# DEEP LEARNING TO SEE

## TOWARDS NEW FOUNDATIONS OF COMPUTER VISION

Alessandro Betti, Marco Gori, and Stefano Melacci
SAILab - University of Siena

forthcoming book (Springer)

To All People Who Love To Ask Questions

ACDL-2021

# OUTLINE

1. Motion is all what you need
2. Focus of attention
3. Principles of motion invariance
4. Foveate neural networks
5. Information-based laws of learning
6. Epilogue

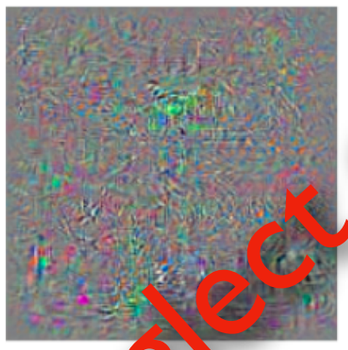
# MOTION
# IS ALL WHAT YOU NEED

MOTIVATIONS

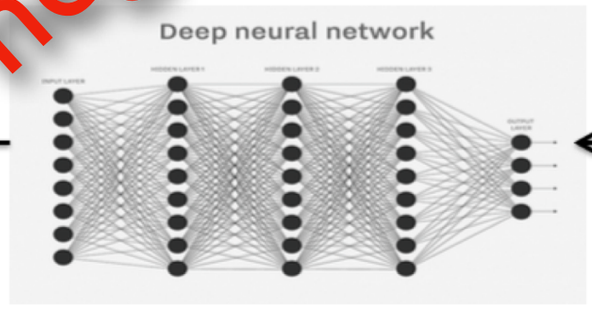
There's not a morning I begin
without a thousand questions running through my mind
...The reason why a bird was given wings
If not to fly, and praise the sky ...
From Yentl, "Where is it Written?" - I.B. Singer, The Yeshiva Boy

# ADVERSARIAL ATTACKS

school bus



Ostrich




Can we really neglect this issue?

# Object Identification And Affordance

my own chair vs "chair concept"

# Motivations For New Theories Of Learning

We have been facing problems harder than those we find in Nature!

The challenge:
Intelligent Agents which interact with their own environment just like humans

**NeurIPS-2020**

# Can Animals See In A World Of Shuffled Frames?

We have been facing problems more difficult

than those posed by Nature!

Maybe time does matter!

# 10 Questions For A Theory Of Vision

**Q1** How can animals conquer visual skills without requiring "intensive supervision"?

**Q2** How can animals gradually conquer visual skills in their own environments?

**Q3** Could children really acquire visual skills in such an artificial world, which is the one we are presenting to machines? Doesn't shuffled visual frames increase the complexity of learning to see?

**Q4** How can humans exhibit such an impressive skill of properly labelling single pixels without having received explicit pixel-wise supervisions? Isn't the case that such a skill is a sort of "visual primitive" that cannot be ignored for efficiently conquering additional skills on object recognition and scene interpretation?

**Q5** Why are the visual mainstreams in the brain of primates organized according to a hierarchical architecture with receptive fields? Is there any reason why this solution has been developed in biology?

# 10 Questions For A Theory Of Vision

**Q6** Why are there two different mainstreams in the primates' brain? What are the reasons for such a different neural evolution?

**Q7** Why do primates and other animals focus attention, whereas others, like the frog, do not?

**Q8** What are the mechanisms that drive eye movements?

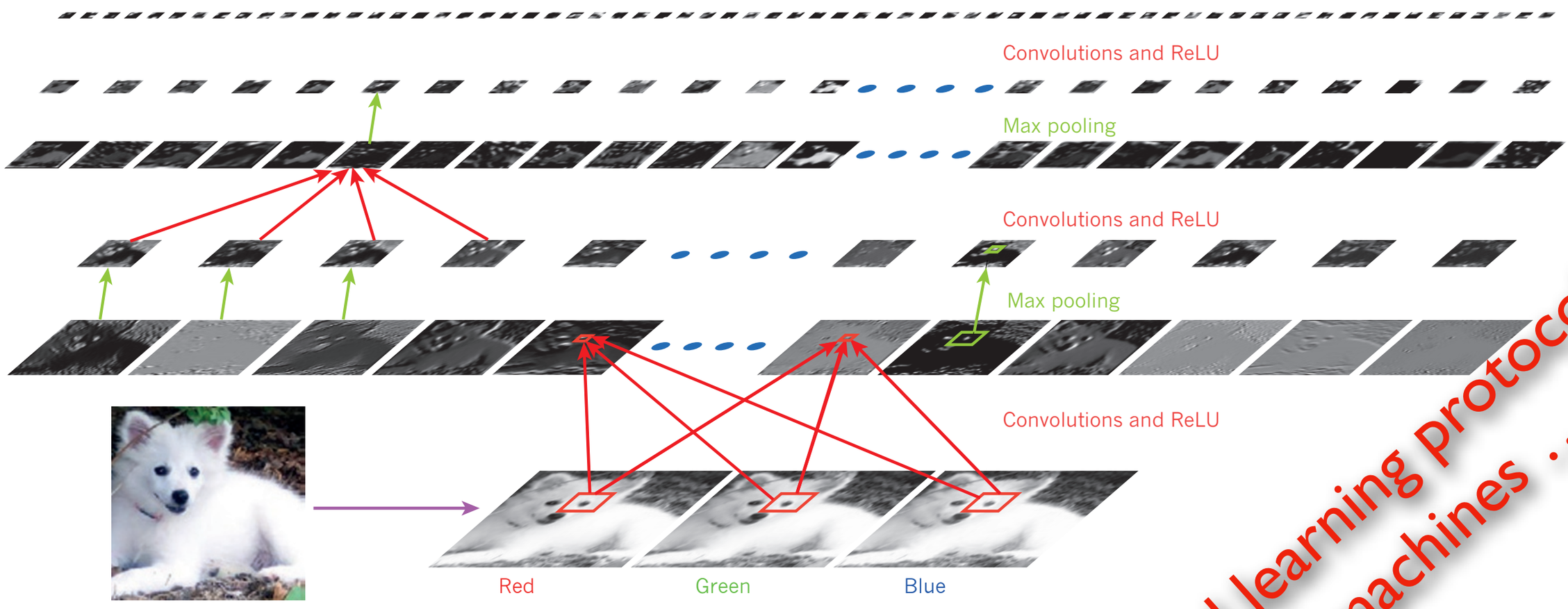
**Q9** Why does it take 8-12 months for newborns to achieve adult visual acuity? Is the development of adult visual acuity a biological issue or does it come from higher level computational laws of vision?

**Q10** How can we develop "linguistic focusing mechanisms" that can drive the process of object recognition?

# DOMINANT ROLE OF SUPERVISED LEARNING

## Deep learning Nature, 2015

Yann LeCun[1,2], Yoshua Bengio[3] & Geoffrey Hinton[4,5]



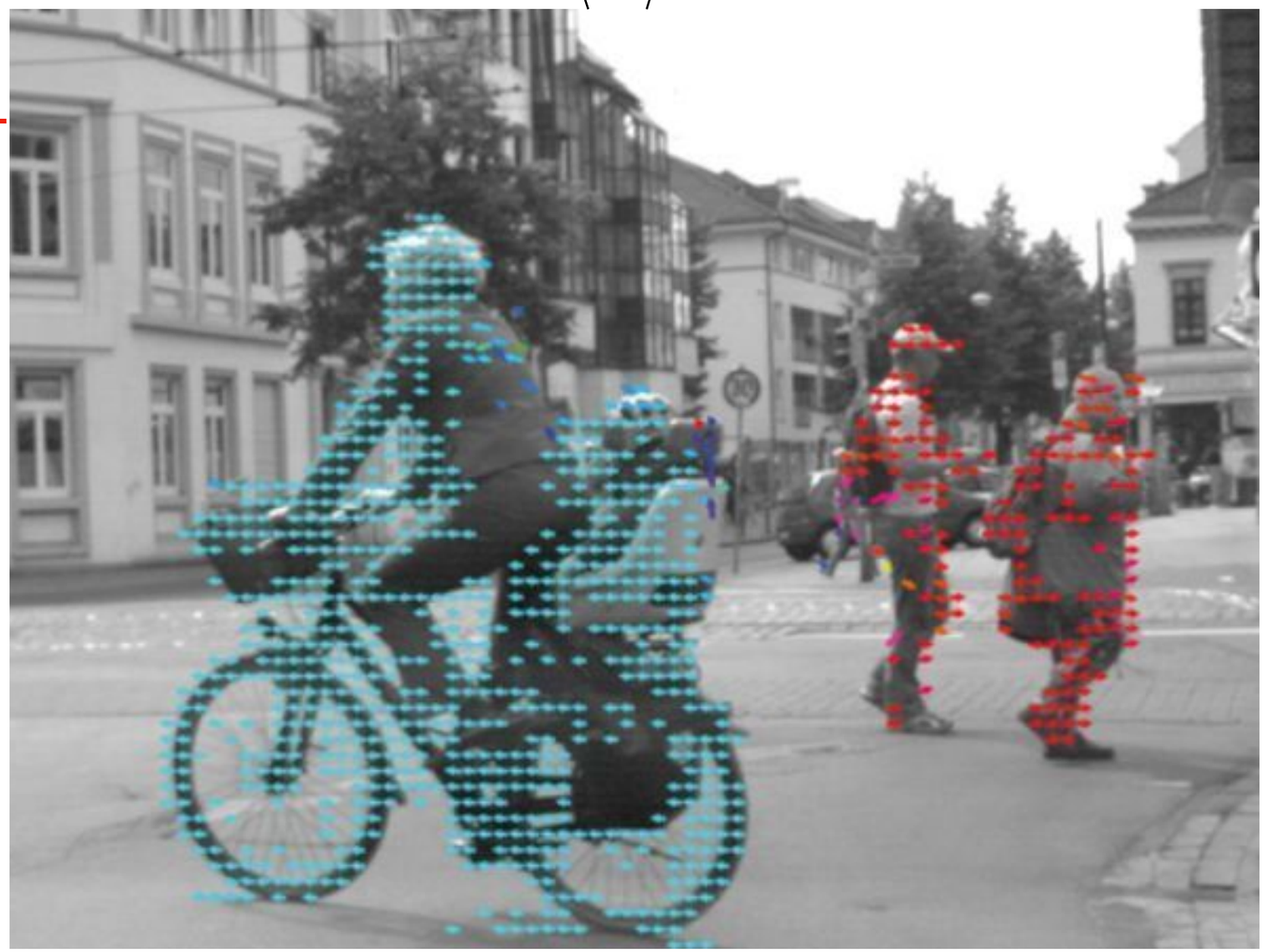

it's an artificial learning protocol
battlefield of machines ...

# MOTION IS ALL WHAT YOU NEED

Truly artificial protocol


motion information

How animals learn ...



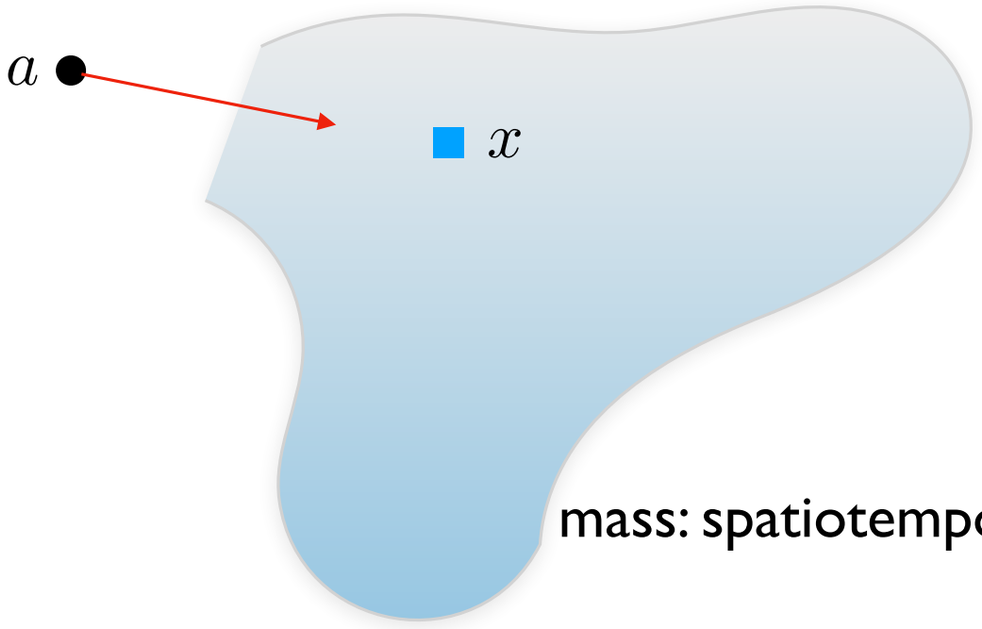
# FOCUS OF ATTENTION

The frog does not seem to see or, at any rate, is not concerned with the detail of stationary parts of the world around him. He will starve to death surrounded by food if it is not moving. His choice of food is determined only by size and movement.
*J. Y. Lettvin, H. R. Maturana, W. S. McCulloch, and W. H. Pitts, 1959*

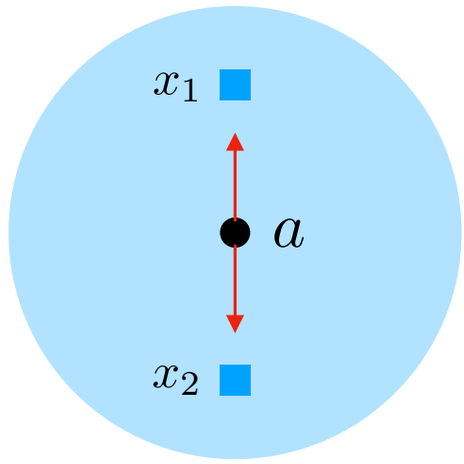
# FOCUS OF ATTENTION

## INFORMATION-BASED ATTRACTION

$$e(a,x) = \frac{1}{2\pi}\frac{x-a}{\|x-a\|^2}$$




$$\|e(a,x_1)\| = \|e(a,x_2)\|$$




mass: spatiotemporal gradient of brightness

$A$ $B$

# FOCUS OF ATTENTION

## INFORMATION-BASED ATTRACTION

virtual mass: spatiotemporal gradient
of the brightness

$$\begin{cases} \ddot{a}(t) + \varpi \dot{a}(t) + \nabla \varphi^0(a(t), t) = 0; \\ a(0) = a_0; \\ \dot{a}(0) = a_1, \end{cases}$$

$$-\nabla^2 \varphi = \mu$$

$$\varphi^0(x,t) := \frac{1}{2\pi} \int_{\mathbb{R}^2} \log \frac{1}{\|x-y\|} \mu(y,t)\, dy$$

inhibition of return

$$\mu(x,t) = \mu_1(x,t)\,(1 - I(x,t)) + \mu_2(x,t)$$

$$I_t + \beta I = \beta \exp(-\|x - a(t)\|^2 / 2\sigma^2)$$

# FOCUS OF ATTENTION

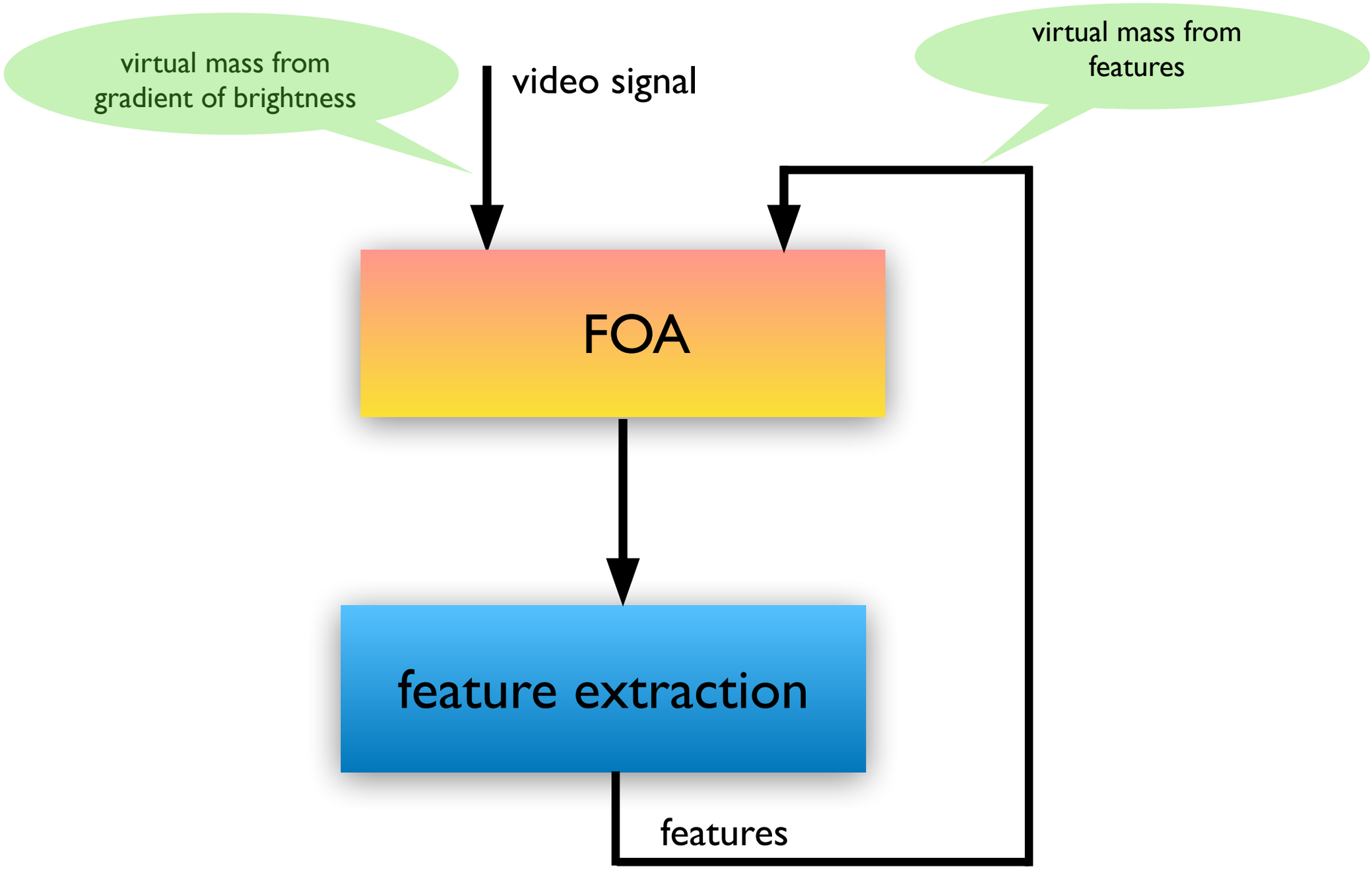
## INFORMATION-BASED ATTRACTION

$$\begin{cases} c^{-1}\varphi_t = \nabla^2\varphi + \mu & \text{in} \quad \mathbb{R}^2 \times (0, +\infty); \\ \varphi(x, 0) = 0, & \text{in} \quad \mathbb{R}^2 \times \{t = 0\}, \end{cases}$$

$$\begin{cases} c^{-2}\varphi_{tt} = \nabla^2\varphi + \mu & \text{in} \quad \mathbb{R}^2 \times (0, +\infty); \\ \varphi(x, 0) = 0, \quad \varphi_t(x, 0) = 0 & \text{in} \quad \mathbb{R}^2 \times \{t = 0\}, \end{cases}$$

$$\begin{cases} \gamma\varphi_{tt}(x, t) + \lambda\varphi_t(x, t) = c^2\nabla^2\varphi(x, t) + \mu(x, t) & \text{in} \quad \mathbb{R}^2 \times (0, +\infty); \\ \varphi(x, 0) = 0, \quad \varphi_t(x, 0) = 0 & \text{in} \quad \mathbb{R}^2 \times \{t = 0\}, \end{cases}$$

# VIRTUOUS LOOP OF FOCUS OF ATTENTION (FOA)



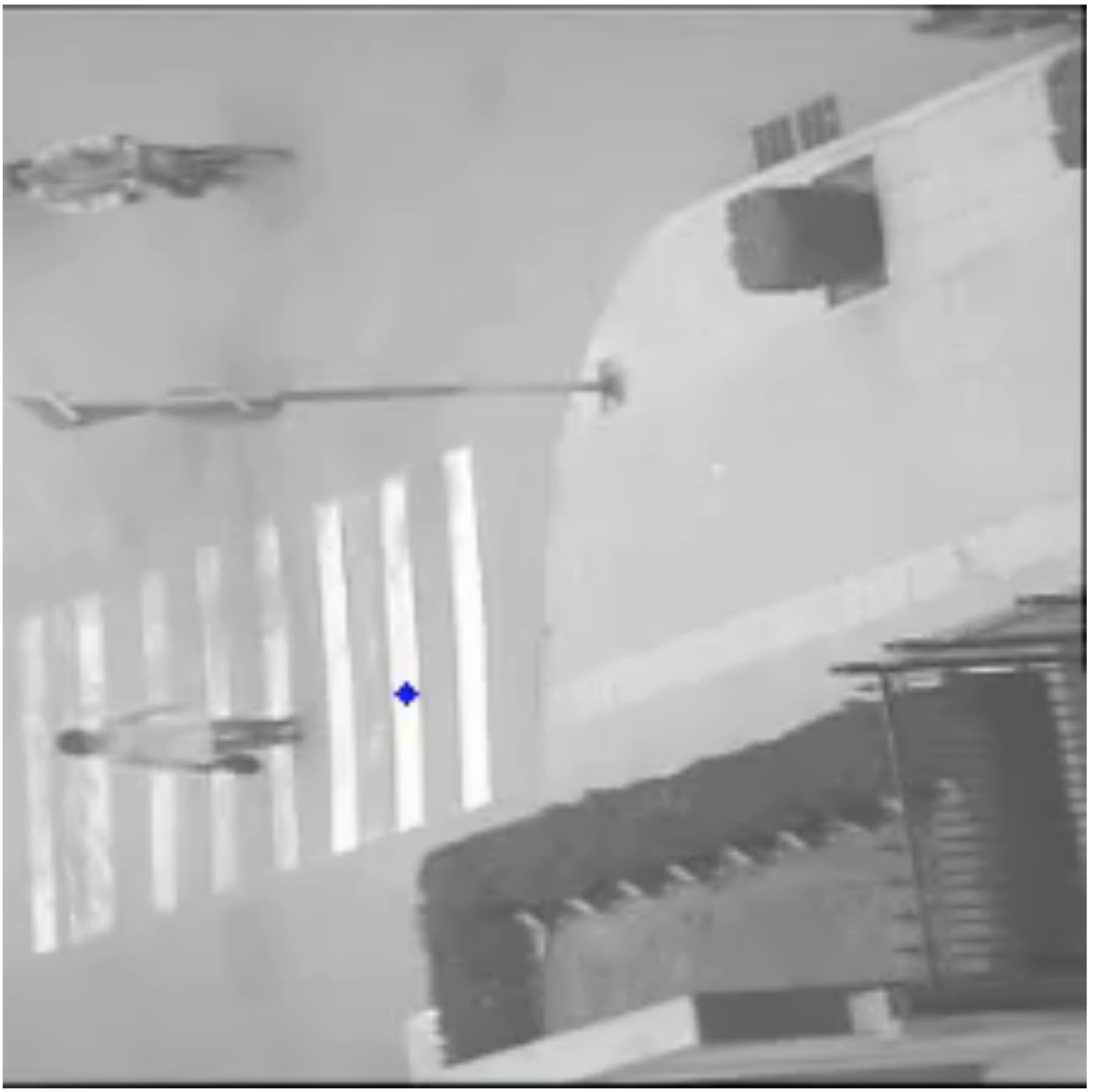
# FOCUS OF ATTENTION

## THERE'S ALWAYS MOTION!

focus of attention in visual scenes

# Variational Laws of Focus of Attention
# SAILab, NeurIPS 2017, TPAMI 2020

EVERYTHING IS MOVING ...



# PRINCIPLES OF MOTION INVARIANCE

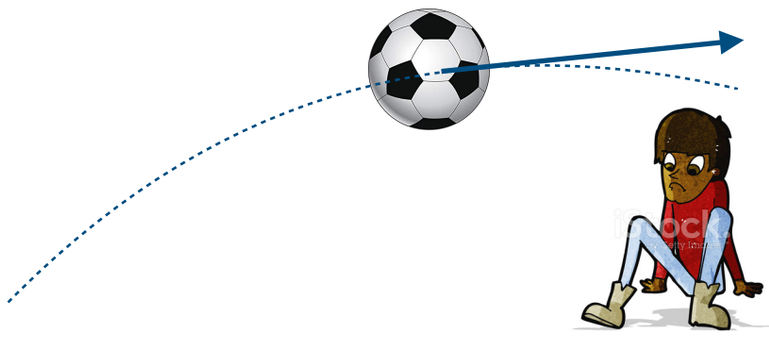
In science there is and will remain a Platonic element which could not be taken away without ruining it. Among the infinite diversity of singular phenomena science can only look for invariants.

Jacques Monod, 1971

# WHICH MOTION?

reference in the eyes!

No distinction between surveillance and egocentric vision



roof, tree, grass: same velocity

# PIXEL-BASED COMPUTATION

How can humans attach semantic labels at pixel level?


linguistic issues

# THE APPROPRIATE WINDOW

## is it well-posed?


linguistic issues

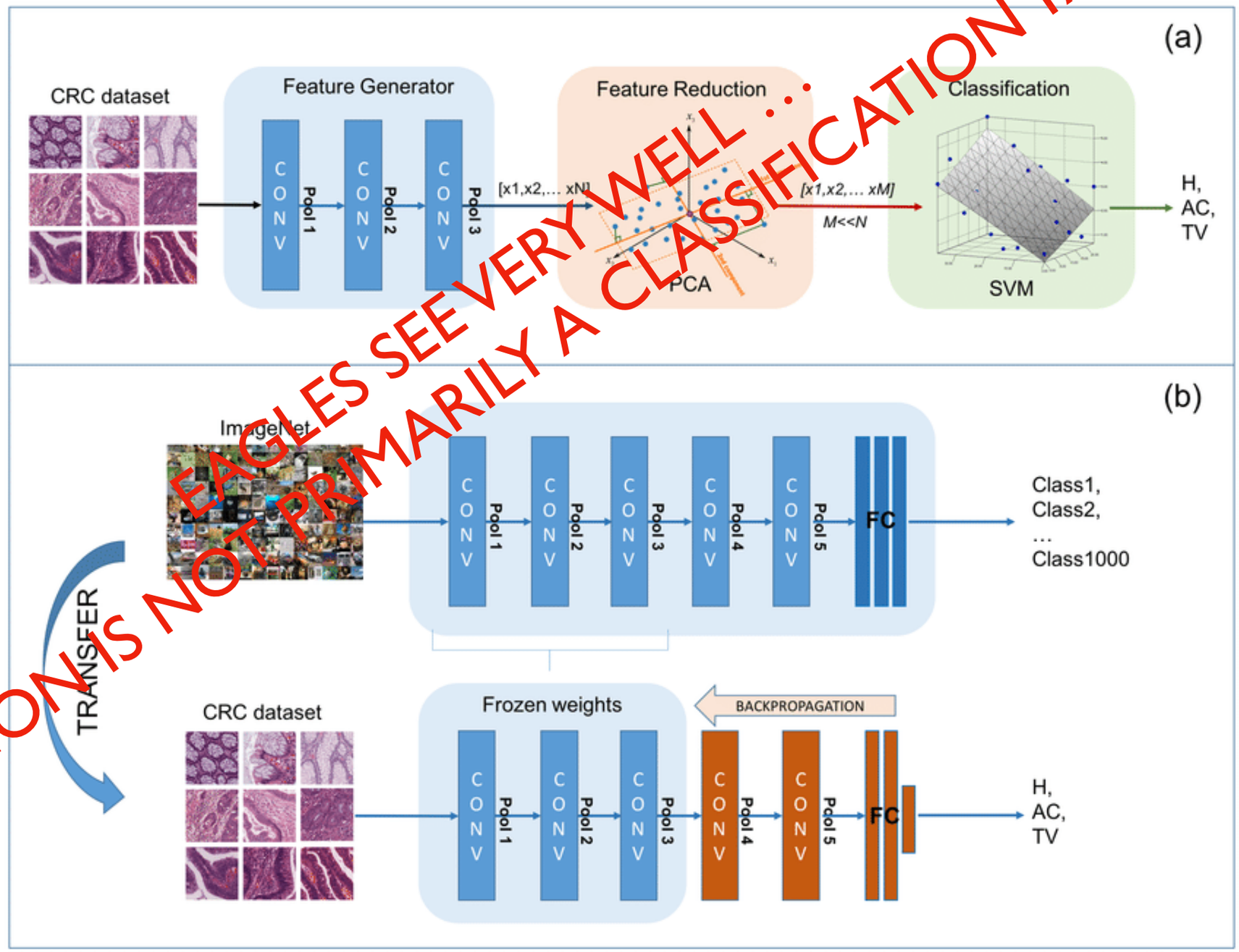
Who knows what is the appropriate context?

The chicken or the egg dilemma …

# FEATURE-BASED COMPUTATION

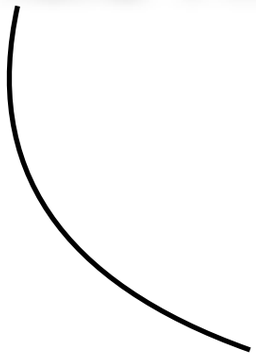
## Transfer learning

by Francesco Ponzio, Enrico Macii, Elisa Ficarra and Santa Di Cataldo



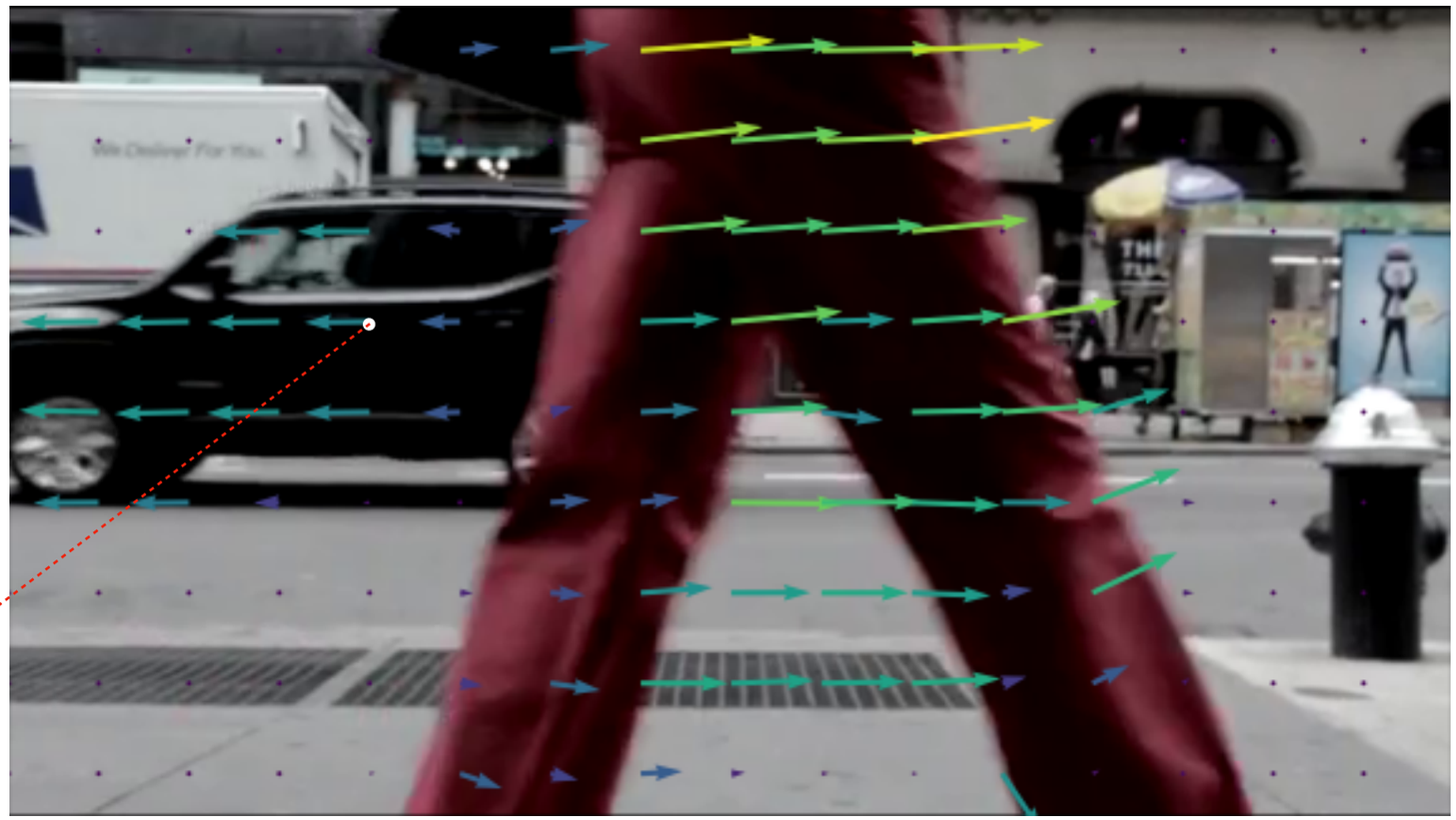
# THE MARRIAGE OF PERCEPTION & ACTION

Motion invariance principle
vision (Field Theory)



predictions, and action skills

# From Material Point To Pixels

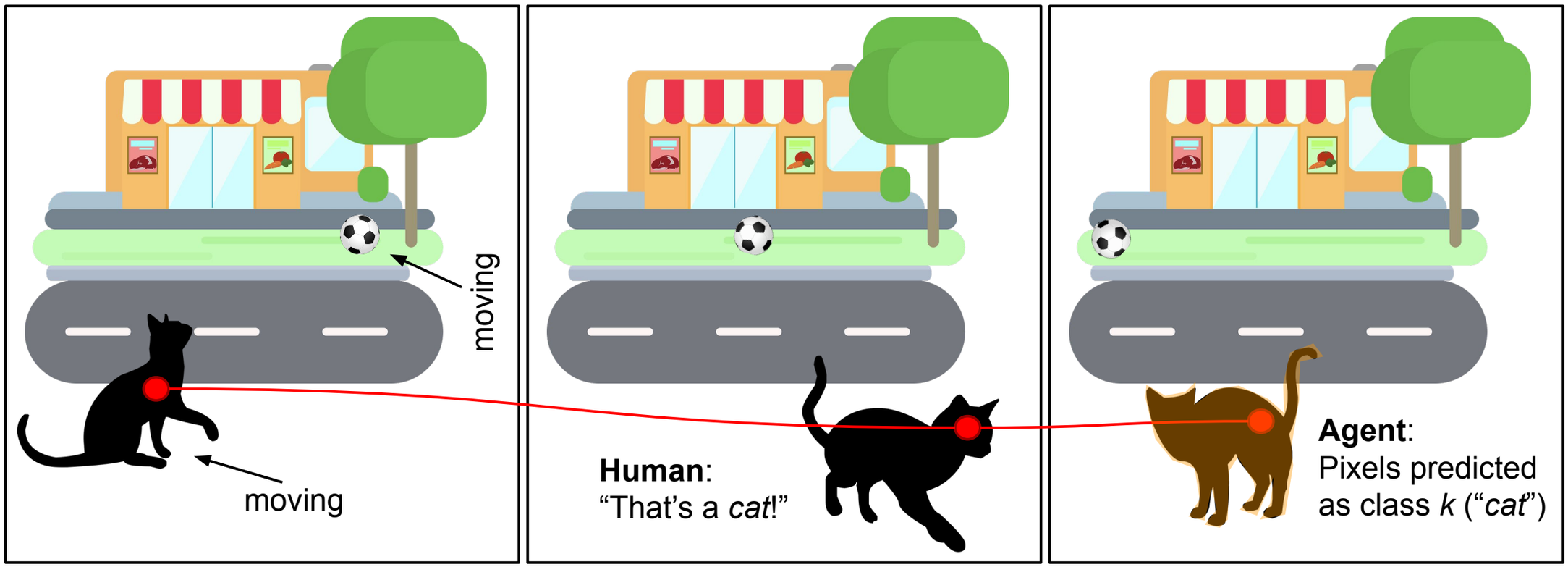



$$\forall t \in (0,T): \quad \frac{db(x(t),t)}{dt} = \nabla_x b \cdot v + \frac{\partial b}{\partial t} = 0$$

$$E(v_1, v_2) = \int_\Omega (\nabla_x v_1)^2 + (\nabla_x v_2)^2$$

# CONSISTENT DECISIONS

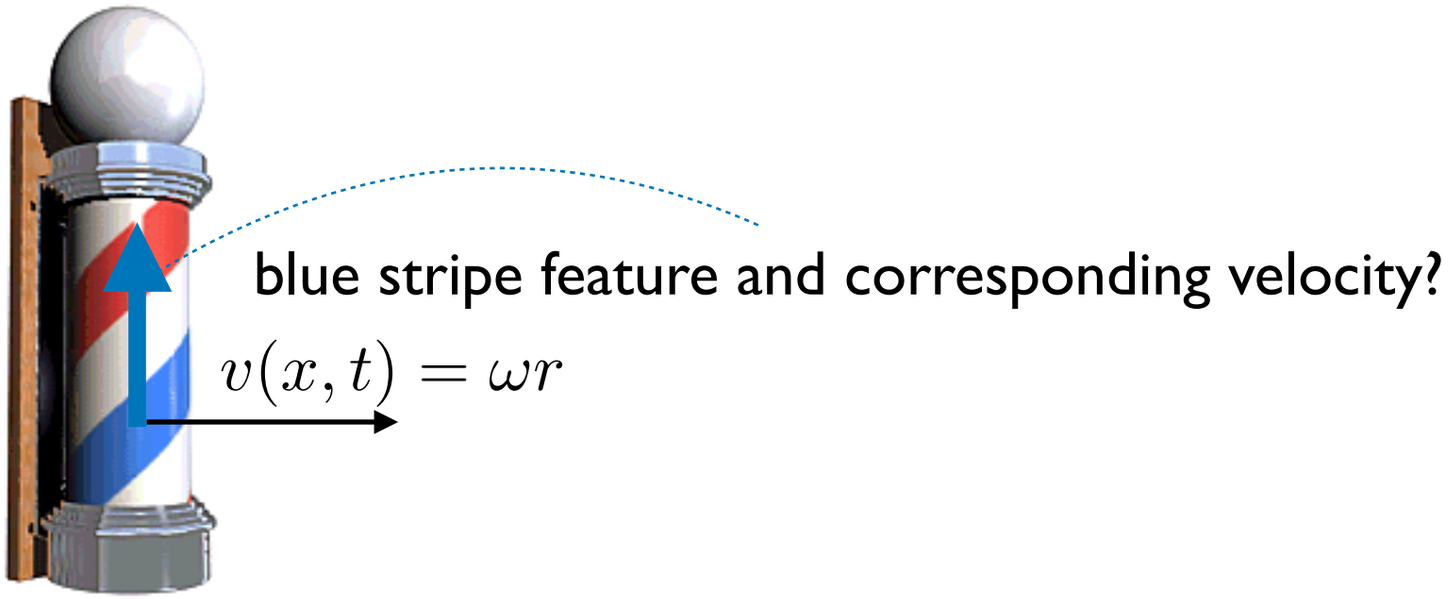
## Material Point Invariance (MPI)




It's a cat … I must take a consistent decision

# FROM MATERIAL POINTS TO PIXELS

The barber pole illusion




Couldn’t be interesting to think
of features and of their own “velocity”?

# I PRINCIPLE OF PERCEPTUAL VISION

## Material Point Invariance (MPI)

$$\forall (x,t) \in \Omega \times [0,T) : \quad \varphi(x_\varphi(t),t) = \varphi(x_\varphi(0),0) = c_\varphi$$

$$\varphi \bowtie v_\varphi := \frac{d\varphi(x(t),t)}{dt} = \nabla_x \varphi \cdot v_\varphi + \frac{\partial \varphi}{\partial t} = 0$$

we say that $\varphi$ is a *conjugate feature* with respect to $v_\varphi$



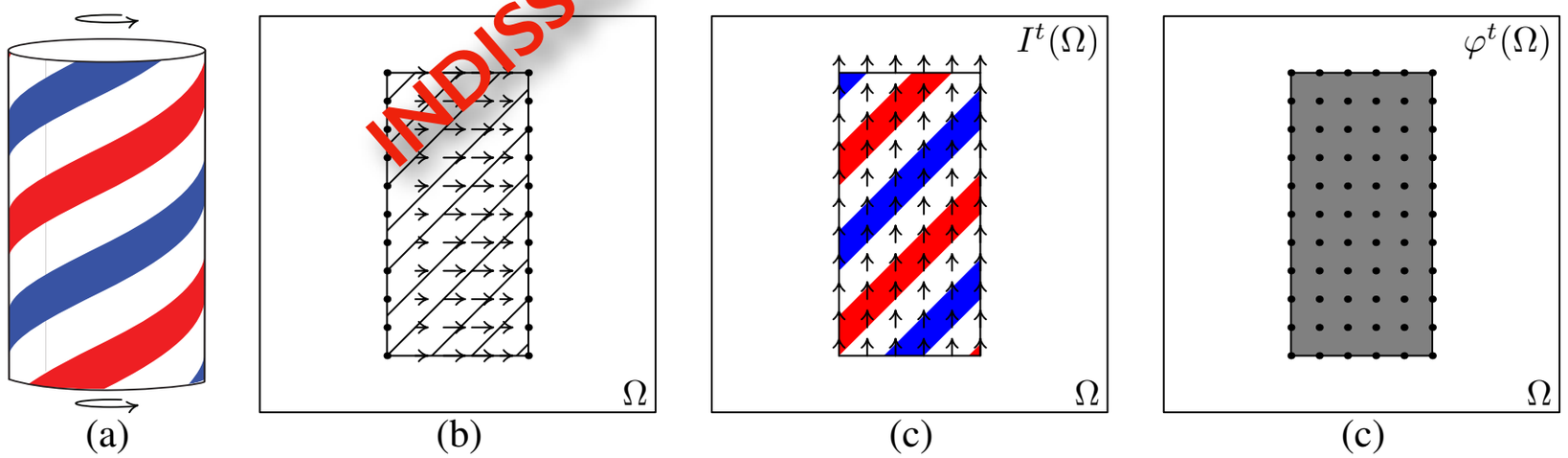
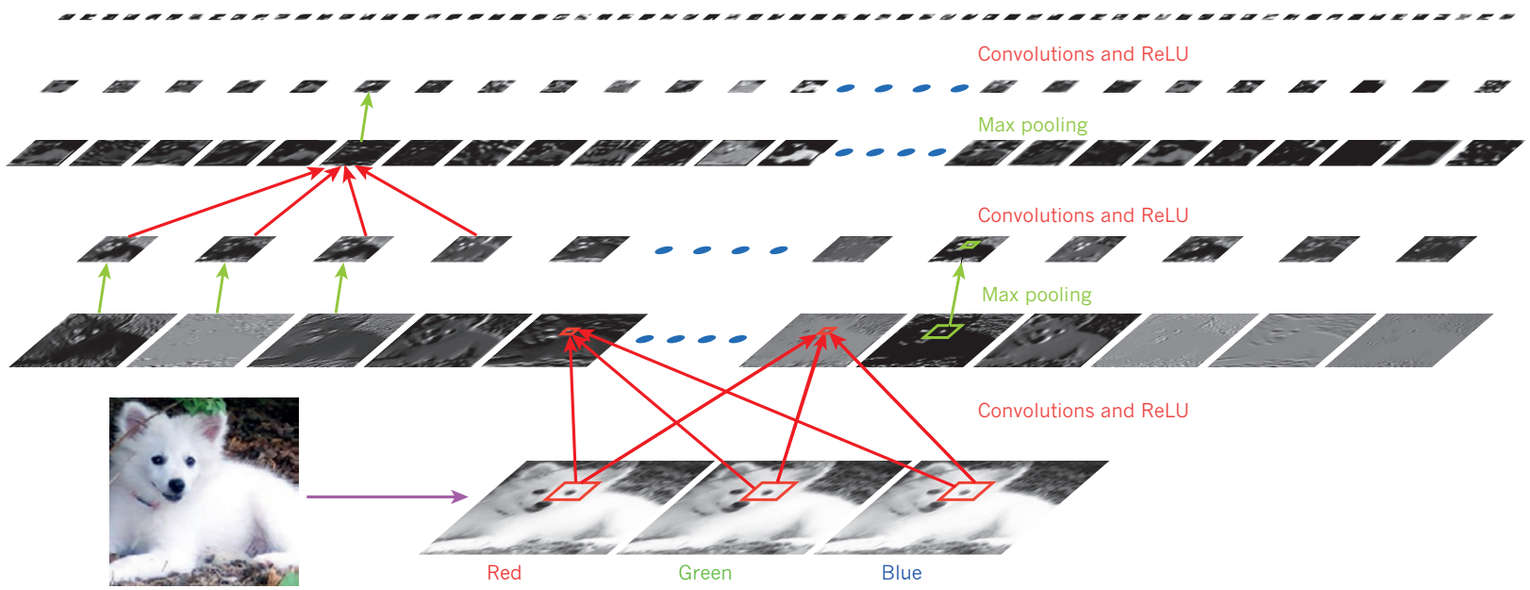



Fig. 3.1: Barber's pole example. (a) The 3-D object spinning counterclockwise.(b) The 2-D projection of the pole and the projected velocity on the retina $X$. (c) The brightness of the image and its optical flow pointing upwards. (d) A feature map that respond to the object and its conjugate (zero) optical flow.

# FEATURE GROUPING




Feature grouping is based on the correspondent receptive field

same velocity at a given layer

$$\forall i = 1, \ldots, m : \ \varphi_i \bowtie v = 0$$ conjugation with the same velocity

conjugate space ↓

$$\phi = \left\{ \varphi : \quad \varphi = \sum_{i=1}^{m} \alpha_i \varphi_i \right\}$$

# Regularization Effect Of Feature Grouping

$$\forall i = 1, \ldots, m : \ \varphi_i \bowtie v = 0$$

$$\nabla_x (\varphi_1 \ldots \varphi_m)' \cdot v + \frac{\partial}{\partial t} (\varphi_1 \ldots \varphi_m)' = 0$$

$\mathbb{R}^{m,2}$

Be careful:

$$\operatorname{rank} \nabla_x (\varphi_1 \ldots \varphi_m)' = 1$$

is still possible!

Features in the last layers (abstract interpretation) are mostly uniform on the retina …

# CANONICAL FORM

$$\frac{\partial \varphi}{\partial t}(x,t) + C\varphi(x,t) = C\alpha_\varphi\big(b(\Omega_{\varphi,x},t),t\big)$$

$$\frac{\partial v_\varphi}{\partial t}(x,t) + Cv_\varphi(x,t) = C\alpha_v\big(\nabla\varphi(\Omega_{\varphi,x},t),t\big)$$

$$\frac{\partial \varphi}{\partial t}(x,t) = -\nabla_x \varphi(x,t)$$ we get rid of this one

in the Lagrangian …

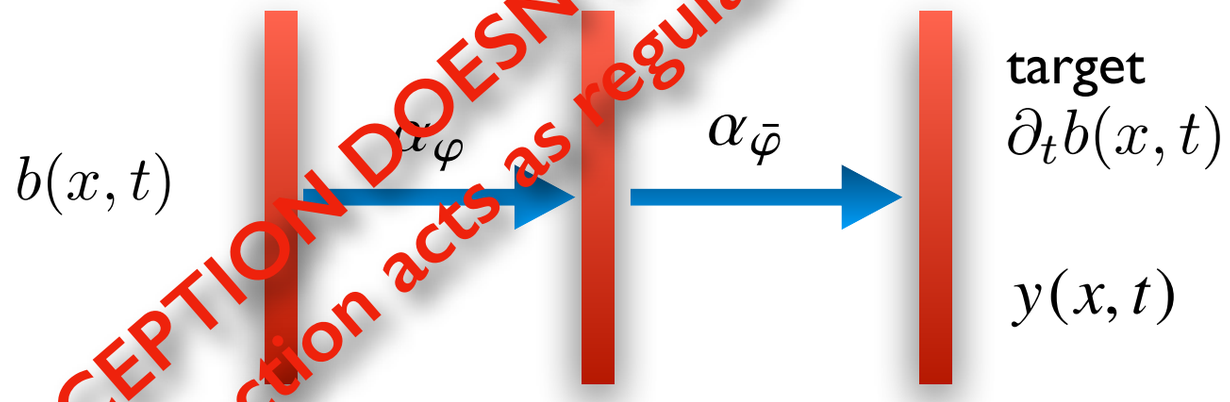
$$C^{-1}\nabla_x\varphi \cdot v_\varphi = \varphi(x,t) - \alpha_\varphi\big(b(\Omega_{\varphi,x},t),t\big)$$

# SUPPORTING VISUAL ACTION

Predict the variation of the brightness!

$$\frac{\partial \varphi}{\partial t}(x,t) + C\varphi(x,t) = C\alpha_{\varphi}\left(b(\Omega_{\varphi,x},t),t\right)$$

$$\frac{\partial y}{\partial t}(t,x) + Cy(x,t) = C\alpha_{\bar{\varphi}}\left(\varphi(\Omega_{\bar{\varphi},x},t),t\right)$$




PERCEPTION DOESN'T COME ALONE!
action acts as regularization

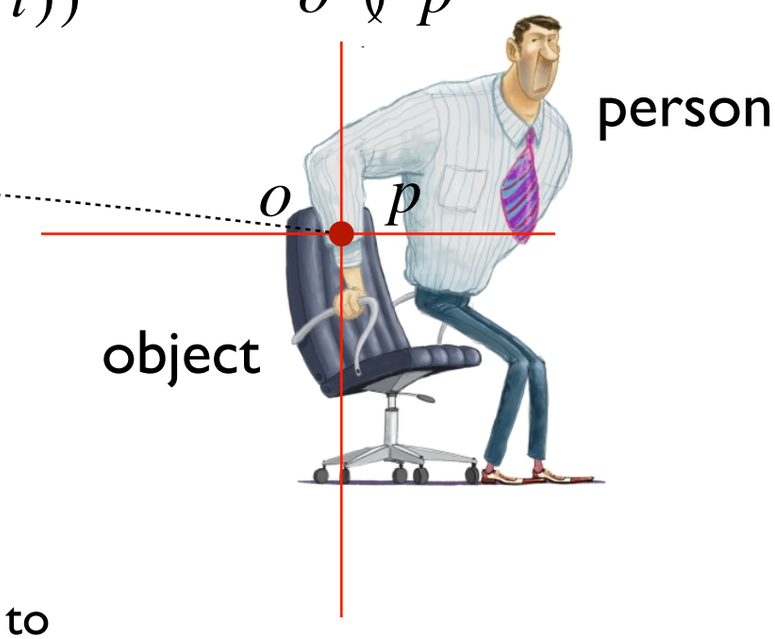
$$R = \frac{1}{2}\int_0^T dt \int_{\Omega} dx\ \partial_t b(x,t)\Big(y(x,t) - \partial_t b(x,t)\Big)^2$$

# II PRINCIPLE: COUPLED MOTION INVARIANCE

… my own chair and the "chair concept" …

$$p(x,t) \between o(x,t) \leftrightarrow \gamma(p(x,t)) \wedge \gamma(o(x,t))$$

$o \between p$




$(x,t)$

$$\texttt{CMI}: \quad \alpha_{op} \bowtie (v_o - v_p) = 0$$

relative velocity of the object with respect to the person who conveys affordance

Reduction to the I Principle: $\alpha_{op} \bowtie v_o = 0$

# II Principle
## Coupled Motion Invariance (con't)

Given $\epsilon > 0$ we say that
the coupling $o \between p$ is $\epsilon$-significant provided that $\mu(\mathscr{C}_{p \between o}) > \epsilon$ and write $o \between_\epsilon p$.

$o_1 \overset{p}{\sim} o_2 \leftrightarrow (p \between_\epsilon o_1) \wedge (p \between_\epsilon o_2)$

also specific actions: e.g. manipulation

$p_1 \overset{o}{\sim} p_2 \leftrightarrow (p_1 \between_\epsilon o) \wedge (p_2 \between_\epsilon o)$

e.g.: the same person can manipulate
different objects

… interesting relations

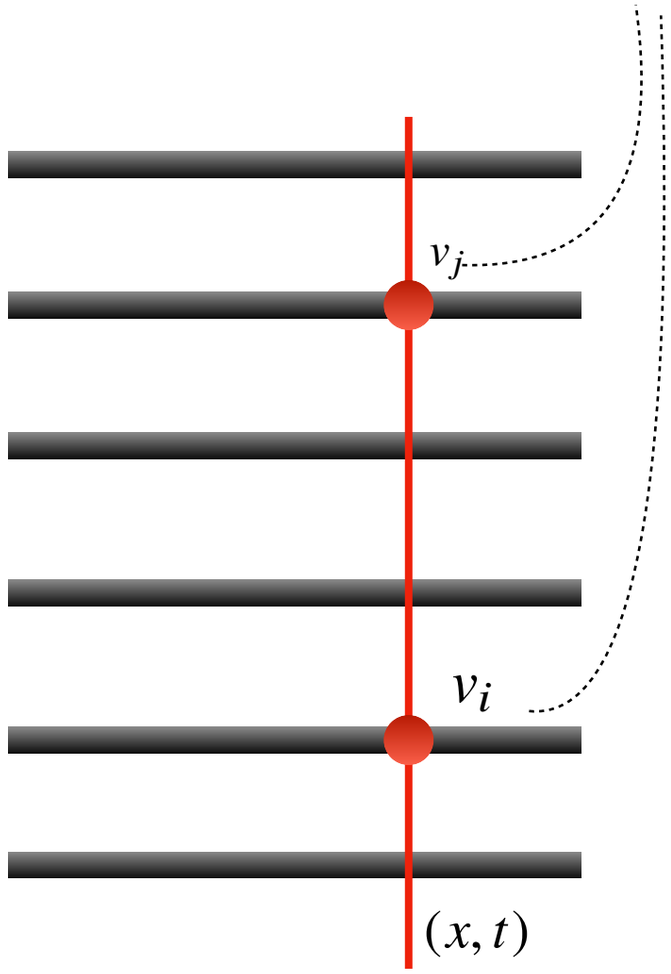
### Inherent Affordance

$$\forall j \in \mathcal{P}_o : \quad \alpha_o \bowtie (v_o - v_j) = 0 \qquad A = \sum_{j \in \mathcal{P}_o} (\alpha_o \bowtie (v_o - v_j))^2$$

objects in this class convey the
inherent environmental affordance

# II PRINCIPLE
## COUPLING OF VISION FIELDS

$$\texttt{CMI}-\texttt{bis}: \quad \psi_{ij} \bowtie (v_i - v_j) = 0$$




feature coupling!

Coupling from feature $j$

can come from different objects or from the same object!

indistinguishable

… coupling more than affordance

# COUPLING OF VISION FIELDS (con't)

Inherent Affordance

$$\forall j \in \mathcal{P}_i : \quad \psi_i \bowtie (v_i - v_j) = 0$$

fixed star coupling

$$\varphi \bowtie v_\varphi = 0 \qquad \varphi \xrightarrow{c} \psi$$

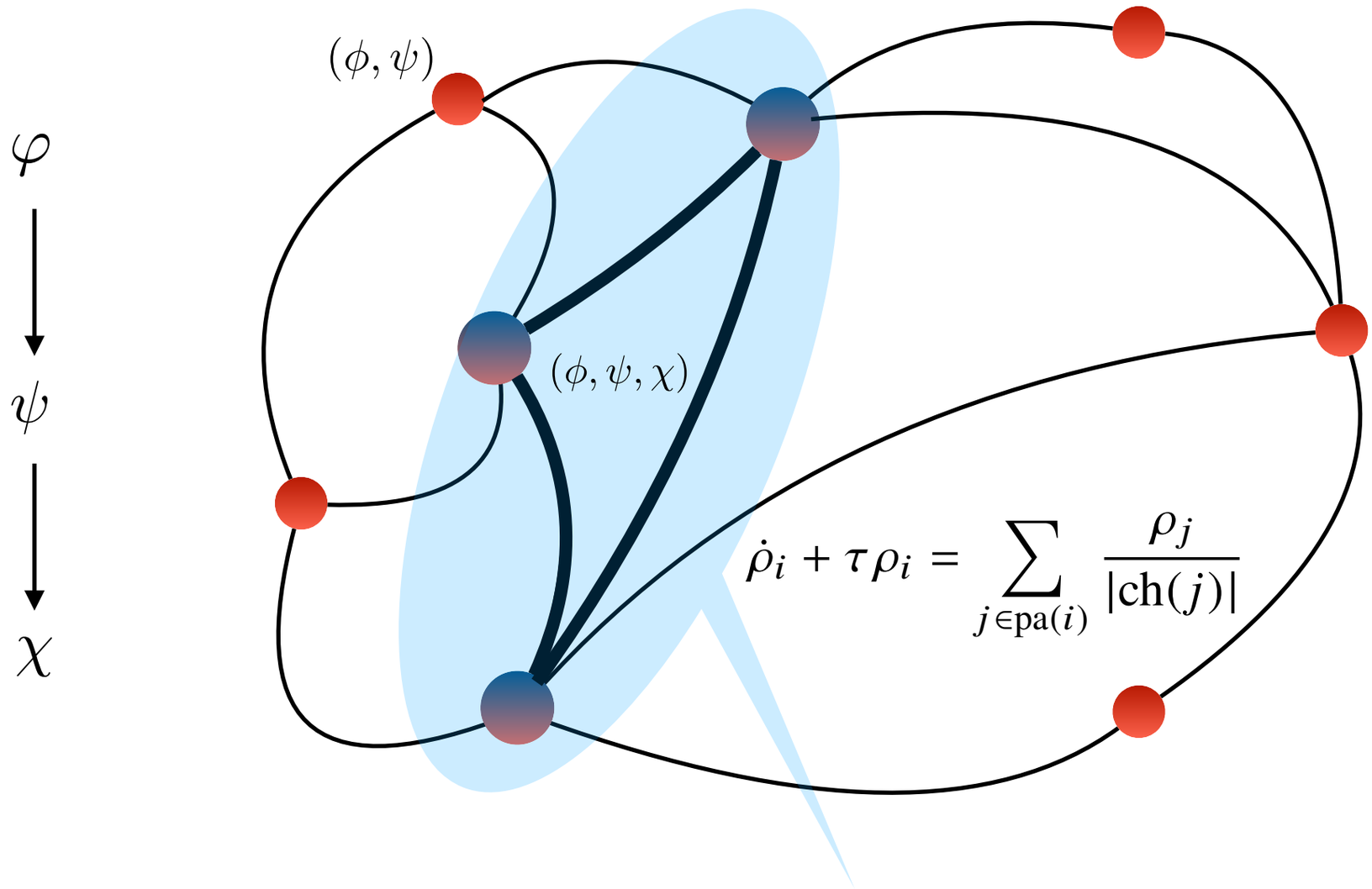
$$\Phi \rightarrow \Psi \quad \neg\Phi \vee \Psi = \neg(\Phi \wedge \neg\Psi)$$

t-norm translation

$$I_\psi = \sum_{i=1}^{n} \int_0^T dt \int_\Omega dx \, (1 - \psi_i(x,t))\varphi_i(x,t)$$

# VISION FIELD INTERACTION GRAPH

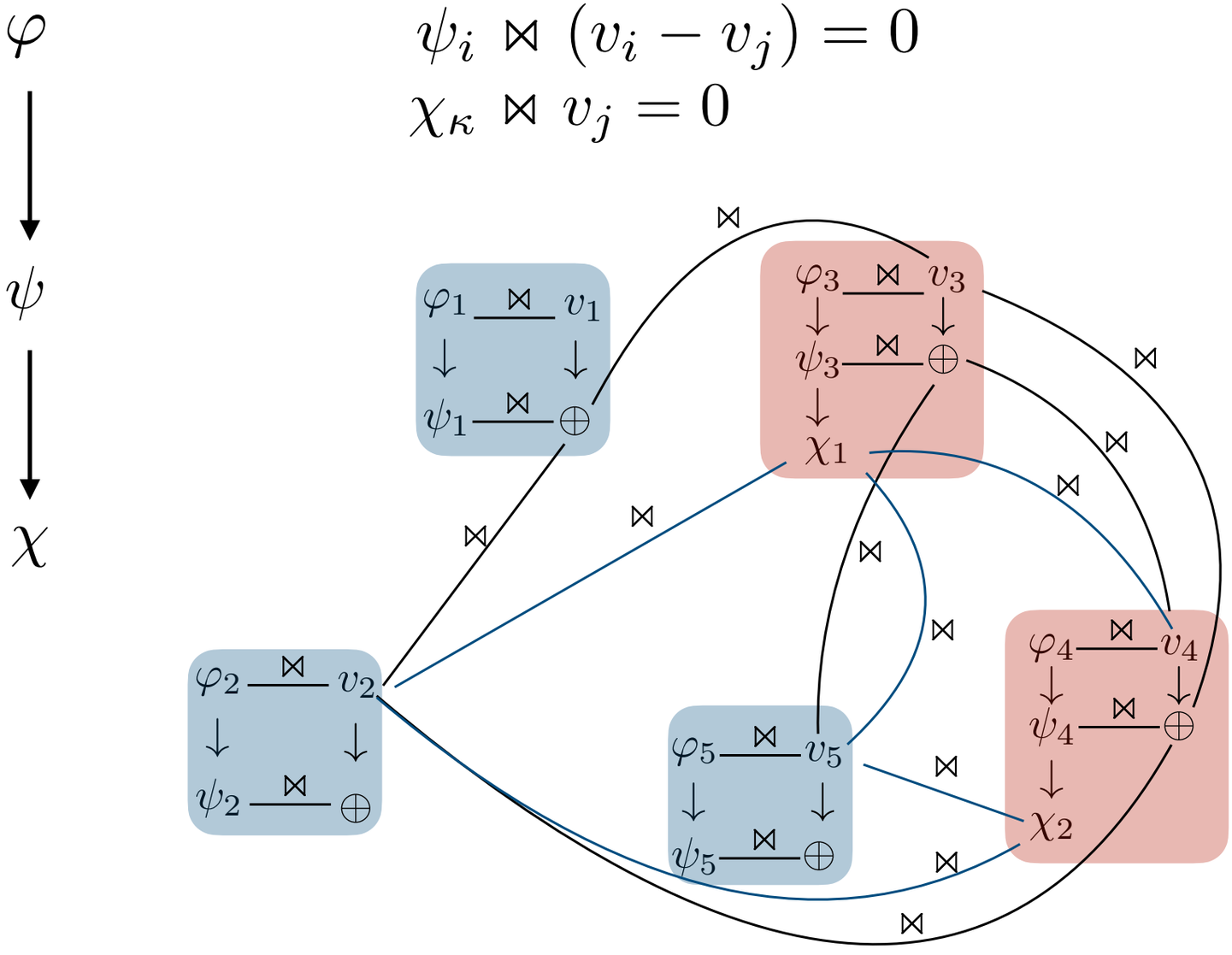
$$\mathcal{V} = (\varphi, \psi, \chi, v)$$




$$\chi_\kappa \bowtie v_j = 0$$

# VISION FIELD INTERACTION GRAPH (con't)

$$\varphi_i \bowtie v_i = 0$$
$$\psi_i \bowtie (v_i - v_j) = 0$$
$$\chi_\kappa \bowtie v_j = 0$$



# FOVEATE NEURAL NETWORKS

The remarkable properties of some recent computer algorithms for neural networks seemed to promise a fresh approach to understanding the computational perspectives of the brain. Unfortunately most of these neural nets are unrealistic in important respects.

Francis Crick, 1989

# FOVEATE NEURAL NETWORKS

$$\frac{\partial \varphi}{\partial t}(x,t) + C\varphi(x,t) = C\alpha_\varphi\big(b(\Omega_{\varphi,x},t),t\big)$$

$$\frac{\partial v_\varphi}{\partial t}(x,t) + Cv_\varphi(x,t) = C\alpha_v\big(\nabla\varphi(\Omega_{\varphi,x},t),t\big)$$

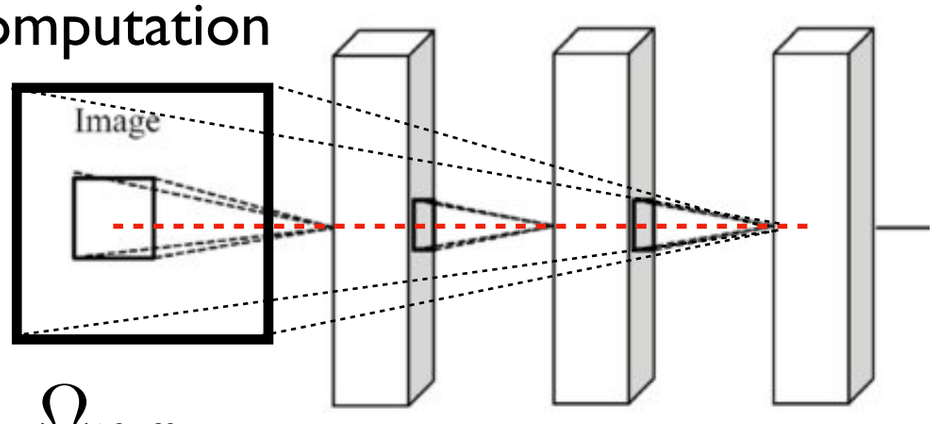
$$\alpha_\varphi(b(\Omega_{\varphi,x},t),t) = \eta_\varphi\big(w_\varphi(x,t), b(\Omega_{\varphi,x},t)\big)$$

$$\alpha_v(\nabla\varphi(\Omega_{\varphi,x},t),t) = \eta_v\big(w_{v_\varphi}(x,t), \nabla\varphi(\Omega_{\varphi,x},t)\big)$$

... neural computation

... biologically plausible

neural computation




$\Omega_{\varphi,x}$

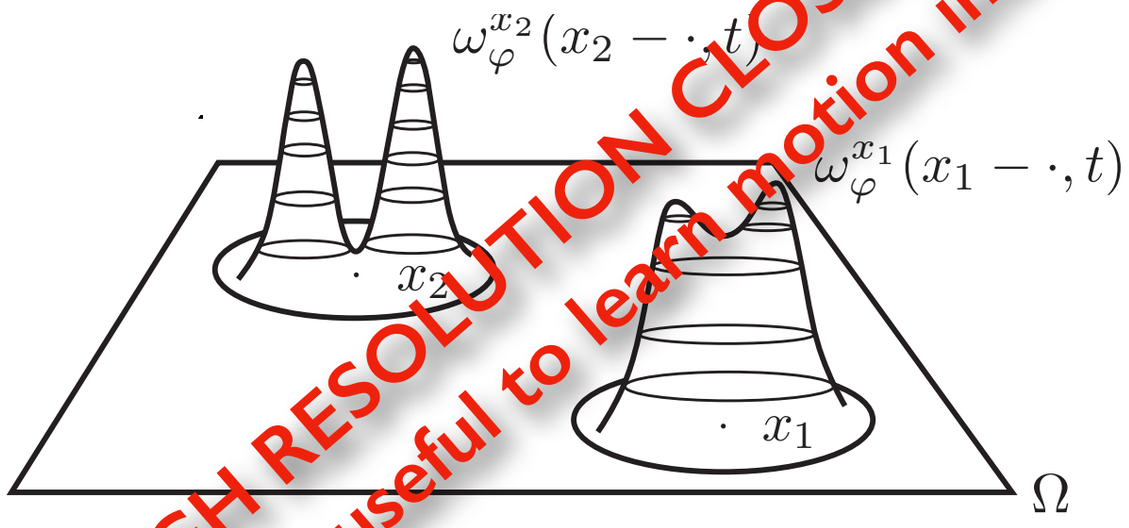
big $C$ "feedforward computation"

$$\varphi(x,t) = \eta_\varphi(w_\varphi(x,t), b(\Omega_{\varphi,x},t)) = 0$$

# FOVEATE NEURAL NETWORKS

"convolutional filters" which depends on
the position in the retina

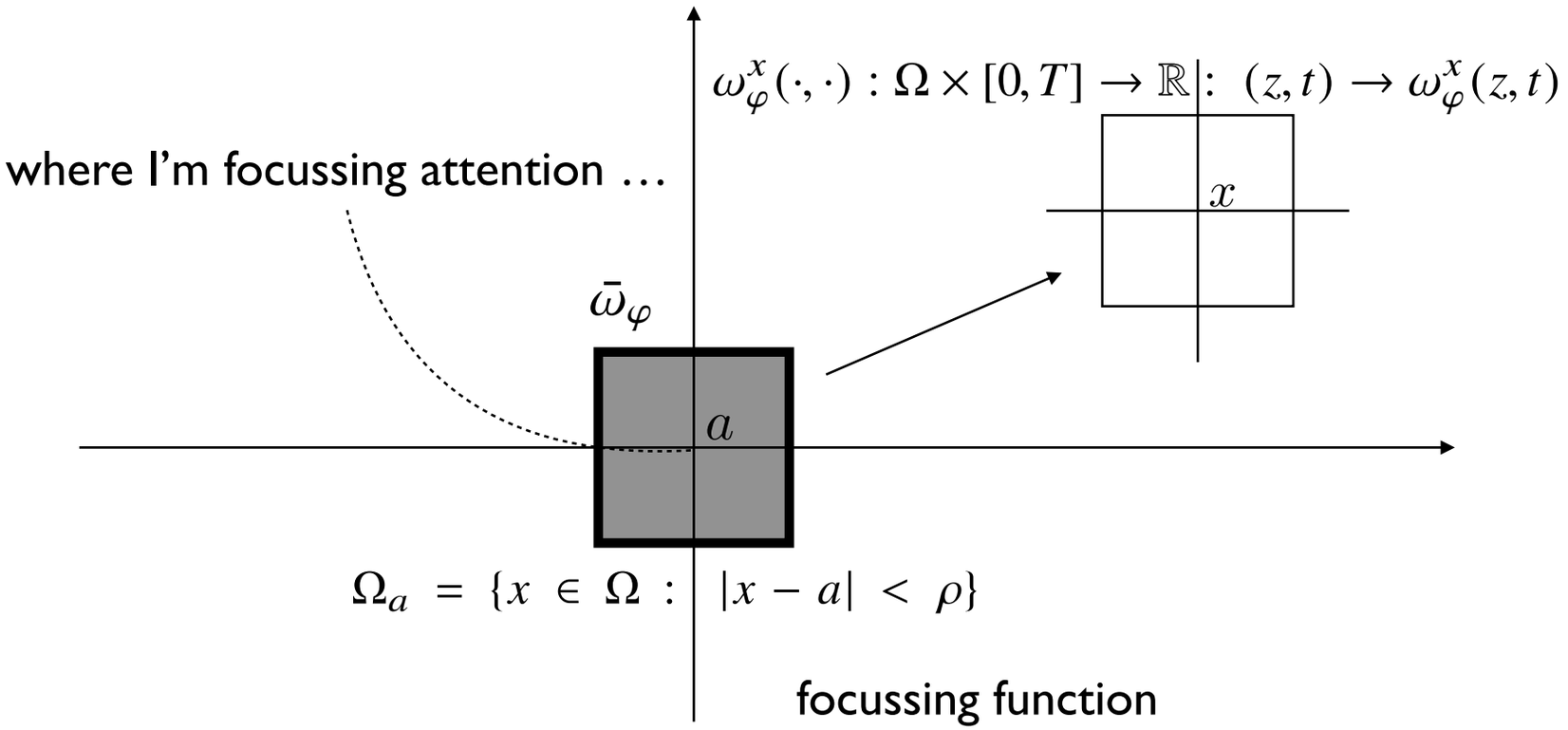
$$\omega_\varphi^x(\cdot,\cdot) : \Omega \times [0,T] \to \mathbb{R} : \ (z,t) \to \omega_\varphi^x(z,t)$$




HIGH RESOLUTION CLOSE TO THE FOA
it's useful to learn motion invariances

# FOVEATE NEURAL NEURAL NETWORKS

How to move a filter on a receptive field to a given position




$$\omega_\varphi^x(z,t) = \int_{\mathbb{R}^2} g(\sigma, x - a(t), z - \xi)\hat{\bar{\omega}}_\varphi(\xi,t)\, d\xi$$

$$\upsilon_\varphi(x,t) = (\omega_\varphi^x(\cdot,t) \star \hat{y}(\cdot,t))(x) := \int_{\mathbb{R}^2} \omega_\varphi^x(x - \xi, t)\hat{y}(\xi,t)\, d\xi$$

# FOVEATE NEURAL NEURAL NETWORKS

extreme cases

$$|x - a(t)| \approx 0$$

$$g(\sigma, d, x) = \delta(x)$$

$$\omega_{\varphi}^{x} \equiv \bar{\hat{\bar{\omega}}}_{\varphi}$$

$$|x - a(t)| \approx \operatorname{diam} \Omega$$

$$\upsilon_{\varphi}(x,t) = C\mathcal{L}^2(K)\langle \bar{\omega}(\cdot,t)\rangle_K \int_{\mathbb{R}^2} \hat{y}(\xi,t)\, d\xi =: C\mathcal{L}^2(K)\langle \bar{\omega}(\cdot,t)\rangle_K \langle y(\cdot,t)\rangle_{\Omega}$$

# INFORMATION-BASED LAWS OF LEARNING

When I was in high school, my physics teacher - whose name was Mr. Bader - called me down one day after physics class and said, “you look bored; I want toătell you something interesting.” Then he told me something which I found absolutely fascinating, and have, since then, always found fascinating. Every time the subject comes up, I work on it.

Richard Feynman, physics lectures about the principle of least action

# Optical Flow

Euler-Lagrange eqs: Horn & Schunck, 1981

$$E = \int_\Omega (b \bowtie v)^2 + \lambda\big((\nabla v_1)^2 + (\nabla v_2)^2\big)$$

$$\nabla^2 v_1 = \frac{1}{\lambda}(b \bowtie v)\ \partial_{x_1} b$$

$$\nabla^2 v_2 = \frac{1}{\lambda}(b \bowtie v)\ \partial_{x_2} b,$$

# OPTICAL FLOW: PLAYING WITH ⋈

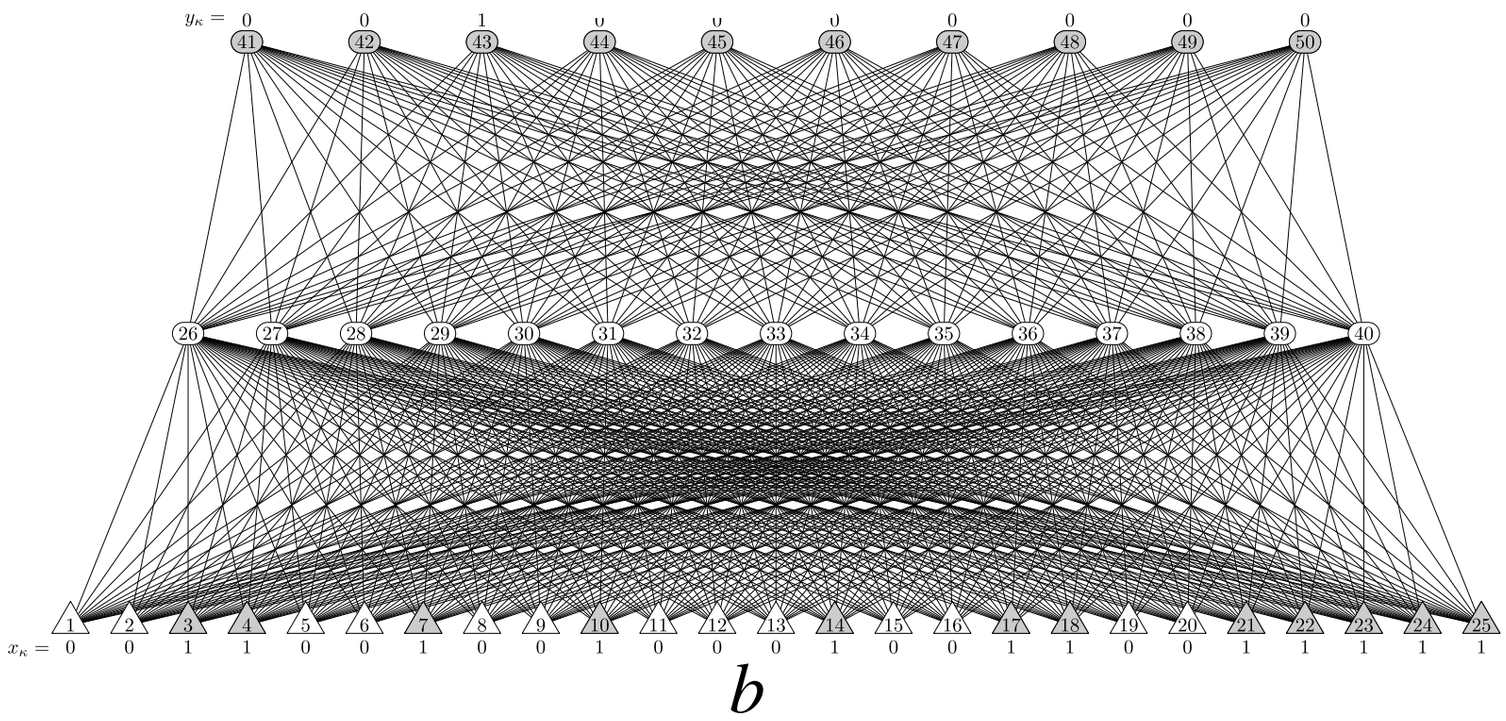
simplest example of conjugation

neural computation yields weight constraints

given!

$$E = \int_\Omega (b \bowtie v)^2 + \lambda\left((\nabla v_1)^2 + (\nabla v_2)^2\right)$$

$v$




$b$

# Optical Flow

Gradient-based learning of the weights

$$F_\varepsilon(\omega) := \int_0^T e^{-t/\varepsilon} \left( \frac{\nu\varepsilon}{2} |\omega'(t)|^2 + \mathrm{HS}(\omega(t), t) \right) dt$$

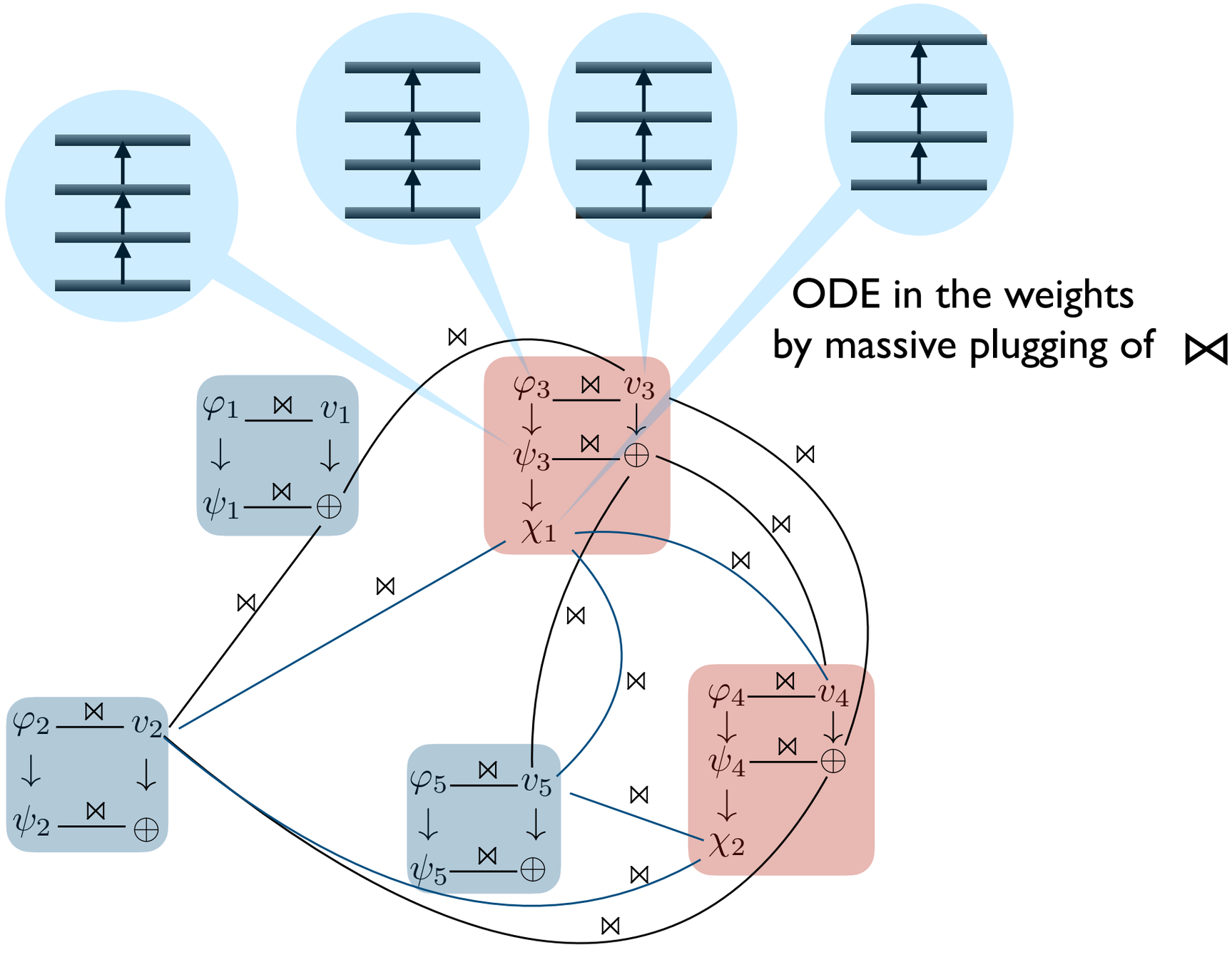
$$\begin{cases} \omega'(t) = -\frac{1}{\nu} \nabla \mathrm{HS}(\omega(t), t) \\ \omega(0) = \omega^0. \end{cases}$$

——————— input "blurring"

$$\overline{F}_\varepsilon(x, \sigma) := \int_0^T e^{-t/\varepsilon} \left( \frac{\nu_\omega \varepsilon}{2} |\omega'(t)|^2 + \frac{\nu_\sigma \varepsilon}{2} (\sigma'(t))^2 + \overline{\mathrm{HS}}(t, \omega(t), \sigma(t)) + \frac{k}{2} (\sigma(t))^2 \right) dt.$$

$$\begin{cases} \omega'(t) = -\frac{1}{\nu_\omega} \nabla \overline{\mathrm{HS}}(t, \omega(t), \sigma(t)) \\ \sigma'(t) = -k\sigma(t) - \frac{1}{\nu_\sigma} \overline{\mathrm{HS}}_\sigma(t, \omega(t), \sigma(t)) \\ \omega(0) = \omega^0, \quad \sigma(0) = \sigma^0. \end{cases}$$

# VISION FIELD AND NEURAL COMPUTATION



# GRADIENT-BASED LEARNING

$$t \mapsto u(t) := (\omega_A(t), \omega_B(t), \omega_C(t), \omega_\Delta(t), \vartheta(t)) \in \mathbb{R}^D$$

All the weights of the neural nets obey this learning equation

$$F(u) := \int_0^T \left( T(u'(t), u''(t)) + \frac{1}{2} |Q(u(t))u'(t) + b(u(t))|^2 + V(u(t), t) \right) dt$$

$$\begin{cases} \big(\nu \mathrm{Id} + \gamma M(u(t))\big) u'(t) + \gamma m(u(t)) + \nabla V(u(t), t) = 0, \quad t \in (0, T); \\ u(0) = u^0. \end{cases}$$

# CONCLUSIONS

- Motion is everything you need
- I and II Principle of Visual Perception
- Vision fields
- Foveate nets: Beyond convolutional nets?
- On-line temporal learning
- What is the interplay with language?

If you are interested in receiving the proto book, please drop me an email at marco.gori@unisi.it (proto-book available by the 31th of July)