: CLASS-SPECIFIC APPROXIMATE INVARIANCE

A. FROM GENERIC TEMPLATES TO CLASS-SPECIFIC TUNING

As discussed in section IIE, approximate invariance for transformations beyond the affine group requires highly tuned templates, and therefore highly tuned simple cells, probably at a level in the hierarchy corresponding to AIT (anterior inferotemporal cortex). According to the considerations of section IIF this is expected to take place in higher visual areas of the hierarchy. In fact, the same localization condition of Equation 4 suggests Gabor-like templates for generic images in the first layers of a hierarchical architecture and specific tuned templates for the last stages of the hierarchy, since class specific modules are needed with each containing highly specific templates, and thus highly tuned cells. This is consistent with the architecture of the ventral stream and the the existence of class-specific modules in primate cortex

such as a face module and a body module [28], [30], [52], [58], [63], [76], [128]. We saw in section IIF that areas in the hierarchy up to V4 and/or PIT provide signatures for larger parts or full objects. Thus we expect

- that the inputs to the class-specific modules are scale and shift invariant
- that the class-specific templates are “large”. For instance in the case of faces, templates should cover significant regions of the face. Notice that only large templates support pose invariance (the image of an isolated eye does not change much under rotations in depth of the face).

B. DEVELOPMENT OF CLASS-SPECIFIC AND OBJECT-SPECIFIC MODULES

A conjecture emerging from our theory offers an interesting perspective [77] on AIT. For transformations that are not affine transformations in 2D (we assume that 3D information is not available to the visual system or used by it, which may not always be true), an invariant representation cannot be computed from a single view of a novel object because the information available is not sufficient. What is lacking is the 3D structure and material properties of the object: thus exact invariance to rotations in depth or to changes in the direction or spectrum of the illumination cannot be obtained. However, as our theory shows, approximate invariance to smooth non-group transformations can still be achieved in several cases (but not always) using the same HW module. The reason this will often approximately work is because it effectively exploits prior knowledge of how similar objects transform. The image-to-image transformations caused by a rotation in depth are not the same for two objects with different 3D structures. However, objects that belong to an object class where all the objects have a similar 3D structure transform their 2D appearance in (approximately) the same way. This commonality is exploited by a HW module to transfer the invariance learned from (unsupervised) experience with template objects to novel objects seen from only a single example view. This is effectively our definition of an object class: a class of objects such that the transformation for a specific object can be approximately inferred from how other objects in the class transform. The necessary condition for this to hold is that the 3D shape is similar between any two objects in the class. The simulation in Figure 5 shows that HW-modules tuned to templates from the same class of the (always novel) test objects provide a signature that tolerates substantial viewpoint changes (plots on the diagonal); it also shows the deleterious effect of using templates from the wrong class (plots off the diagonal). There are of course several other class-specific transformations besides depth-rotation, such as face expression and body pose transformations.

In [77] we argue that the visual system is continuously and automatically clustering objects and their transformations - observed in an unsupervised way - in class-specific modules. New images are added to an existing module only if their

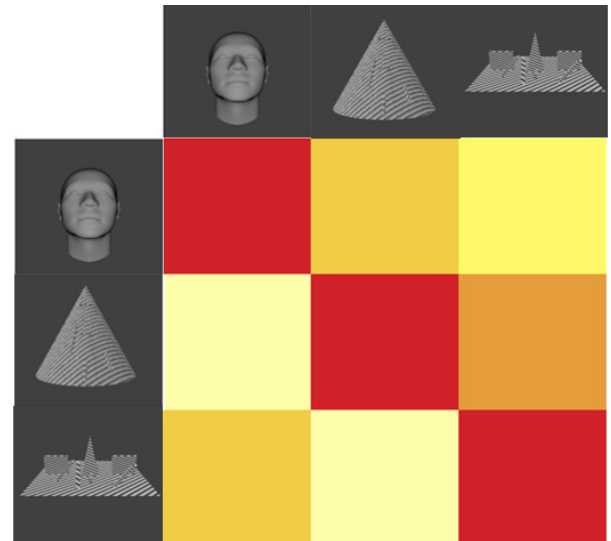


FIGURE 5. Class-specific transfer of depth-rotation invariance for images from three classes (faces, A, cylinders, B and composed, C). The left column of the matrix shows the results of the test for invariance for a random image of a face (A) in different poses w.r.t. 3D rotation using 3D rotated templates from A,B,C; similarly the middle and the right column shows the invariance results for class B and C tested on rotated templates of A,B,C respectively. The colors in the matrix show the maximum invariance range (degrees of rotation away from the frontal view). Only the diagonal values of the matrix (train A - test A, train B - test B, train C - test C) show an improvement of the view-based model over the pixel representation. That is, only when the test images transform similarly to the templates is there any benefit from pooling [77].

transformation are well predicted by it. If no module can be found with this property the new image and all its transformations will be the seed of a new object cluster/module.

For the special case of rotation in depth, [77], ran a simulation using 3D modelling / rendering software to obtain the transformations of objects for which there exist 3D models. Faces had the highest degree of clustering of any naturalistic category - unsurprising since recognizability likely influenced face evolution. A set of chair objects had broad clustering, implying that little invariance would be obtained from a chair-specific region. A set of synthetic “wire” objects, very similar to the “paperclip” objects used in several classic experiments on view-based recognition, e.g., ([12], [82], and [83]) were found to have the smallest index of clusterability: experience with familiar wire objects does not transfer to new wire objects (because the 3D structure is different for each individual paperclip object).

It is instructive to consider the limit case of object classes that consist of single objects - such as individual paperclips. If the object is observed under rotation several frames are memorized as transformations of a single template (identity is implicitly assumed to be conserved by a Foldiak-like rule, as long as there is continuity in time of the transformation). The usual HW module pooling over them will allow view-independent recognition of the specific object. A few comments:

- 1) remarkably, the HW module described above for class-specific transformations - when restricted to

- multiple-views, single-object – is equivalent⁴ to the Edelman-Poggio model for view invariance [100];
- 2) the class-specific module is also effectively a “gate”: in addition to providing a degree of invariance it also performs a template matching operation with templates that can effectively “block” images of other object classes. This gating effect may be important for the system of face patches discovered by Tsao and Freiwald [37] and it is especially obvious in the case of a single object module;
 - 3) from the point of view of evolution, the use of the HW module for class-specific invariances can be seen as a natural extension from its role in single-objects view invariance. The latter case is computationally less interesting, since it implements effectively a look-up table, albeit with interpolation power. The earlier case is more interesting since it allows generalization from a single view of a novel object. It also represent a clear case of transfer learning.

C. DOMAIN-SPECIFIC REGIONS IN THE VENTRAL STREAM

As discussed by [77], there are other domain-specific regions in the ventral stream besides faces and bodies. It is possible that additional regions for less-common or less transformation-compatible object classes will appear with higher resolution imaging techniques. One example may be the fruit area, discovered in macaques with high-field fMRI [69]. Others include the body area and the Lateral Occipital Complex (LOC) which according to recent data [85] is not really a dedicated region for general object processing but a heterogeneous area of cortex containing many domain-specific regions too small to be detected with the resolution of fMRI (but see also [99] and [135]). The Visual Word Form Area (VWFA) [19], [22] seems to represent printed words. In addition to the generic transformations that apply to all objects, printed words undergo several non-generic transformations that never occur with other objects. For instance, our reading is rather invariant to font transformations and can deal with hand-written text. Thus, VWFA is well-accounted for by the invariance hypothesis, as words are a frequently-viewed stimuli which undergo class-specific transformations.

The justification - really a prediction - by our theory for domain-specific regions in cortex is different from other proposals. However, it is complementary w.r.t. some of them, rather than exclusive. For instance, it would make sense that the clustering depends not only on the index of compatibility but also on the relative frequency of each object class. We conjecture that a) that transformation compatibility is the critical factor driving the development of domain-specific regions, and b) that there are separate modules for object classes that transform differently from one another.

⁴In [100] the similarity operation was the Gaussian of a distance - instead of the dot product required by our theory. Notice that for normalized vectors, l_2 norms and dot products are equivalent.

D. TUNING OF “SIMPLE” CELLS IN IT

In the case of “simple” neurons in the AL face patch [37], [39], [77], exposure to several different faces – each one generating several images corresponding to different rotations in depth – yields a set of views with a covariance function which has eigenvectors (PCs) that are either even or odd functions (because faces are bilaterally symmetric).

The Class-specific result together with the Spectral pooling proposition suggests that square pooling (over these face PCs) provides approximate invariance to rotations in depth. The full argument goes as follows. Rotations in depth of a face around a certain viewpoint - say $\theta = \theta^0$ - can be well approximated by linear transformations (by transformations in the general linear group, $g \in GL(2)$). The HW algorithm can then provide invariance around $\theta = \theta^0$. Finally, if different sets of “simple” cells are plastic at somewhat different times, exposure to a partly different set of faces yields different eigenvectors summarizing different sets of faces. The different sets of faces play the role of different object templates in the standard theory.

The limit case of object classes that consist of single objects is important to understand the functional architecture of most of IT. If an object is observed under transformations, several images of it can be memorized and linked together by continuity at time of the transformation. As we mentioned, the usual HW module pooling over them will allow view-independent recognition of the specific object. Since this is equivalent to the Edelman-Poggio model for view invariance [100] there is physiological support for this proposal (see [83], [84] and [123]).

E. MIRROR SYMMETRIC TUNING IN THE FACE PATCHES AND POOLING OVER PCs

The theory then offers a direct interpretation of the Tsao-Freiwald data (see [37] [38], and [39]) on the face patch system. The most posterior patches (ML/MF) provide a view and identity specific input to the anterior patch AL, where most neurons show tuning that is an even function of the rotation angle around the vertical axis. AM, which receives inputs from AL, is identity-specific and view-invariant. The puzzling aspect of this data is the mirror symmetric tuning in AL: why does this appear in the course of a computation that leads to view-invariance? According to our theory this result should be expected if AL contains “simple” cells that are tuned by a synaptic Hebb-like Oja rule and the output of the cells is roughly a squaring nonlinearity as required by the Spectral pooling proposition. In this interpretation, cells in AM pool over several of the squared eigenvector filters to obtain invariant second moments (see Figure 6). Detailed models from V1 to AM show properties that are consistent with the data and also perform well in invariant face recognition [75], [77], [81], [91].

V. DISCUSSION

Several different levels of understanding. Our theory addresses several different levels, including the computational

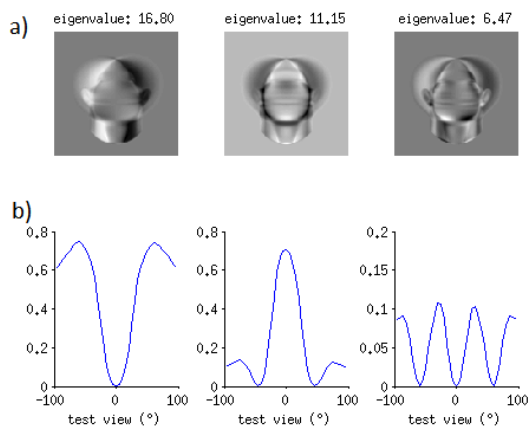


FIGURE 6. Face identity is represented in the macaque face patches (Freiwald and Tsao, 2010). Neurons in the middle areas of the ventral stream face patch (middle lateral and fundus (ML, MF)) are view specific, while those in the most anterior (anterior medial patch (AM)) are view invariant. Neurons in an intermediate area (anterior lateral patch (AL)) respond similarly to mirror-symmetric views. In our theory view invariance is obtained by pooling over “simple” neurons whose tuning corresponds to the PCAs of a set of faces previously experienced each under a range of poses. Due to the bilateral symmetry of faces, the eigenvectors of the associated covariance matrix are even or odd. This is shown in a) where the first 3 PCAs of set of grey-level faces under different poses are plotted: the same symmetry arguments apply to “neural” images of faces. Figure b shows the response of 3 model AL units to a face stimulus as a function of pose under different poses (From [77]).

goal of the ventral stream, the algorithms used, the architecture of visual cortex, its hierarchical architecture, and the neural circuits underlying tuning of cells. This is unlike most other models or theories.

Predictions. From the point of view of neuroscience, the theory makes a number of predictions, some obvious, some less so. One of the main predictions is that simple and complex cells should be found in all visual and auditory areas, not only in V1. Our definition of simple cells and complex cells is different from the traditional ones used by physiologists; for example, we propose a broader interpretation of complex cells, which in the theory represent invariant measurements associated with histograms (or moments of their values) of the outputs of simple cells. The theory implies that, under some conditions, exact or approximate invariance to all geometric image transformations can be learned, either during development or in adult life. It is, however, also consistent with the possibility that basic invariances may be genetically encoded by evolution and possibly refined and maintained by unsupervised visual experience. A single cell model for simple complex cells follows from the theory as an interesting possibility. Our theory also makes predictions about the architecture of the ventral stream:

- the output of V2, V4, and PIT should access memory either via connections that bypass higher areas or indirectly via equivalent neurons in higher areas (because of the argument in a previous section about clutter).
- areas V1, V2, V4 and possibly PIT are mainly dedicated to computing signatures that are invariant to translation,

scale and their combinations - as experienced in past visual experience.

- IT is a complex of parallel class-specific modules for a large number of large and small object classes. These modules receive position and scale invariant inputs (invariance in the inputs greatly facilitates unsupervised learning of class specific transformations). We recall that, from the perspective of the theory, the data of [83] concern single object modules and strongly support the prediction that exposure to a transformation lead to neuronal tuning to several “frames” of it.

Object-based vs 3D vs view-based recognition. We should mention here an old controversy about whether visual recognition is based on views or on 3D primitive shapes called geons. In the light of our theory image views retain the main role but ideas related to 3D shape may also be valid. The psychophysical experiments of Edelman and Buelthoff [32] concluded that generalization for rotations in depth was limited to a few degrees ($\approx \pm 30$ degrees) around a view (independently of whether 2D or 3D information was provided to the human observer (psychophysics in monkey [82], [83] yielded similar results). The experiments were carried out using “paperclip” objects with random 3D structure (or similar but smoother objects). For this type of objects, class-specific learning is impossible (they do not satisfy the second condition in the class-specific result) and thus our theory predicts the result obtained by Edelman and Buelthoff. For other objects, however, such as faces, the generalization that can be achieved from a single view by our theory can span a much larger range than ± 30 degrees, effectively exploiting 3D-like information from templates of the same class.

Genes or learning. Our theory shows how the tuning of the “simple” cells in V1 and other areas could be learned in an unsupervised way. It is possible however that the tuning - or the ability to quickly develop it in interaction with the environment - may have been partly compiled during evolution into the genes.⁵ Notice that this hypothesis implies that most of the times the specific function is not fully encoded in the genes: genes facilitate learning but do not replace it completely. It is then be expected in the “nature vs nurture debate” that usually nature needs nurture and nurture is made easier by nature. An interesting result in this respect comes from a recent paper [66] where the authors propose a genetic

⁵If a function learned by an individual represents a significant evolutionary advantage we could expect that aspects of learning the specific function may be encoded in the genes, since an individual who learns more quickly has a significant advantage. In other words, the hypothesis implies a mix of nature and nurture in most competencies that depend on learning from the environment (like perception). This is an interesting implication of the “Baldwin effect” - a scenario in which a character or trait change occurring in an organism as a result of its interaction with its environment becomes gradually assimilated into its developmental genetic or epigenetic repertoire [27]. In the words of Daniel Dennett, “Thanks to the Baldwin effect, species can be said to pretest the efficacy of particular different designs by phenotypic (individual) exploration of the space of nearby possibilities. If a particularly winning setting is thereby discovered, this discovery will create a new selection pressure: organisms that are closer in the adaptive landscape to that discovery will have a clear advantage over those more distant.” (p. 69, quoting [27]).

bottleneck as an effective regularizer that enables evolution to select simple circuits that can be readily adapted to important real-world tasks (see also [48]).

Computational structure of the HW module. The HW module computes the CDF of $\langle I, g_i t^k \rangle$ over all $g_i \in G$. The computation consists of

$$\mu_h^k(I) = \frac{1}{|G|} \sum_{i=1}^{|G|} \sigma(\langle I, g_i t^k \rangle + h\Delta) \quad (10)$$

with $h = 0, \cdot, H$ and $k = 1, \cdot, K$; the main forms of the nonlinearity σ are either a threshold function or a power $n = 1, \cdot, \infty$ of its argument. Several known networks are special cases of this module. One interesting case is when G is the translation group and $\sigma(\cdot) = \|\cdot\|^2$: then the equation is equivalent (for $H = 0$) to a unit in a convolutional network with max pooling. In another noteworthy case (we always assume that I and t^k are normalized) the equation is very similar to the Radial Basis Functions network proposed by Edelman and Poggio [100] for view classification. In this spirit, note that the equation for a unit in a convolutional network is

$$\frac{1}{|G|} \sum_{i=1}^{|G|} c_i \sigma(\langle I, g_i t \rangle + h\Delta) \quad (11)$$

where I is the input vector, c_i, t, Δ are parameters to be learned, in supervised mode, from labeled data and $g_i t(x) = t(x - i\delta_x)$. Thus units in convolutional network could learn to become units of the our theory (by learning $c_i = 1$) but only when G is the translation group (in our theory G is the full affine group for the first layers and can be a non group such as the transformation induced by rotations in depth). We also note that the same computational model can be extended to other sensory modalities (I can be, e.g., a soundwave). Moreover, the learned representation can be used not only for image recognition but for a variety of different tasks that integrate multisensory information.

Relations to Deep Learning networks and unsupervised learning. Most of the best performing deep learning networks have convolutional layers as well as densely connected layers in their architecture. Our theory applies to the convolutional but not the densely connected, classification stage. Historically, hardwired invariance to translation was first introduced in the Neocognitron by Fukushima and later in LeNet [70] and in HMAX ([106]; HMAX had also invariance to scale). These architectures are early examples of convolutional networks. Our theory provides a general theory for them⁶ that also offers two significant algorithmic and architectural extensions: a) it ensures, within the same algorithm, invariances to other groups beyond translation and provides approximate invariances to certain non-group transformations; b) it provides a way to learn arbitrary invariances from unsupervised learning. Learning the correct symmetries

⁶In the case of the translation group the HW module (see Equation 1) consists of (non-linear) pooling of the convolution of the image with a template.

can potentially give an advantage in terms of sample complexity w.r.t. hardwired translation invariant convolutional networks. Note however that, differently from state of art CNNs (Convolutional Neural Networks), our proposed learning algorithm is unsupervised and a direct comparison could be potentially misleading. In general, supervised methods are nowadays superior to unsupervised ones although unsupervised models based on contrastive embeddings [139] or biologically plausible backpropagation [57] and contrastive predictive coding [11] might offer an alternative. A more appropriate comparison can be done with architecture where the signal representation is learned in an unsupervised way and supervision is only used to adapt the representation to the specific task(s) [68]. Finally note that our unsupervised model differs from those cited above in that we make the hypothesis that the input is a collection of group transformed sensory signals. We take advantage of this data structure deriving the equivariant and invariant properties of the representation.

Invariance in 2D and 3D vision. We have assumed here that “images” as well as templates are in 2D. This is the case if possible sources of 3D information such as stereopsis and/or motion are eliminated. Interestingly, it seems that stereopsis does not facilitate recognition, suggesting that 3D information, even when available, is not used by the human visual system (see [15]).⁷

Explicit or implicit gating of object classes. The second stage of the recognition architecture consists of a large set of object-class specific modules of which probably the most important is the face system. It is natural to think that signals from lower areas should be gated, in order to route access only to the appropriate module. In fact, Tsao [129], but see also [18] postulated a gate mechanism for the network of face patches. The structure of the modules however suggests that the module themselves automatically provides a gating function even if their primary computational function is invariance. This is especially clear in the case of the module associated with a single object (the object class consists of a single object as in the case of a paperclip). The input to the module is subject to dot products with each of the stored views of the object: if none match well enough the output of the module will be close to zero, effectively gating off the signal and switching off subsequent stages of processing.

Invariance to X and estimation of X. The description of our theory focuses on the problem of recognition as estimating identity or category invariantly to a transformation X - such as translation or scale or pose. Often however the complementary problem, of estimating X , for instance pose, is also important. The same neural population may be able to support both computations and multiplex the representations of their

⁷This hypothesis should however be checked further since our theory implies that if 3D information is available, rotation in depth is a group and therefore generalization from a single view could be available simply by having stored 3D templates of a few arbitrary objects and their 3D transformations. This is not what psychophysics (for instance on the paperclips) shows; however, the mathematical claim of perfect invariance is only true in the absence of self-occlusions, a clearly unrealistic assumption for most objects.

outcome as shown in IT recordings and model simulations ([55], [118] but see also [117]). As human observers, we are certainly able to estimate position, rotation, and illumination of an object without eye movements. HW modules pooling over the same units in different way - pooling over identities for each pose or pooling over pose for each identity - can provide the different types of information using the same “simple” cells and different “complex” cells. Anselmi ([5], fig 45) show simulations of recognizing a specific body invariantly to pose and estimating pose-out of a set of 32 possibilities-of a body invariantly to the identity.

PCAs vs ICAs. Independent Component Analysis [56] and similar unsupervised mechanisms describe plasticity rules similar to the basic Oja flow analyzed in this paper. They can generate Gabor-like receptive fields and they may not need the assumption of different sizes of Gaussian distributions of LGN synapses. We used PCA simply because its properties are easier to analyze and should be indicative of the properties of similar Hebbian-like mechanisms.

Parsing a scene. Full parsing of a scene cannot be done in a single feedforward processing step in the ventral stream. It requires task-dependent top-down control, in general multiple fixations and therefore observation times longer than ≈ 100 msec. This also follows from the limited high resolution region of the inverted pyramid model of the visual system, which theory predicts as a consequence of simultaneous invariance to shift and scale (see [102] for details). In any case, full parsing of a scene is beyond what a purely feedforward model can provide.

Feedforward and feedback. We have reviewed a forward theory of recognition and some of the related evidence. Our theory does not address top-down, recurrent, or horizontal connectivity and their computational role. It however makes it easier to consider plausible hypothesis. The inverted pyramid architecture that follows from scale and position invariance requires a tight loop between different fixations in which an efficient control module drives eye movements by combining task requirements with memory access. However, within a single fixation the space-scale inverted pyramid cannot be shifted in space. What could be controlled in a feedback mode are parameters of pooling, including the choice of which scales to use depending on the results of classification or memory access. The most obvious limitation of feedforward architectures is recognition in clutter and the most obvious way around the problem is the attentional masking of large parts of the image under top-down control. More generally, a realistic implementation of the present theory requires top-down control signals and circuits, supervising learning and possibly fetching signatures from different areas and at different locations in a task-dependent way. An even more interesting hypothesis is that backprojections update local signatures at lower levels depending on the scene class currently detected at the top (an operation similar to the top-down pass of Ullman [14]). In summary, the output of the feedforward pass is used to retrieve labels and routines associated with the image; backprojections may implement

an attentional focus of processing to reduce clutter effects and also to run visual routines [118] at various levels of the hierarchy. An interesting, but not biologically plausible, alternative might be offered by the recently introduced transformers architecture, [136].

Motion helps learning isolated templates. Ideally templates and their transformations should be learned without clutter. It can be argued that if the background changes between transformed images of the same template then the averaging effect intrinsic to pooling will mostly “average out” the effect of clutter during the unsupervised learning stage. Though this is correct and we have computer simulations that provide empirical support to the argument, it is interesting to speculate that motion could provide a simple way to eliminate most of the background. Sensitivity to motion is one of the earliest visual computations to appear in the course of evolution and one of the most primitive. Stationary images on the retina tend to fade away. Detection of relative movement is a strong perceptual cue in primate vision as well as in insect vision, probably with similar normalization-like mechanisms [49], [103]. Motion induced by the transformation of a template may then serve two important roles:

- To bind together images of the same template while transforming: continuity of motion is implicitly used to ensure that identity is preserved;
- To eliminate background and clutter by effectively using relative motion.

The required mechanisms are probably available in the retina and early visual cortex.

Despite significant advances in sensory neuroscience over the last five decades, a true understanding of the basic functions of the ventral stream in visual cortex has proven to be elusive. Thus it is interesting that the theory used in this paper follows from a novel hypothesis about the main computational function of the ventral stream: the representation of new objects/images in terms of a signature which is invariant to transformations learned during visual experience, thereby allowing recognition from very few labeled examples—in the limit, just one. This view of the cortex may also represent a novel theoretical framework for the next major challenge in learning theory beyond the relatively-mature supervised learning: the problem of representation learning, formulated here as the unsupervised learning of invariant representations that significantly reduce the sample complexity of the supervised learning stage.

CONFLICT OF INTERESTS

The authors declare no conflict of interests

ACKNOWLEDGMENT

The authors would like to thank Danny Harari, Lorenzo Rosasco, and especially to Gabriel Kreiman for discussions, and also would like to thank Ryan Pyle for reading the manuscript.

REFERENCES

- [1] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 1248–1252.
- [2] Y. S. Abu-Mostafa, "Hints and the VC dimension," *Neural Comput.*, vol. 5, no. 2, pp. 278–288, Mar. 1993.
- [3] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 2, no. 2, p. 284, 1985.
- [4] A. Achille and S. Soatto, "Emergence of invariance and disentanglement in deep representations," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2018, pp. 1–34.
- [5] F. Anselmi, J. Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio, "Magic materials: A theory of deep hierarchical architectures for learning sensory representations," Massachusetts Inst. Technol., Cambridge, MA, USA, CBCL Paper, Tech. Rep., 2013.
- [6] F. Anselmi, L. Rosasco, and T. Poggio, "On invariance and selectivity in representation learning," *Inf. Inference*, vol. 5, no. 2, pp. 134–158, Jun. 2016.
- [7] F. Anselmi, J. Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio, "Unsupervised learning of invariant representations," *Theor. Comput. Sci.*, vol. 633, pp. 112–121, Jun. 2016.
- [8] T. Poggio and F. Anselmi, *Visual Cortex and Deep Networks: Learning Invariant Representations*. Cambridge, MA, USA: MIT Press, 2016.
- [9] F. Anselmi, G. Evangelopoulos, L. Rosasco, and T. Poggio, "Symmetry-adapted representation learning," *Pattern Recognit.*, vol. 86, pp. 201–208, Feb. 2019.
- [10] F. Anselmi, A. Patel, and L. Rosasco, "Neurally plausible mechanisms for learning selective and invariant representations," *J. Math. Neurosci.*, vol. 10, no. 1, Dec. 2020.
- [11] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, [arXiv:1807.03748](https://arxiv.org/abs/1807.03748).
- [12] M. Bar, E. Aminoff, and D. L. Schacter, "Scenes unseen: The parahippocampal cortex intrinsically subserves contextual associations, not scenes or places per se," *J. Neurosci.*, vol. 28, no. 34, pp. 8539–8544, Aug. 2008.
- [13] G. Benton, M. Finzi, P. Izmailov, and A. G. Wilson, "Learning invariances in neural networks from training data," in *Proc. NeurIPS*, 2020, pp. 17605–17616.
- [14] E. Borenstein and S. Ullman, "Combined top-down/bottom-up segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2109–2125, Dec. 2008.
- [15] E. Bricolo, "On the representation of novel objects: human psychophysics, monkey physiology and computational models," Ph.D. dissertation, MIT, Cambridge, MA, USA, 1996.
- [16] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," 2021, [arXiv:2104.13478](https://arxiv.org/abs/2104.13478).
- [17] S. A. Cadena, G. H. Denfield, E. Y. Walker, L. A. Gatys, A. S. Tolias, M. Bethge, and A. S. Ecker, "Deep convolutional models improve predictions of macaque v1 responses to natural images," *PLOS Comput. Biol.*, vol. 15, no. 4, Apr. 2019, Art. no. e1006897.
- [18] L. Chang, B. Egger, T. Vetter, and D. Y. Tsao, "Explaining face representation in the primate brain using different computational models," *Current Biol.*, vol. 31, no. 13, pp. 2785–2795, Jul. 2021.
- [19] L. Chen, D. Wassermann, D. A. Abrams, J. Kochalka, G. Gallardo-Diez, and V. Menon, "The visual word form area (VWFA) is part of both language and attention circuitry," *Nature Commun.*, vol. 10, no. 1, p. 5601, Dec. 2019.
- [20] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn. (PMLR)*, Jul. 2020, pp. 1597–1607.
- [21] S. Chen, E. Dobriban, and J. H. Lee, "A group-theoretic framework for data augmentation," in *Proc. Adv. Inf. Process. Syst.*, vol. 33, 2020, pp. 21321–21333.
- [22] L. Cohen, S. Dehaene, and L. Naccache, "The visual word form area," *Brain*, vol. 123, no. 2, p. 291, 2000.
- [23] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 2990–2999.
- [24] T. S. Cohen, "Equivariant convolutional networks," Ph.D. thesis, Dept. Comput. Sci., Univ. Amsterdam, Amsterdam, The Netherlands, 2021.
- [25] T. Cohen and M. Welling, "Steerable CNNs," in *Proc. ICLR*, 2017.
- [26] N. Dehmamy, R. Walters, Y. Liu, D. Wang, and R. Yu, "Automatic symmetry discovery with lie algebra convolutional network," in *Proc. NeurIPS*, 2021.
- [27] D. Dennett, "The Baldwin effect: A crane, not a skyhook," in *Evolution and Learning: The Baldwin Effect Reconsidered*, B. H. Weber and D. J. Depew, Eds. Cambridge, MA, USA: MIT Press, 2003, pp. 69–106.
- [28] K. Dobs, J. Martinez, A. J. E. Kell, and N. Kanwisher, "Brain-like functional specialization emerges spontaneously in deep neural networks," *Sci Adv.*, vol. 8, no. 11, 2022, Art. no. eabl8913.
- [29] D. L. Donoho and P. B. Stark, "Uncertainty principles and signal recovery," *SIAM J. Appl. Math.*, vol. 49, no. 3, pp. 906–931, Jun. 1989.
- [30] P. E. Downing, Y. Jiang, M. Shuman, and N. Kanwisher, "A cortical area selective for visual processing of the human body," *Science*, vol. 293, no. 5539, pp. 2470–2473, Sep. 2001.
- [31] A. Deza, Q. Liao, A. Banburski, and T. Poggio, "Hierarchically compositional tasks and deep convolutional networks," 2020, [arXiv:2006.13915](https://arxiv.org/abs/2006.13915).
- [32] H. H. Bülthoff and S. Edelman, "Psychophysical support for a two-dimensional view interpolation theory of object recognition," *Proc. Nat. Acad. Sci. USA*, vol. 89, no. 1, pp. 60–64, Jan. 1992.
- [33] B. Eleseedy, "Group symmetry in PAC learning," in *Proc. Workshop Geometrical Topol. Represent. Learn. (ICLR)*, 2022.
- [34] R. Epstein and N. Kanwisher, "A cortical representation of the local visual environment," *Nature*, vol. 392, no. 6676, pp. 598–601, Apr. 1998.
- [35] P. Földiák, "Learning invariance from transformation sequences," *Neural Comput.*, vol. 3, no. 2, pp. 194–200, Jun. 1991.
- [36] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, Apr. 1980.
- [37] W. A. Freiwald and D. Y. Tsao, "Functional compartmentalization and viewpoint generalization within the macaque face-processing system," *Science*, vol. 330, no. 6005, pp. 845–851, Nov. 2010.
- [38] W. A. Freiwald, D. Y. Tsao, and M. S. Livingstone, "A face feature space in the macaque temporal lobe," *Nature Neurosci.*, vol. 12, no. 9, pp. 1187–1196, Sep. 2009.
- [39] W. Freiwald and H. Hosoya, "Neuroscience: A face's journey through space and time," *Current Biol.*, vol. 31, no. 1, pp. R13–R15, 2021.
- [40] D. Gabor, "Theory of communication. Part 1: The analysis of information," *J. Inst. Electr. Eng. III, Radio Commun. Eng.*, vol. 93, no. 26, pp. 429–441, Nov. 1946.
- [41] J. L. Gallant, J. Braun, and D. C. Van Essen, "Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex," *Science*, vol. 259, no. 5091, pp. 100–103, Jan. 1993.
- [42] J. L. Gallant, C. E. Connor, S. Rakshit, J. W. Lewis, and D. C. Van Essen, "Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey," *J. Neurophysiol.*, vol. 76, no. 4, pp. 2718–2739, Oct. 1996.
- [43] I. Gauthier and M. J. Tarr, "Becoming a 'Greeble' expert: Exploring mechanisms for face recognition," *Vis. Res.*, vol. 37, no. 12, pp. 1673–1682, Jun. 1997.
- [44] R. Gens and P. M. Domingos, "Deep symmetry networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2537–2545.
- [45] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014, [arXiv:1409.4842](https://arxiv.org/abs/1409.4842).
- [46] G. CG and J. Sergent, "Face recognition," *Current Opinion Neurobiol.*, vol. 2, pp. 156–161, Apr. 1992, doi: [10.1016/0959-4388\(92\)90004-5](https://doi.org/10.1016/0959-4388(92)90004-5).
- [47] H. Hasani, M. Soleymani, and H. Aghajan, "Surround modulation: A bio-inspired connectivity structure for convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 18335–18345.
- [48] U. Hasson, S. A. Nastase, and A. Goldstein, "Direct fit to nature: An evolutionary perspective on biological and artificial neural networks," *Neuron*, vol. 105, no. 3, pp. 416–434, 2020.
- [49] D. J. Heeger, "Normalization of cell responses in cat striate cortex," *Vis. Neurosci.*, vol. 9, no. 2, pp. 181–197, Aug. 1992.
- [50] D. O. Hebb, *The Organization of Behaviour: A Neuropsychological Theory*. Hoboken, NJ, USA: Wiley, 1949.
- [51] J. Hegd e and D. C. Van Essen, "Selectivity for complex shapes in primate visual area V2," *J. Neurosci.*, vol. 20, no. 5, p. RC61, Mar. 2000.
- [52] J. K. Hesse and D. Y. Tsao, "The macaque face patch system: A turtle's underbelly for the brain," *Nature Rev. Neurosci.*, vol. 21, no. 12, pp. 695–716, Dec. 2020.
- [53] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, no. 1, pp. 106–154, Jan. 1962.

- [54] D. H. Hubel and T. N. Wiesel, "Uniformity of monkey striate cortex: A parallel relationship between field size, scatter, and magnification factor," *J. Comparative Neurol.*, vol. 158, no. 3, pp. 295–305, Dec. 1974.
- [55] C. Hung, G. Kreiman, T. Poggio, and J. DiCarlo, "Fast readout of object identity from macaque inferior temporal cortex," *Science*, vol. 310, no. 5749, pp. 863–866, 2005.
- [56] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, nos. 4–5, pp. 411–430, Jun. 2000.
- [57] T. P. Lillicrap, A. Santoro, L. Marris, C. J. Akerman, and G. Hinton, "Backpropagation and the brain," *Nature Rev. Neurosci.*, vol. 21, no. 6, pp. 335–346, Jun. 2020.
- [58] H. L. Kosakowski, M. A. Cohen, A. Takahashi, B. Keil, N. Kanwisher, and R. Saxe, "Selective responses to faces, scenes, and bodies in the ventral visual pathway of infants," *Current Biol.*, vol. 32, no. 2, pp. 265–274, Jan. 2022.
- [59] N. Kriegeskorte, "Deep neural networks: A new framework for modeling biological vision and brain information processing," *Annu. Rev. Vis. Sci.*, vol. 1, no. 1, pp. 417–446, Nov. 2015.
- [60] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, CA, USA, vol. 25, 2012, pp. 1106–1114.
- [61] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex," *J. Neurophysiol.*, vol. 58, no. 6, pp. 1233–1258, Dec. 1987.
- [62] K. N. Kay, "Principles for models of neural information processing," *NeuroImage*, vol. 180, pp. 101–109, Oct. 2018.
- [63] N. Kanwisher, "Functional specificity in the human brain: A window into the functional architecture of the mind," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 25, pp. 11163–11170, Jun. 2010.
- [64] J. Karhunen, "Stability of Oja's PCA subspace rule," *Neural Comput.*, vol. 6, no. 4, pp. 739–747, Jul. 1994.
- [65] M. Kouh and T. Poggio, "A canonical neural circuit for cortical non-linear operations," *Neural Comput.*, vol. 20, no. 6, pp. 1427–1451, Jun. 2008.
- [66] A. Koulakov, S. Shuvaev, and A. Zador, "Encoding innate ability through a genomic bottleneck," *bioRxiv*, Jan. 2021, doi: 10.1101/2021.03.16.435261.
- [67] E. Y. Ko, J. Z. Leibo, and T. Poggio, "A hierarchical model of perspective-invariant scene identification," in *Proc. Soc. Neurosci.*, Washington, DC, USA, 2011, pp. 30–40.
- [68] D. Krotov and J. J. Hopfield, "Unsupervised learning by competing hidden units," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 16, pp. 7723–7731, Apr. 2019.
- [69] S. Ku, A. Tolia, N. Logothetis, and J. Goense, "fMRI of the face processing network in the ventral temporal lobe of awake and anesthetized macaques," *Neuron*, vol. 70, no. 2, pp. 352–362, 2011.
- [70] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.
- [71] Y. LeCun, F. Jie Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Dec. 2004, pp. 11097–11104.
- [72] Y. LeCun and Y. Bengio, "The handbook of brain theory and neural networks," in *Convolutional Networks for Images Speech Time Series*. Cambridge, MA, USA: MIT Press, 1995, pp. 255–258.
- [73] Q. V. Le, "Building high-level features using large scale unsupervised learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 8595–8598.
- [74] T. Lee and S. Soatto, "Video-based descriptors for object recognition," *Image Vis. Comput.*, vol. 29, no. 10, pp. 639–652, Sep. 2011.
- [75] J. Z. Leibo, J. Mutch, and T. Poggio, "Why the brain separates face recognition from object recognition," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Granada, Spain, 2011, pp. 9950–9965.
- [76] J. Z. Leibo, J. Mutch, and T. Poggio, "How can cells in the anterior medial face patch be viewpoint invariant," presented at the COSYNE, Salt Lake City, UT, USA, 2011.
- [77] J. Z. Leibo, Q. Liao, F. Anselmi, and T. Poggio, "The invariance hypothesis implies domain-specific regions in visual cortex," *PLoS Comput. Biol.*, vol. 11, no. 10, 2015.
- [78] J. Z. Leibo, F. Anselmi, J. Mutch, A. F. Ebiyara, W. Freiwald, and T. Poggio, "View-invariance and mirror-symmetric tuning in a model of the macaque face-processing system," in *Proc. Comput. Syst. Neurosci.*, Salt Lake City, UT, USA, 2013.
- [79] K. Lenc and A. Vedaldi, "Understanding image representations by measuring their equivariance and equivalence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 456–476.
- [80] M. Lessmann and R. P. Würtl, "Learning invariant object recognition from temporal correlation in a hierarchical network," *Neural Netw.*, vol. 54, pp. 70–84, Jun. 2014.
- [81] Q. Liao, J. Z. Leibo, and T. Poggio, "Learning invariant representations and applications to face verification," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Lake Tahoe, NV, USA, 2013.
- [82] N. K. Logothetis, J. Pauls, H. H. Bülthoff, and T. Poggio, "View-dependent object recognition by monkeys," *Current Biol.*, vol. 4, no. 5, pp. 401–414, May 1994.
- [83] N. K. Logothetis, J. Pauls, and T. Poggio, "Shape representation in the inferior temporal cortex of monkeys," *Current Biol.*, vol. 5, no. 5, pp. 552–563, May 1995.
- [84] N. K. Logothetis and D. L. Sheinberg, "Visual object recognition," *Annu. Rev. Neurosci.*, vol. 19, pp. 577–621, Oct. 1996.
- [85] R. Malach, J. B. Reppas, R. R. Benson, K. K. Kwong, H. Jiang, W. A. Kennedy, P. J. Ledden, T. J. Brady, B. R. Rosen, and R. B. Tootell, "Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex," *Proc. Nat. Acad. Sci. USA*, vol. 92, no. 18, pp. 8135–8139, Aug. 1995.
- [86] D. Marr and T. Poggio, "From understanding computation to understanding neural circuitry," *Res. Program. Bull.*, vol. 15, pp. 470–488, Jun. 1976.
- [87] H. N. Mhaskar and T. Poggio, "Deep vs. shallow networks: An approximation theory perspective," *Anal. Appl.*, vol. 14, no. 6, pp. 829–848, 2016.
- [88] B. W. Mel, "SEEMORE: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition," *Neural Comput.*, vol. 9, no. 4, pp. 777–804, May 1997.
- [89] T. Micconi, "Hebbian learning with gradients: Hebbian convolutional neural networks with modern deep learning frameworks," 2021, *arXiv:2107.01729*.
- [90] J. Mutch and D. G. Lowe, "Multiclass object recognition with sparse, localized features," in *Proc. Comput. Vis. Pattern Recognit.*, 2006, pp. 11–18.
- [91] J. Mutch, J. Z. Leibo, S. Smale, L. Rosasco, and T. Poggio, "Neurons that confuse mirror-symmetric object views," MIT, Cambridge, MA, USA, Tech. Rep. MIT-CSAIL-TR-2010-062, CBCL-295, 2010.
- [92] J. Mutch, F. Anselmi, A. Tacchetti, L. Rosasco, J. Z. Leibo, and T. Poggio, "Invariant recognition predicts tuning of neurons in sensory cortex," in *Proc. Comput. Cogn. Neurosci. Vis.*, 2017, pp. 85–104.
- [93] C. M. Niell and M. P. Stryker, "Highly selective receptive fields in mouse visual cortex," *J. Neurosci.*, vol. 28, no. 30, pp. 7520–7536, Jul. 2008.
- [94] P. Niyogi, F. Girosi, and T. Poggio, "Incorporating prior information in machine learning by creating virtual examples," *Proc. IEEE*, vol. 86, no. 11, pp. 2196–2209, 1998.
- [95] E. Oja, "Simplified neuron model as a principal component analyzer," *J. Math. Biol.*, vol. 15, no. 3, pp. 267–273, Nov. 1982.
- [96] E. Oja, "Principal components, minor components, and linear neural networks," *Neural Netw.*, vol. 5, no. 6, pp. 927–935, Nov. 1992.
- [97] D. I. Perrett and M. W. Oram, "Neurophysiology of shape processing," *Image Vis. Comput.*, vol. 11, no. 6, pp. 317–333, Jul. 1993.
- [98] N. Pinto, J. M. DiCarlo, J. James, and D. D. Cox, "How far can you get with a modern face recognition test set using only simple features?" in *Proc. Comput. Vis. Pattern Recognit.*, 2009, pp. 2591–2598.
- [99] D. Pitcher, G. Ianni, and L. G. Ungerleider, "A functional dissociation of face-, body- and scene-selective brain areas based on their response to moving and static stimuli," *Sci. Rep.*, vol. 9, no. 1, p. 8242, Dec. 2019.
- [100] T. Poggio and S. Edelman, "A network that learns to recognize three-dimensional objects," *Nature*, vol. 343, no. 6255, pp. 263–266, Jan. 1990.
- [101] T. Poggio, J. Mutch, F. Anselmi, A. Tacchetti, L. Rosasco, and J. Z. Leibo, "Does invariant recognition predict tuning of neurons in sensory cortex?" Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep. MIT-CSAIL-TR-2013-019, CBCL-313, 2013.
- [102] T. Poggio, J. Mutch, and L. Isik, "Computational role of eccentricity dependent cortical magnification," 2014, *arXiv:1406.1770*.
- [103] T. Poggio and W. Reichardt, "Considerations on models of movement detection," *Kybernetik*, vol. 13, no. 4, pp. 223–227, Dec. 1973.

- [104] M. C. Potter, "Meaning in visual search," *Science*, vol. 187, no. 4180, pp. 965–966, Mar. 1975.
- [105] T. Poggio, A. Banburski, and Q. Liao, "Theoretical issues in deep networks," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 48, pp. 30039–30045, Dec. 2020.
- [106] M. Riesenhuber and T. Poggio, "Models of object recognition," *Nature Neurosci.*, vol. 3, no. 11, pp. 1199–1204, 2000.
- [107] D. L. Ringach, "Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex," *J. Neurophysiol.*, vol. 88, no. 1, pp. 455–463, Jul. 2002.
- [108] R. Rodriguez, E. Dokladalova, and P. Dokladal, "Rotation invariant CNN using scattering transform for image classification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 654–658.
- [109] E. T. Rolls, "Learning invariant object and spatial view representations in the brain using slow unsupervised learning," *Frontiers Comput. Neurosci.*, vol. 15, Jul. 2021.
- [110] E. T. Rolls, *Brain Computations: What and How*. Oxford, U.K.: Oxford Univ. Press, 2021.
- [111] D. Ruderman, "The statistics of natural images," *Netw., Comput. Neural Syst.*, vol. 5, pp. 517–548, Dec. 1994.
- [112] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," 2014, *arXiv:1409.0575*.
- [113] T. D. Sanger, "Optimal unsupervised learning in a single-layer linear feedforward neural network," *Neural Netw.*, vol. 2, no. 6, pp. 459–473, Jan. 1989.
- [114] S. Saremi and T. J. Sejnowski, "Hierarchical model of natural images and the origin of scale invariance," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 8, pp. 3071–3076, Feb. 2013.
- [115] A. M. Saxe, M. Bhand, R. Mudur, B. Suresh, and A. Y. Ng, "Unsupervised learning models of primary cortical receptive fields and receptive field plasticity," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 1971–1979.
- [116] D. L. Schacter and D. R. Addis, "On the nature of medial temporal lobe contributions to the constructive simulation of future events," *Phil. Trans. Roy. Soc. B, Biol. Sci.*, vol. 364, no. 1521, pp. 1245–1253, May 2009.
- [117] B. R. Sheth and R. Young, "Two visual pathways in primates based on sampling of space: Exploitation and exploration of visual information," *Frontiers Integrative Neurosci.*, vol. 10, pp. 10–37, Nov. 2016.
- [118] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, Mar. 2007.
- [119] J. Sokolic, G. Giryas, G. Sapiro, and M. Rodrigues, "Generalization error of invariant classifiers," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2017, pp. 1094–1103.
- [120] J. Sohl-Dickstein, C. Ming Wang, and B. A. Olshausen, "An unsupervised algorithm for learning lie group transformations," 2010, *arXiv:1001.1027*.
- [121] C. F. Stevens, "Preserving properties of object shape by computations in primary visual cortex," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 43, pp. 15524–15529, Oct. 2004.
- [122] S. M. Stringer and E. T. Rolls, "Invariant object recognition in the visual system with novel views of 3D objects," *Neural Comput.*, vol. 14, no. 11, pp. 2585–2596, Nov. 2002.
- [123] M. P. Stryker, "Temporal associations," *Nature*, vol. 354, pp. 108–109, Jan. 1991.
- [124] M. J. Tarr, "Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects," *Psychonomic Bull. Rev.*, vol. 2, no. 1, pp. 55–82, Mar. 1995.
- [125] M. J. Tarr and I. Gauthier, "FFA: A flexible fusiform area for subordinate-level visual processing automatized by expertise," *Nature Neurosci.*, vol. 3, no. 8, pp. 764–769, Aug. 2000.
- [126] S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381, no. 6582, pp. 520–522, Jun. 1996.
- [127] A. Torralba and A. Oliva, "Statistics of natural image categories," *Network: Comput. Neural Syst.*, vol. 14, no. 3, pp. 391–412, Jan. 2003.
- [128] D. Tsao and W. Freiwald, "Faces and objects in macaque cerebral cortex," *Nature*, vol. 6, no. 9, pp. 989–995, 2003.
- [129] D. Y. Tsao and M. S. Livingstone, "Mechanisms of face perception," *Annu. Rev. Neurosci.*, vol. 31, no. 1, pp. 411–437, Jul. 2008.
- [130] G. G. Turrigiano and S. B. Nelson, "Homeostatic plasticity in the developing nervous system," *Nature Rev. Neurosci.*, vol. 5, no. 2, pp. 97–107, Feb. 2004.
- [131] S. Ullman and R. Basri, "Recognition by linear combinations of models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 10, pp. 992–1006, Aug. 1991.
- [132] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 23, pp. 8619–8624, Jun. 2014.
- [133] D. L. K. Yamins and J. J. DiCarlo, "Using goal-driven deep learning models to understand sensory cortex," *Nature Neurosci.*, vol. 19, no. 3, pp. 356–365, Mar. 2016.
- [134] M. van der Wilk, M. Bauer, S. John, and J. Hensman, "Learning invariances using the marginal likelihood," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9960–9970.
- [135] W. Vanduffel, Q. Zhu, and G. A. Orban, "Monkey cortex through fMRI glasses," *Neuron*, vol. 83, no. 3, pp. 533–550, Aug. 2014.
- [136] S. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
- [137] K. S. Weiner and K. Grill-Spector, "The evolution of face processing networks," *Trends Cognit. Sci.*, vol. 19, no. 5, pp. 240–241, May 2015, doi: [10.1016/j.tics.2015.03.010](https://doi.org/10.1016/j.tics.2015.03.010).
- [138] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, 2014, pp. 818–833.
- [139] C. Zhuang, S. Yan, A. Nayebi, M. Schrimpf, M. C. Frank, J. J. DiCarlo, and D. L. K. Yamins, "Unsupervised neural network models of the ventral visual stream," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 3, pp. 1–11, Jan. 2021.



FABIO ANSELM was born in Tregnago, Verona, Italy, in 1976. He received the B.S. and M.S. degrees in theoretical physics from the University of Padova, in 2000, and the Ph.D. degree in quantum physics from Hertfordshire University, U.K., in 2004. Since 2020, he has been an Assistant Professor at the Baylor College of Medicine in Houston, TX, USA. His research interests include the edge between machine learning, in particular deep learning, and computational neuroscience.

In 2016, together with Tomaso Poggio he published the book *Visual Cortex and Deep Networks: Learning Invariant Representations*.



TOMASO POGGIO is currently an Eugene McDermott Professor with the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, where he is also an Investigator at the McGovern Institute for Brain Research. He has been a member of the MIT Computer Science and Artificial Intelligence Laboratory and the Director of the Center for Biological and Computational Learning, MIT, and the Center for Brains, Minds, and Machines, a multi-institutional collaboration headquartered at the McGovern Institute, since 2013. He is also a Core Founding Scientific Advisor for the MIT Quest for Intelligence. He introduced regularization as a mathematical framework to approach the ill-posed problems of vision and the key problem of learning from data. He received the 2009 Okawa Prize and the sixth Pattern Analysis and Machine Intelligence Azriel Rosenfeld Lifetime Achievement Award. His research interests include interdisciplinary, between brains and computers, and currently include the mathematics of deep learning and the computational neuroscience of the visual cortex.

...