frontiers in Computer Science

ORIGINAL RESEARCH
published: 03 March 2022
doi: 10.3389/fcomp.2021.701248

Ten Questions for a Theory of Vision

Marco Gori[1,2]*

[1]SAILab, Università di Siena, Siena, Italy, [2]3iA Chair Université Côte d'Azur, Nice, France

By and large, the remarkable progress in visual object recognition in the last few years has been fueled by the availability of huge amounts of labelled data paired with powerful, bespoke computational resources. This has opened the doors to the massive use of deep learning, which has led to remarkable improvements on new challenging benchmarks. While acknowledging this point of view, in this paper I claim that the time has come to begin working towards a deeper understanding of visual computational processes that, instead of being regarded as applications of general purpose machine learning algorithms, are likely to require tailored learning schemes. A major claim of in this paper is that current approaches to object recognition lead to facing a problem that is significantly more difficult than the one offered by nature. This is because of learning algorithms that work on images in isolation, while neglecting the crucial role of temporal coherence. Starting from this remark, this paper raises ten questions concerning visual computational processes that might contribute to better solutions to a number of challenging computer vision tasks. While this paper is far from being able to provide answers to those questions, it contains some insights that might stimulate an in-depth re-thinking in object perception, while suggesting research directions in the control of object-directed action.

Keywords: compter vision, convolutional networks, biological plausibility, motion invariance, optical flow

OPEN ACCESS

Edited by:
Marcello Pelillo,
Ca' Foscari University of Venice, Italy

Reviewed by:
Giuseppe Boccignone,
University of Milan, Italy
Kaleem Siddiqi,
McGill University, Canada
Fabio Cuzzolin,
Oxford Brookes University,
United Kingdom

***Correspondence:**
Marco Gori
marco.gori@unisi.it

Specialty section:
This article was submitted to Computer Vision, a section of the journal Frontiers in Computer Science

Received: *27 April 2021*
Accepted: *17 November 2021*
Published: *03 March 2022*

Citation:
Gori M (2022) Ten Questions for a Theory of Vision. Front. Comput. Sci. 3:701248. doi: 10.3389/fcomp.2021.701248

1 INTRODUCTION

The construction of huge supervised visual data bases has significantly contributed to the spectacular performance of deep learning. However, the extreme exploitation of the truly artificial communication protocol of supervised learning has its drawbacks, including the vulnerability to adversarial attacks, which might also be tightly connected to the negligible role typically played to the temporal structure. Are not we missing something? It looks like Nature did a great job by using time to "sew all the video frames", whereas it goes unnoticed to our eyes! At the dawn of pattern recognition, when we also began to cultivate the idea of interpreting natural video, in order to simplify the problem of dealing with a huge amount of information we removed time, the connecting thread between frames. As a consequence, all tasks of pattern recognition were turned into problems formulated on collections of images, where we only exploited spatial regularities and neglected the crucial role of temporal coherence. Interestingly, when considering the general problem of object recognition and scene interpretation, the joint role of the computational resources and the access to huge visual databases of supervised data has contributed to erect nowadays "reign of computer vision". At a first glance this is reasonable, especially if you consider that video were traditionally heavy data sources to be played with. However, a closer look reveals that we are in fact neglecting a fundamental clue to interpret visual information, and that we have ended up facing problems where the extraction of the visual concepts is mostly based on spatial regularities. On the other hand, reigns have typically consolidated rules from which it's hard to escape. This is common in novels and real life. "The Three Princes of Serendip" is the English version of "Peregrinaggio di tre giovani figliuoli

del re di Serendippo," published by Michele Tramezzino in Venice on 1557. These princes journeyed widely, and as they traveled they continually made discoveries, by accident and sagacity, of things they were not seeking. A couple of centuries later, in a letter of 28 January 1754 to a British envoy in Florence, the English politician and writer Horace Walpole coined a new term: serendipity, which is succinctly characterized as the art of finding something when searching for something else. Couldn't similar travels open new scenario in computer vision? Couldn't the visit to well-established scientific domains, where time is dominating the scene, open new doors to an in-depth understanding of vision? We need to stitch the frames to recompose the video using time as a thread, the same thread we had extracted to work on the images at the birth of the discipline. We need to go beyond a peaceful interlude and think of reinforcing the currently few contributions on learning theories based on video more than on images!

This paper is a travel towards the frontiers of the science of vision with special emphasis on object perception. We drive the discussion by a number of curious questions that mostly arise as one tries to interpret and disclose natural vision processes in a truly computational framework. Unfortunately, as yet, we are far away from addressing the posed questions. However, this paper takes the position that asking right questions on the discipline might themselves stimulate its progress.

2 CUTTING THE UMBILICAL CORD WITH PATTERN RECOGNITION

In the eighties, Satosi Watanabe wrote a seminal book (Watanabe, 1985) in which he pointed out the different facets of pattern recognition. Most of the modern work on computer vision for object perception fits with Watanabe's view of pattern recognition as statistical decision making and pattern recognition as categorization. Based on optimization schemes with billions of variables and universal approximation capabilities, the spectacular results of deep learning have elevated this view of pattern recognition to a position where it is hardly debatable. While the emphasis on a general theory of vision was already the main objective at the dawn of the discipline (Marr, 1982), its evolution has been mostly marked by significant experimental achievements. Most successful approaches seem to be the natural outcome of a very well-established tradition in pattern recognition methods working on images, which have given rise to the present emphasis on collecting big labelled image databases (e.g., Deng et al., 2009). However, in spite of these successful results, this could be the time of an in-depth rethinking of what we have been doing, especially by considering the remarkable traditions of the overall field of vision. Couldn't it be the right time to exploit the impressive literature in the field of vision to conquer a more unified view of object perception?

In the last few years, a number of studies in psychology and cognitive science have been pushing truly novel approaches to vision. In (Kingstone et al., 2010), it is pointed out that a critical problem that continues to bedevil the study of human cognition is related to the remarkably successes gained in experimental psychology, where one is typically involved in simplifying the experimental context with the purpose to discover causal relationships. In so doing we minimize the complexity of the environment and maximize the experimental control, which is typically done also in computer vision when we face object recognition. However, one might ask whether such a simplification is really adequate and, most importantly, if it is indeed a simplification. Are we sure that treating vision as a collection of unrelated frames leads a simplification of learning visual tasks? In this paper we argue this not the case, since cognitive processes vary substantially with changes in context. When promoting the actual environmental interaction, Kingstone et al. (2010) introduce a novel research approach, called, "Cognitive Ethology", where one opts to explore first how people behave in a truly naturally situation. Once we have collected experience and evidence in the actual environment then we can move into the laboratory to test hypotheses. This strongly suggests that also machines should learn in the wild!

Other fundamental lessons come from the school of robotics for whatever involves the control of object-directed actions. In (Benjamin et al., 2011), it is pointed out that "the purpose of vision is very different when looking at a static scene to when engaging in real-world behavior." The interplay between extracting visual information and coordinating the motor actions is a crucial issue to face for gaining an in-depth understanding of vision. One early realizes that manipulation of objects is not something that we learn from a picture; it looks like you definitely need to act yourself if you want to gain such a skill. Likewise, the perception of the objects you manipulate can nicely get a reinforcement form such as mechanical feedback. The mentioned interplay between perception and action finds an intriguing convergence in the natural processes of gaze control and, overall, on the focus of attention (Ballard, 1991). It looks like *animate vision* goes beyond passive information extraction and plays an important role in better posing most vision tasks.

The studies in computer vision might benefit significantly also from the exploration of the links with predictive coding Rao and Ballard (1999) that have had a remarkable impact in neuroscience. In that framework one is willing to study theories of brain in which it constantly generates and updates a "mental model" of the environment. Overall, the model is supposed to generate its own predictions of sensory input and compare them to the actual sensory input. The prediction error is expected to be used to update and revise the mental model. While most of the studies in deep learning have been focused on the direct learning of object categories, there are a few contributions also in the direction of performing a sort of predictive coding by means of auto-encoding architectures (Ronneberger et al., 2015). This neural network, which is referred to as an *U-net*, is used for medical image segmentation.

3 DEALING WITH VIDEO INSTEAD OF IMAGES

While the processing of video has been massively investigated for a number of computer vision tasks, including tracking and action

recognition, the emphasis on methods rooted on still images has currently been dominating the state of the art in object recognition approaches. In this paper we argue that there are strong arguments to start exploring the more natural visual interaction that animals experience in their own environment for all perception tasks. The idea of shifting to video is very much related to the growing interest of *learning in the wild* that has been explored in the last few years[1]. The learning processes in the wild have a different nature with respect to those that are typically considered in machine learning. While ImageNet (Deng et al., 2009) is a collection of unrelated images, a video supports information only when motion is involved. In presence of still images that last for awhile, the corresponding stream of equal frames only conveys the information of a single image–apart from the duration of the interval in which the video has been kept constant. As a consequence, visual environments mostly diffuse information only when motion is involved. As time goes by, the information is only carried out by motion, which modifies one frame to the next one according to the optical flow. Once we deeply capture this fundamental feature of vision, we realize that a different theory of machine learning is needed that must be capable of naturally processing streams instead of a collection of independent images. An important ingredient for such a theory is that of emphasizing the role of the position (pixel) on which the decision is carried out and, even more, the role of time in the recognition processes doesn't seem to play a central role. It looks like we are mostly ignoring that we are in front of spatiotemporal information, whose reduction to isolated patterns might not be a natural approach especially for the complex tasks that we have been recently tackling. While there are already remarkable contributions on computer vision approaches that perform semantic labelling, most methods struggle for massive labeling that is difficult to achieve and definitely far away from natural human skills.

It is worth mentioning that pixel-based computations and segmentation have been successfully addressed in important real-world applications. In the last decades, the massive production of electronic documents, along with their printed version, has given rise to specialized software tools to extract textual information from optical data. Most optical documents, like tax forms or invoices, are characterized by a certain layout which dramatically simplifies the process of information extraction. Basically, as one recognizes the class of a document, its layout offers a significant prior on what we can expect to find in its different areas. For those documents, the segmentation process can often be given a somewhat formal description, so as most of the problems are reduced to deal with the presence of noise. Basically, the knowledge on the document layout typically offers the opportunity of providing robust solutions. The noise doesn't compromise significantly the presence of segmentation, that is in fact very well driven by the expectations provided in each pixel of the documents. These guidelines have been fueling the field of document analysis and recognition (DAR), whose growth in the last few years has led to impressive results (Marinai et al., 2005). Unfortunately, in most real-world problems, as we move to natural images and vision, the methodology used in DAR is not really effective. The reason is that there is no longer a reliable anchor to which one can cling for segmenting the objects of a scene. While we can provide a clear description of characters and lines in optical documents, the same doesn't hold for the picture of a car which is mostly hidden by a truck during the overtaking. Humans exhibit a spectacular detection ability by simply relying on small glimpses at different scale and rotations. In no way are those cognitive processes reducible to the well-posed segmentation problems of chars and lines in optical documents. As we realize that there is a car, we can in fact provide its segmentation. Likewise, if an oracle gives us the segmented portion of a car, we can easily classify it. Interestingly, we don't really know which of the two processes is given a priority–if any. We are trapped into the chicken-egg dilemma on whether classification of objects must take place first of segmentation or vice versa. Amongst others, this issue has been massively investigated by in (Borenstein and Ullman, 2002) and pointed out in (Ullman, 1979). This intriguing dilemma might be connected with the absence of focus of attention, which necessarily leads to holistic mechanisms of information extraction. Unfortunately, while holistic mechanisms are required at a certain level of abstraction, the segmentation is a truly local process that involves low level features.

The bottom line is that most problems of computer vision are posed according to the historical evolution of the applications more than via an in-depth analysis of the underlying computational processes. While this choice has been proven to be successful in many real-world cases, stressing this research guideline might lead, on the long run, to sterile directions. Somewhat outside the mainstream of massive exploration of supervised learning, Poggio and Anselmi (Poggio and Anselmi, 2016) pointed out the crucial role of incorporating appropriate visual invariances into deep nets to go beyond the simple translation invariance that is currently characterizing convolutional networks. They propose an elegant mathematical framework on visual invariance and enlighten some intriguing neurobiological connections. Couldn't it be the case that the development of appropriate invariances might be exactly what is needed to go one step beyond?

4 QUESTIONS AND INSIGHTS

A good way to attack important problems is to pose the right question. To quote Tukey:

> Far better an approximate answer to the right question, which is often vague, than the exact answer to the wrong question, which can always be made precise.

Overall, posing appropriate questions can open a debate and solicit answers. However, the right questions cannot be easily posed since, while they need to have a big pictures in mind, we

[1]See, e.g., https://sites.google.com/site/wildml2017icml/. Of course, the spirit of learning in the wild goes beyond video, but there is in fact a natural match between them

also need the identification of reasonable intermediate steps. It is often the case that while addressing little problems, inconsistencies arise that suggest the formulation of better questions. In this section, we formulate ten questions on the emergence of visual skills in nature that might contribute to the development of a new approach to computer vision that is based in processing of video instead of huge collections of images.

4.1 How can we Go Beyond "Intensive Supervision"?

We start with a fundamental question that reveals a commonly recognized limitation of current studies in computer vision:

> ***Q1:*** *How can animals acquire visual skills without requiring "intensive supervision"?*

Everybody working in the broad area of artificial intelligence and in the field of cognitive science has been stimulated by this question. Yet, it has not been addressed as it definitely deserves! For example, it is well-known that the acquisition of the abstract notion of objects goes well beyond shape interpretation. It has been pointed out that a fundamental abstraction process arises when considering the object interactions in the environment, which does convey affordance (Gibson, 1950; Gibson, 1966; Gibson and Boston 1979). While this cognitive property is clearly of fundamental importance, we have not seen its significant exploitation in state of the art object recognition systems, yet.

The recent remarkable achievements in computer vision come from a different mechanism with respect to human vision which is based on tons of supervised examples—of the order of millions. The different environmental interactions that one can conceive in computers makes it possible to stress this artificial protocol of learning that is supported by mathematical foundations from decades and, more recently, by professional software tools. Interestingly, humans could not be exposed to such a boring interaction. Hence, strictly speaking, the mechanisms behind supervised learning are artificial and offer a space for machines to conquer visual skills that humans couldn't replicate in such a context. Of course, there is no need to be surprised by the spectacular capabilities that machines can achieve under such an artificial communication protocol, just like nobody gets surprised by the computer speed in performing multiplications.

On the other hand, because of the completely different environmental interaction, humans conquer the capability of recognizing objects just by a few supervisions. If we are interested in the scientific foundations of vision, we should aspire for an in-depth explanation of this remarkable difference between humans recognition capabilities and present computer vision technologies. Because of the expected saturation of performance of systems based on state of the art technologies, the time has already come to face the question. This has been advocated in a number of papers (see e.g., Lee et al., 2009; Ranzato et al., 2007; Goroshin et al., 2015; Tavanaei et al., 2016), and the interest in this kind of exploration is growing.

It turns out that when the posed question is analyzed more carefully, one easily comes into the conclusion that the answer must be searched in the different learning protocols in humans and computers (e.g., active learning and active vision, see Aloimonos et al., 1988). In a sense, we need to focus on the role of a continuous environmental interaction during the "life of the agent." The interaction is not only carried out through symbolic communication but also it seems to rely primarily on continuous-based information exchange. The reward for successful actions is dominating the visual learning process in animals of any specie. These actions range from drinking milk from the mother's breast to obstacle avoidance. Hence, those actions basically result in reinforcement learning processes that contribute to the development of the visual skills. An eagle, like any predator, is early driven by the objective of capturing the prey, that strongly contributes to the development of the visual skills. It is worth mentioning that for newborns, like for other primates, the acquisition of visual skills also benefits from the joint development of motion control. As soon as a baby begins the first experiences of object manipulation, such a task is paired with visual development. Successful manipulations are due to the correct visual interpretation which, in turn, is reinforced by the concrete acts of touching and moving. Hence, this is just a way to "supervise" and reinforce visual skills as the outcome of other processes taking place in the environment. Interestingly, at that time, linguistic skills are nearly absent, which clearly indicates that, just for other animals, the learning of vision undergoes its own developmental phases, while the interplay with language comes later on.

Of course, in nature the discussion on the acquisition of vision cannot neglect the fundamental role of genetic inheritance, that has a fundamental impact on the cognitive developmental steps. It is worth mentioning that remarkable differences have been discovered between different species of animals, in terms of the balance between their genome and their learning acquisition from the environment. For example, while chicks acquire significant visual skills early on (Wood and Wood, 2020), it takes months for humans, that clearly need to conquer more sophisticated skills at the perceptual level.

Interestingly there are intriguing connections with recent achievements in machine learning, where we have been constantly underlying the crucial role of transfer learning mechanisms (Pan and Yang, 2010) in many real-world problems. Clearly, whenever we transfer knowledge to the learning agent, this simplifies its own visual tasks and, as a consequence, the agent can learn by relying on less supervised examples. However, one should bear in mind that the current generation of neural networks that are currently used in transfer learning relies on the development of "individuals" that are strongly limited from a "genetic viewpoint", since they haven't been exposed to visual natural interactions. The learned features that are "genetically" transmitted are typically developed under supervised learning on classification that, even for large databases, might be biased by the specific benchmark.

BOX 1 | The bottom line is that for the posed question to be addressed one must dramatically change the agent environmental interactions. Interestingly, it is not the lack of technology for mimicking human interaction that mostly prevents us from going beyond "intensive supervision", but the lack of a theory to support the appropriate computational mechanisms. From one side, this is a stimulating scientific challenge. From the other side, the development of a similar theory would likely contribute to open new technological perspectives where machines learn to seen by a on-line scheme without needing visual databases.

4.2 What is the Role of Time?

As stated at the end of the previous section, a remarkable distinction between humans and state of the art machines is that they operate within a very different learning environment. The visual interaction that we experiment in nature leads us to face the following fundamental question:

> ***Q2:*** *How can animals gradually acquire visual skills in their own environments?*

Apparently, such a gradual learning scheme also takes place in machine learning. But, can we really state that the "gradual" process of human learning is somewhat related to the "gradual" weight updating of neural networks? A closer look at the mechanisms that drive learning in vision tasks suggests that current models of machine learning mostly disregard the fundamental role of "time", which shouldn't be confused with the iteration steps that mark the weight update. First, notice that learning to see in nature takes place in a context where the classic partition into learning and test environment is arguable. On the other hand, this can be traced back to early ideas on statistical machine learning and pattern recognition, which are dominated by the principles of statistics. In the extreme case of batch mode learning, the protocol assumes that the agent possesses information on its life before coming to life. Apparently this doesn't surprise computer vision researchers, whereas it sounds odd for the layman, whose viewpoint shouldn't be neglected, since we might be trapped into an artificial world created when no alternative choice was on the horizon. The adoption of mini-batches and even the extreme solution of on-line stochastic gradient learning are still missing a true incorporation of time. Basically, they pass through the whole training data many times, a process which is still far from natural visual processes, where the causal structure of time dictates the gradual exposition to video sources. There is a notion of life-long learning that is not captured in current computational schemes, since we are ignoring the role of time which imposes causality.

Interestingly, when we start exploring the alternatives to huge collections of labelled images, we are immediately faced with a fundamental choice, which arises when considering their replacement with video collections. What about their effectiveness? Suppose you want to prepare a collection to be used for learning the market segment associated with cars (luxury vehicles, sport cars, SUVs/off-road vehicles, . . .). It could be the case that a relatively small image database composed of a few thousands of labelled examples is sufficient to learn the concept. On the other hand, in a video setting, this corresponds with a few minutes of video, a time interval in which it is unreasonable to cover the variety of car features that can be extracted from 10,000 images! Basically, there will be a lot of near-repetitions of frames which support scarse information with respect to the abrupt change from picture to picture. This is what motivates a true paradigm shift in the formulation of learning for vision. In nature, it is unlikely to expect the emergence of vision from the accumulation of video. Hence, couldn't we do same? A new communication protocol can be defined where the agent is simply expected to learn by processing the video as time goes by without its recording and by handling human-like vocal interactions. Interestingly, this gets rid of the need to accumulate and properly handle huge collections of labelled images, which would represent a paradigm shift in computer vision and might opens great opportunities to all research centers to compete in another battlefield.

At any stage of child development, it looks like only the visual skills that are required to face the current tasks are acquired. One might believe that this is restricted to natural processes, but we conjecture that the temporal dimension plays a crucial role in the well-positioning of most challenging cognitive tasks, regardless of whether they are faced by humans or machines. The formulation of learning in the temporal dimension likely becomes more and more important when we begin to address a number of challenges that are also outlined in the other questions posed in the paper. The role of time becomes crucial when considering the extraction of good features. This is in fact an issue that, as the interest in transfer learning has been demonstrating, has becoming more and more relevant. While in the literature we have been typically concerned with feature extraction that is independent of classic geometric transformation, it looks like we are still missing the astonishing human skill of capturing distinctive features when looking at ironed and rumpled shirts! There is no apparent difficulty to recognize shirts by keeping the recognition coherence in case we roll up the sleeves, or we simply curl them up into a ball for the laundry basket. Of course, there are neither rigid transformations, like translations and rotation, nor scale maps that transforms an ironed shirt into the same shirt thrown into the laundry basket. Is there any natural invariance?

In this paper, we claim that motion invariance is in fact the only one that we need. Translation, scale, and rotation invariances, that have been the subject of many studies, are in fact instances of invariances that can be fully gained whenever we develop the ability to detect features that are invariant under motion. If my finger moves closer and closer to my eyes then any of its representing features that is motion invariant will also be scale invariant. The finger will become bigger and bigger as it approaches my face, but it is still my finger! Clearly, translation, rotation, and complex deformation invariances derive from motion invariance. Humans life always experiments motion, so as the gained visual invariances naturally arise from motion. Studies on different types of invariances in vision have been so rich and massively investigated that one can suspect that there is something missing in the claim that enforcing motion invariance only suffices to learn. The

emergence of information-based laws of learning can take place and, consequently, a natural formulation of learning to see in the temporal dimension relies on the principle of *Material Point Invariance*, which blesses the pairing of any feature of a given object, including the brightness, with its own velocity.

We can somewhat parallel the idea of brightness invariance that is used for the estimation of the optical flow for imposing a fundamental invariance condition on any visual feature φ. Hence, the following consistency condition holds true:

$$\varphi\big(x_\varphi(t),t\big) = \varphi\big(x_\varphi(0),0\big) = c_\varphi, \quad \forall t \in [0,T] \tag{1}$$

where $x_\varphi(t)$ denotes a trajectory of the associated feature φ and $c_\varphi \in \mathbb{R}$. It is worth mentioning that motion invariance is not alway desirable since perception and action also need to develop features that provide different reactions in front of motion. For example, this makes it possible to learn the meaning of "vertical" and "horizontal" positions and to react to moving objects. When thinking of the trajectory $x_\varphi(t)$ we must bear in mind that if we consider that feature extraction takes place under focus of attention mechanisms then the visual agent always experiences motion. Formally, for any pair (x, t) (pixel-time), let $x_\varphi(t) = x$ and $\dot{x}_\varphi(t) = v_\varphi(x,t)$ be. Given the optical flow v_φ. Then motion invariance of feature φ can be expressed by

$$\frac{d\varphi\big(x_\varphi(t),t\big)}{dt} = \nabla\varphi(x,t)\cdot v_\varphi(x,t) + \varphi_t(x,t) = 0. \tag{2}$$

This shares the formal structure of the transport equation of brightness invariance, but it is important to notice that this equation is *stating a principle which is not supposed to be violated*. While brightness invariance represents an approximation, the conjunction stated by **Eq. 2** is expected to hold perfectly. Basically (φ, v_φ) is an *indissoluble pair* that plays a fundamental role in the learning of the visual features that characterizes the object. Notice that, just like the color components of a color video typically share the same velocity, also different features can be aggregated with the same velocity.

What if an object is gradually deformed into another one? One can think of cat which is very slowly transformed to a dog! This relates to the metaphysical question of whether and how things persist over time. Philosophers regards things as concrete material objects as well as pure abstract objects which are associated with concepts and ideas[2]. Concepts can drift as video changes very slowly, which can deceive any intelligent agent who relies on motion invariance only, provided that such an agent fails at detecting slow motion. While one can always argue about the possibility of achieving certain thresholds for slow motion detection, when trusting motion invariance one must be ready to accept the problem of concept drift. Interestingly, as will be pointed out in the reminder of the paper, the development of focus of attention mechanisms helps facing also this problem.

It could be the case that the extraction of very efficient information from visual processes that has captured the attention for decades of computer scientists and engineers benefits from a sort of *pre-algorithmic phase* where we primarily need to understand basic *perceptual laws of vision* that hold regardless of the nature of the agent. This research guideline has been stimulating the conception of variational laws of learning that can capture the elegance and the simplicity of natural behavior (Betti and Gori, 2016; Betti et al., 2018). When following this approach one promotes the role of time by following principles that have a very well-established tradition in physics. Most importantly, one begins to challenge the indisputable principle that learning can be regarded as the outcome of an optimization process that operates on the risk function. Clearly, there is something wrong with this principle in nature, since the agent doesn't possess the risk at the time of its birth! While such a risk can be gradually constructed and the adoption of stochastic gradient is a poweful idea for capturing the underlying statistics, we are basically attacking a different problem with respect to what all species of animals are expected to face in nature.

BOX 2 | The bottom line is that while we struggle for the acquisition of huge labeled databases, the true incorporation of time might led to a paradigm shift in the process of feature extraction. We promote the study of the agent life based on the ordinary notion of time, which emerges in all its facets. The incorporation of motion invariance might be the key for overcoming the artificial protocol of supervised learning. We claim that such an invariance is in fact the only one that we need.

4.3 Can Animals see in a World of Shuffled Frames?

One might figure out what human life could have been in a world of visual information with shuffled frames.

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

A related issue has been faced in (Wood, 2016) for the acquisition of visual skills in chicks. It is pointed out that "when newborn chicks were raised with virtual objects that moved smoothly over time, the chicks developed accurate color recognition, shape recognition, and color-shape binding abilities." Interestingly, the authors notice that in contrast, "when newborn chicks were raised with virtual objects that moved non-smoothly over time, the chicks' object recognition abilities were severely impaired." When exposed to a video composed of independent frames taken from a visual database, like ImageNet, that are presented at classic cinema frame rate of 24 fps, humans seem to experiment related difficulties in non-smooth visual presentation.

[2]This discussion was stimulated by Marcello Pelillo who pointed out intriguing links with The "Ship of Theseus Puzzle" and "The Puzzle of the Statue and the Clay."

Hence, it turns out that our spectacular visual skills completely collapse in a task that is successfully faced in computer vision! As a consequence, one might start formulating conjectures on the inherent difficulty of artificial versus natural visual tasks. The remarkably different performance of humans and machines has stimulated the curiosity of many researchers in the field. Of course, you can start noticing that in a world of shuffled frames, a video requires an order of magnitude more information for its storing than the corresponding temporally coherent visual stream. This is a serious warning that is typically neglected in computer vision, since it suggests that any recognition process is likely to be more difficult when shuffling frames. One needs to extract information by only exploiting spatial regularities in the retina, while disregarding the spatiotemporal structure that is offered by nature. The removal of the thread that nature used to sew the visual frames might prevent us from the construction of a good theory of vision. Basically, we need to go beyond the current scientific peaceful interlude and abandon the safe model of restricting computer vision to the processing of images. Working with video was discouraged at the dawn of computer vision because of the heavy computational resources that it requires, but the time has come to reconsider significantly this implicit choice. Not only it is the case that humans and animals cannot see in a world of shuffled frames, but it is likely that they could not learn to see in such an environment. Shuffling visual frames is the implicit assumption of most of present vision technology that, as stated in the previous section, corresponds with neglecting the role of time in the discovery of visual regularities. No matter what computational scheme we conceive, the presentation of frames where we have removed the temporal structure exposes visual agents to a problem where a remarkable amount of information is delivered at any presentation of new examples. When going back to the previous discussion on time, one clearly see its natural environmental flow that must be somehow synchronized with the agent's computational capability. The need for this synchronization is in fact one of the reasons for focussing attention at specific positions in the retina, which confers the agent also the gradual capability of extracting information at pixel label. Moreover, as already pointed out, we need to abandon the idea of recording a data base for statistical assessment. There is nothing better than human evaluation in perceptual tasks, which could stimulate new ways of measuring the scientific progress of the discipline (see **Section 5**).

The reason for formulating a theory of learning on video instead of on images is not only rooted in the curiosity of grasping the computational mechanisms that take place in nature. A major claim in this paper is that those computational mechanisms are also fundamental in most of computer vision tasks.

BOX 3 | It appears that, while ignoring the crucial role of temporal coherence, the formulation of most of present computer vision tasks lead us to tackle problems that are remarkably more difficult than those nature has prepared for us!

4.4 How can Humans Perform Pixel Semantic Labeling?

Many object recognition systems are based on an opportune pre-processing of video information represented by a vector, which is subsequently processed for class prediction. Surprisingly enough, the state of the art approaches that follow this guideline already offer quite accurate performance in real-world contexts, without relying on the semantic labelling of each pixel. Basically, a global computational scheme emerges that is typically made more and more effective when the environment in which the machine is supposed to work is quite limited, and it is known in advance. The number of the classes that one expects to recognize in the environment affects the performance, but very high accuracy can be achieved without necessarily being able to perform the object segmentation and, therefore, without needing to perform pixel semantic labeling. However, for an agent to conquer visual capabilities in a broad context, it seems to be very useful to rely on more specific visual skills. When thinking of a video, the information that one can extract is not only driven by time but also by spatial information. We humans can easily describe a scene by locating the objects in specific positions, and we can describe their eventual movement. This requires a deep integration of visual and linguistic skills, that are required to come up with compact, yet effective video descriptions. However, in any case humans can successfully provide a very accurate labeling of single pixels, which leads us to pose the following question:

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

Interestingly, in humans semantic pixel labelling is by driven by the focus of attention, another fundamental features that, as we will see in the remainder of the paper, is at the core of all important computational processes of vision. While pixel-based decisions are inherently interwound with a certain degree of ambiguity, they are remarkably effective. The linguistic attributes that we can extract are related to the context of the pixel that is taken into account for label attachment, while the ambiguity is mostly a linguistic more than a visual issue. In a sense, this primitive is likely in place for conquering higher abstraction levels. How can this be done? The focus on single pixels allows us to go beyond object segmentation based on sliding windows. Instead of dealing with object proposals (Zitnick and Dollár, 2014), a more primitive task is that of attaching symbols to single pixels in the retina. The task of semantic pixel labelling leads to focussing attention on the given pixel, while considering the information in its neighborhood. This clearly opens the doors to an in-depth re-thinking of pattern recognition processes. It is not only the frame content, but also where we focus attention in the retina that does matter.

Human ability of exhibiting semantic labeling at pixel level is really challenging. The visual developmental processes conquer this ability nearly without pixel-based supervisions. It seems that such a skill is mostly the outcome of the acquisition of the capability to perform object segmentation. This is obtained by constructing the appropriate memberships of the pixels that define the segmented regions. When thinking of the classic human communication protocols, one early realizes that even though it is rare to provide pixel-based supervision, the information that is linguistically conveyed to describe visual scenes makes implicit reference to the focus of attention. This holds regardless of the scale of the visual entity being described. Hence, the emergence of the capability of performing pixel semantic label seems to be deeply related to the emergence of focus of attention mechanisms. The most striking question, however, is how can humans construct such a spectacular segmentation without a specific pixel-based supervision! Interestingly, we can focus on a pixel and attach meaningful labels, without having been instructed for that task.

BOX 4 | The primitive of pixel semantic labelling is likely crucial for the construction of human-like visual skills. There should be a hidden supervisor in nature that, so far, has nearly been neglected. We conjecture that it is the optical flow which plays the central role for object recognition. The decision on its recognition must be invariant under motion, a property that does require a formulation in the temporal direction.

4.5 What is the Role of Receptive Fields and Hierarchical Architectures?

Beginning from early studies on the visual structure of the cortex (Hubel and Wiesel, 1962), neuroscientists have gradually gained evidence that it presents a hierarchical structure and that neurons process the video information on the basis of inputs restricted to receptive fields. Interestingly, the recent spectacular results of convolutional neural networks suggests that hierarchical structures based on neural computation with receptive fields play a fundamental role also in artificial neural networks (LeCun et al., 2015). The following questions naturally arise:

> **Q5:** *Why are the visual mainstreams organized according to a hierarchical architecture with receptive fields? Is there any reason why this solution has been developed in biology? Why is its "replication" in neural networks so successful?*

First of all, we can promptly realize that, even though neurons are restricted to compute over receptive fields, deep structures rely on large virtual contexts for their decision. As we increase the depth of the neural network, the consequent pyramidal dependence that is established by the receptive fields increases the virtual input window used for the decision, so that higher abstraction is progressively gained as we move towards the output. Hence, while one gives up to exploit all the information available at a certain layer, the restriction to receptive field does not prevent from considering large windows for the decision. The marriage of receptive fields with deep nets turns out to be an important ingredient for a parsimonious and efficient implementation of both biological and artificial networks. In convolutional neural networks, the assumption of using receptive fields comes with the related hypothesis of *weight sharing* on units that are supposed to extract the same feature, regardless of where the neurons are centered in the retina. In so doing we enforce the extraction of the same features across the retina. This makes sense whereas, in general, it does not make sense to extract features depending on the pixel in the retina. The same visual clues are clearly positioned everywhere in the retina and the equality constraints on the weights turn out to be a precise statement for implementing a sort of *invariance under translation*.

Clearly, this constraint has neither effect on invariance under scale nor under rotation. Any other form of invariance that is connected with deformable objects is clearly missed and is supposed to be learned. The current technology of convolutional neural networks in computer vision typically gains these invariances thanks to the power of supervised learning by "brute force." Notice that since most of the tasks involve object recognition in a certain environment, the associated limited amount of visual information allows to go beyond the principle of extracting visual features at pixel level. Visual features can be shared over small windows in the retina by the process of pooling, thus limiting the dimension of the network. Basically, the number of features to be involved has to be simply related to the task at hand, and we can go beyond the association of the features with the pixels. However, the acquisition of human-like visual skills is not compatible with this kind of simplifications since, as stated in the previous section, humans can perform pixel semantic labeling. There is a corresponding trend in computer vision where convolutional nets are designed to keep the connection with each pixel in the retina at any layer so as to carry out segmentation and semantic pixel label. Interestingly, this is where we need to face a grand challenge. So far, very good results have been made possible by relying on massive labelling of collections of images. While image labeling for object classification is a boring task, human pixel labeling (segmentation) is even worse! Instead of massive supervised labelling, one could realize that motion and focus of attention can be massively exploited to learn the visual features mostly in an unsupervised way. A recent study in this direction is given in (Betti and Gori, 2018), where the authors provide evidence of the fact that receptive fields do favor the acquisition of motion invariance which, as already stated, is the fundamental invariance of vision. The study of motion invariance leads to dispute the effectiveness and the biological plausibility of convolutional networks. First, while weight sharing is directly gained by translational invariance on any neuron, the vice versa clearly does not hold. Hence, we can think of receptive field based neurons organized in a hierarchical architecture that carry out translation invariance without sharing their weights. This is strongly motivated also by the arguable biological plausibility of the mechanism of weight sharing (Ott et al., 2020). Such a lack of plausibility is more serious than the supposed lack of a local computational scheme in Backpropagation, which mostly comes

from the lack of delay in the forward model of the neurons (Betti and Gori, 2019).

BOX 5 | Hierarchical architectures and receptive fields seems to be tightly connected in the development of abstract representations. However, we have reasons to doubt that weight sharing happens in biological networks and that the removal of this constraint facilitates the implementation of motion invariance. The architectural incorporation of this fundamental invariance property, as well as the match with the need for implementing the focus of attention mechanisms likely needs neural architectures that are more sophisticated than current convolutional neural networks. In particular, neurons which provide motion invariance likely benefit from dropping the weight sharing constraint.

4.6 Why Two Different Main Visual Processing Streams?

In **Section 2** we have emphasized the importance of bearing in mind the neat functional distinction between vision for action and vision for perception. A number of studies in neuroscience lead to the conclusion that the visual cortex of humans and other primates is composed of two main information pathways that are referred to as the ventral stream and the dorsal stream (Goodale and Milner, 1992; Goodale and Keith Humphrey, 1998). We typically refer to the ventral "what" and the dorsal "where/how" visual pathways. The ventral stream is devoted to perceptual analysis of the visual input, such as object recognition, whereas the dorsal stream is concerned with providing spatial localization and motion ability in the interaction with the environment. The ventral stream has strong connections to the medial temporal lobe (which stores long-term memories), the limbic system (which controls emotions), and the dorsal stream. The dorsal stream stretches from the primary visual cortex (V1) in the occipital lobe forward into the parietal lobe. It is interconnected with the parallel ventral stream which runs downward from V1 into the temporal lobe.

> ***Q6:*** *Why are there two different mainstreams? What are the reasons for these a different neural evolutions?*

This neurobiological distinction arises for effectively facing visual tasks that are very different. The exhibition of perceptual analysis and object recognition clearly requires computational mechanisms that are different with respect to those required for estimating the scale and the spatial position. Object recognition requires the ability of developing strong invariant properties that mostly characterize the objects themselves. By and large scientists agree that objects must be recognized independently of their position in the retina, scale, and orientation. While we subscribe to this point of view, a more careful analysis of our perceptual capabilities indicates that these desirable features are likely more adequate to understand the computational mechanisms behind the perception of rigid objects. The elastic deformation and the projection into the retina gives in fact rise to remarkably more complex patterns that can hardly be interpreted in the framework of geometrical invariances. We reinforce the claim that *motion invariance is in fact the only invariance which does matter*. Related studies in this direction can be found (Bertasius et al., 2021). As the nose of a teddy bear approaches child's eyes it becomes larger and larger. Hence, scale invariance is just a byproduct of motion invariance. The same holds true for rotation invariance. Interestingly, as a child deforms the teddy bear a new visual pattern is created that, in any case, is the outcome of the motion of "single object particles." The neural enforcement of motion invariance likely takes place in the "what" neurons. Of course, neurons with built-in motion invariance are not adequate to make spatial estimations or detection of scale/rotation. Unlike the "what" neurons, in this case motion does matter and the neural response must be affected by the movement.

BOX 6 | These analyses are consistent with neuroanatomical evidence and suggest that "what" and "where" neurons are important also in machines. The anatomical difference between the two processing streams is in fact the outcome of a different functional role. While one can ignore such a difference and rely on the rich representational power of big deep networks, the underlined difference stimulates the curiosity of discovering canonical neural structures to naturally incorporate motion invariance, with the final purpose being that of discovering different features for perception and action.

4.7 Why do Some Animals Focus Attention?

It is well-known that the presence of the fovea in the retina leads to focus attention on details in the scene. Such a specialization of the visual system is widespread among vertebrates, it is present in some snakes and fishes, but among mammals is restricted to haplorhine primates. In some nocturnal primates, like the owl monkey and in the tarsier, the fovea is morphologically distinct and appears to be degenerate. An owl monkey's visual system is somewhat different from other monkeys and apes. As its retina develops, its dearth of cones and its surplus of rods mean that this focal point never forms. Basically, a fovea is most often found in diurnal animals, thus supporting the idea that it is supposed to play an important role for capturing details of the scene (Ross, 2004). But why haven't many mammals developed such a rich vision system based on foveate retinas? Early mammals, which emerged in the shadow of the dinosaurs, were likely forced to live nocturnal lives so as to avoid to become their pray (Sohn, 2019). In his seminal monograph, Gordon Lynn Walls (Walls, 1942) proposed that there has been a long nocturnal evolution of mammals' eyes, which is the reason of the remarkable differences with respect to those of other vertebrates. The idea became known as the "nocturnal bottleneck" hypothesis (Gerkema et al., 2013). Mammals' eyes tended to resemble those of nocturnal birds and lizards, but this does not hold for humans and closely related monkeys and apes. It looks they re-evolved features useful for diurnal living after they abandoned a nocturnal lifestyle upon dinosaur extinction. It is worth mentioning that haplorhine primates are not the only mammals which focus attention in the visual environment. Most mammals have quite a well-developed visual system for dealing with details. For example, it has been shown that dogs possess quite a good visual system that share many features with

those of haplorhine primates (Beltran et al., 2014). A retinal region with a primate fovea-like cone photoreceptor density has been identified but without the excavation of the inner retina. A similar anatomical structure, that has been observed in rare human subjects, has been named fovea-plana. Basically, the results in (Beltran et al., 2014) challenge the dogma that within the phylogenetic tree of mammals, haplorhine primates with a fovea are the sole lineage in which the retina has a central bouquet of cones. In non-primate mammals, there is a central region of specialization, called the *area centralis*, which also is often located temporal to the optic axis and demonstrates a local increase in photoreceptor and retinal ganglion cell density that plays a role somehow dual with respect to the fovea. Like in haplorhine primates, in those non-primate mammals we experience focus of attention mechanisms that are definitely important from a functional viewpoint.

This discussion suggests that the evolution of animals' visual system has followed many different paths that, however, are related to focus of attention mechanisms, that are typically more effective for diurnal animals. There is, however, an evolution path which is definitely set apart, in which the frog is most classic representer. More than 60 years ago, the visual behavior of the frog posed an interesting question (Lettvin et al., 1959) which is mostly still on the table. In the words of the authors:

> 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.

No mammal experiments such a surprising behavior! However, the frog is not expected to eat like mammals. When tadpoles hatch and get free, they attach themselves to plants in the water such as grass weeds, and cattails. They stay there for a few days and eat tiny bits of algae. Then the tadpoles release themselves from the plants and begin to swim freely, searching out algae, plants and insects to feed upon. At that time their visual system is ready. Their food requirements are definitely different from what mammals need and their visual system has evolved accordingly for catching flying insects. Interestingly, unlike mammals, the studies in (Lettvin et al., 1959) already pointed out that the frogs' retina is characterized by uniformly distributed receptors with neither fovea nor *area centralis*. Interestingly, this means that the frog does not focus attention by eye movements. When the discussion focuses on functional issues the following natural questions arise:

> **Q7:** *Why are the fovea and the area centralis convenient? Why do primates and other animals focus attention, whereas others, like the frog, do not?*

One can easily argue that any action that animals carry out needs to prioritize the frontal view. On the other hand, this leads to the detriment of the peripheral vision, that is also very important. In addition, this could apply for the dorsal system whose neurons are expected to provide information that is useful to support movements and actions. Apparently, the ventral mainstream, with neurons involved in the "what" function, does not seem to benefit from foveate eyes. Apart from recent developments, most state of the art computer vision models for object recognition, just like frogs, do not focus attention, since they carry out a uniform massive parallel computation on the retina. Just like frogs, the cameras used in computer vision applications are uniformly distributed, but machines seem to conquer human-like recognition capabilities on still images. Interestingly, unlike frogs, machines recognize quite well food properly served in a bowl. This capability might be due to the current strongly artificial communication protocol. Machines benefit from supervised learning of tons of supervised pairs, a process which, as already pointed out, cannot be sustained in nature. On the other hand, as already pointed out, in order to attack the task of understanding what is located in a certain position, it is natural to think of eyes based on fovea or on area centralis. The eye movements with the corresponding trajectory of the focus of attention (FOA) is also clearly interwound with the temporal structure of video sources. In particular, humans experience eye movements when looking at fixed objects, which means that they continually experience motion. Hence, also in case of fixed images, conjugate, vergence, saccadic, smooth pursuit, and vestibulo-ocular movements lead to the acquisition of visual information from relative motion. We claim that the production of such a continuous visual stream naturally drives feature extraction, since the corresponding convolutional filters, charged with representing features for object recognition, are expected not to change during motion. The enforcement of this consistency condition creates a mine of visual data during animal life! Interestingly, the same can happen for machines. Of course, we need to compute the optical flow at the pixel level so as to enforce the consistency of all the extracted features. Early studies on this problem (see Horn and Schunck 1981), along with related improvements (see e.g., Baker et al., 2011) suggests to determine the velocity field by enforcing brightness invariance. As the optical flow is gained, it can be used to enforce motion consistency on the visual features. These features can be conveniently combined with those responsible of representing objects. Early studies driven by these ideas are reported in Gori et al., (2016), where the authors propose the extraction of visual features as a constraint satisfaction problem, mostly based on information-theoretic principles and early ideas on motion invariance.

The following remarks on focus of attention coming from nature seem to be important for conquering efficient visual skills for any intelligent agent. Basically, it looks like we are faced with functional issues which mostly obeys information-based principles.

- *The FOA drives the definition of visual primitives at pixel level.* The already mentioned visual skill that humans possess to perform pixel semantic labeling clearly indicates the capability of focusing on specific points in the retina with high resolution. Hence, FOA is needed if we want to perform such a task.

- *Eye movements and FOA help estimating the probability distribution on the retina*. At any time a visual agent clearly needs to possess a good estimation of the probability distribution over the pixels of the retina. This is important whenever we consider visual tasks for which the position does matter. This involves both the *where* and *what* neurons. In both cases it is quite obvious that any functional risk associated with the given task should avoid reporting errors in regions of the retina where there is a uniform color. The probability distribution is of fundamental importance and it is definitely related to saliency maps built on focus of attention trajectories.
- *FOA very well fits the need for receptive fields and deep nets*. We have already discussed the marriage between receptive fields and deep networks. Interestingly, the FOA mechanisms emphasize the role of a single receptive field in the computational process that takes place at any time. The saccadic movements contribute to perform "temporally segmented computations" over the retina on the different sequences produced by micro-saccadic movements. In addition, in (Betti and Gori, 2018), the authors provide evidence of the fact that receptive fields do favor the development of biologically-plausible models based on local differential equations.
- *Eye movements and FOA are the basis for establishing invariant laws.*. The interplay between the FOA and the invariance properties is the key for understanding human vision and general principles that drive object recognition and scene interpretation. In order to understand the nice circle that is established during the processes of learning in vision, let us start exploring the very nature of eye movements in humans. Basically, they produce visual sequences that are separated by saccadic movement, during which no information is acquired[3]. Interestingly, each of those sequences is composed of pixels that somehow share common visual features. In case of micro-saccades the corresponding micro-movements explore regions with a remarkable amount of details that are somehow characterized by certain features. The same holds true for smooth pursuit, where the invariance of the extracted feature turns out to be a sort of primitive consistency property: objects do not change during their motion. Hence, any visual feature associated with "what neurons" must be invariant under any eye movement, apart from saccadic movements. Clearly, such an invariance has a true unsupervised nature. A deep net based on the discussed convolutional structure can in fact learn a set of latent features to be motion invariant This results in an impressive collection of "labelled data" that nature offers for free. The eye movements and the FOA significantly contribute to enhance the motion invariance since, as already pointed out, humans always experience motion in a frame of reference located on the retina. When thinking of the "what" neurons for which invariances need to be imposed, we can promptly realize that those which are close to the output are better suited for the enforcing of invariances. It is in fact quite obvious that for many of those invariances to take place we need a strong computational capability of the "what neuron." Interestingly, this seems to suggest that "where" neurons could better be located in the early layers of the hierarchy whereas "what" neurons require higher abstraction.
- *FOA helps disambiguation at learning time*. A puzzle is offered at learning time when two or more instance of the same object are present in the same frame, maybe with different poses and scales. The FOA in this case helps disambiguating the enforcement of motion invariance. While the enforcement of weight sharing is ideal for directly implementing translation invariance, such a constraint doesn't facilitate other more complex invariances that can better be achieved by its removal.
- *FOA drives the temporal interpretation of scene understanding*. The importance of FOA is not restricted to feature and object invariance, since it involves also the interpretation of visual scenes. It is in fact the way FOA is driven which sequentially selects the information for conquering the scene interpretation. Depending on the purpose of the agent and on its level of scene understanding the FOA is consequently moved. This process clearly shows the fundamental role of the selection of the points where to focus attention, an issue which is described in the following section.
- *FOA helps disambiguating illusions*. Depending on where an agent with foveate eyes focuses attention, concepts that, strictly speaking, don't exist can emerge, thus creating an illusion. A noticeable example is the Kanizsa's triangle, but it looks like other illusions arise for related reasons. You can easily experiment that as you approach any detail, it is perfectly perceived without any ambiguity. A completion mechanism arises that leads us to perceive the triangle as soon as you move away from figure and the mechanism is favored by focussing attention on the barycenter. Interestingly, the different views coming from different points where an agent with foveate eyes focuses attention likely helps disambiguating illusions, a topic that has been recently studied in classic convolutional networks Kim et al. (2019), Baker et al. (2018).
- *FOA helps to address the problem of "concept drift"*. When discussing motion invariance we mentioned the problem of concept drift. Clearly, this could dramatically affect the practical implementation of the motion invariance. However, amongst different types of FOA trajectories, the saccadic movements play the fundamental role of resetting the process, which clearly faces directly problems of concept drift.

The analysis on foveated-based neural computation nicely explains also the reason why humans cannot see video with a number of frames per second that exceeds the classic sampling

[3]There is in fact a rich literature on this topic, from which it is clearly stated that subject cannot see his own saccades in a mirror, that is there is in fact *saccadic suppression* (Matin, 1974).

threshold. It turns out that this number is clearly connected with the velocity of the scan paths of the focus of attention. Of course, this is a computational issue which goes beyond biology and clearly affects machines as well.

BOX 7 | The above items provide strong evidence for the reasons why foveate eyes turn out to be very effective for scene understanding. Interestingly, we can export the information-based principle of focussing attention to computer retinas by simulating eye movements. There is more: machines could provide multiple focuses of attention which could increase their visual skills significantly.

4.8 What Drives Eye Movements?

Foveate animals need to move their eyes to properly focus attention. The previous discussion has emphasized the importance of performing appropriate movements, which motivates the following question naturally arises:

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

Human eyes make jerky saccadic movements during ordinary visual acquisition. One reason for these movements is that the fovea provides high-resolution in portions of about 1, 2°. Because of such a small high resolution portions, the overall sensing of a scene does require intensive movements of the fovea. Hence, the foveate movements do represent a good alternative to eyes with a uniformly high resolution retina. The information-based principles discussed so far lead us to conclude that foveate retinas with saccadic movements is in fact a solution that is computationally sustainable and very effective. Fast reactions to changes in the surrounding visual environment require efficient attention mechanisms to reallocate computational resources to most relevant locations in the visual field. While current computational models keep improving their predictive ability thanks to the increasing availability of data, they are still far away from the effectiveness and efficiency exhibited by foveate animals. An in-depth investigation on biologically-plausible computational models of focus of attention that exhibit spatiotemporal locality is very important also for computer vision, where one relies on parallel and distributed implementations. The research carried out by (Faggi et al., 2020) suggests an interpretation based on a computational model where attention emerges as a wave propagation process originated by visual stimuli corresponding to details and motion information. The resulting field obeys the principle of *inhibition of return*, so as not to get stuck in potential holes, and extend previous studies in (Zanca et al., 2020) with the main objective of providing spatiotemporal locality. In particular, the idea of modeling the focus of attention by a gravitational process finds its evolution in the corresponding local model based on the Poisson equation on the corresponding potential. Interestingly, Newtonian gravity yields an instantaneous propagation of signals, so as a sudden change in the mass density of a given pixel immediately affects the focus of attention, regardless of its location on the retina. These studies are driven by the principle that there are in fact sources which drive attention (e.g., masses in a gravitational field). At early cognitive stages, attention mechanisms are mostly driven by the presence of details and movements. This is the reason why the mentioned masses for modeling the focus of attention have been based on the magnitude of the gradient of the brightness in the retina and on the optical flow. Interestingly, in children the mechanisms that drive the focus of attention are strongly connected with the developmental stages. Newborns and children in their early stages of evolution only focus attention on details and movements and on a few recurrent visual patterns like faces. As time goes by, visual features acquire a semantic value and, consequently, the focus of attention is gradually driven by specific intentions and corresponding plans. Of course, this is only possible after having acquired some preliminary capability of recognizing objects. Interestingly, as the forward process that facilitate high level cognitive tasks from the focus of attention becomes effective a corresponding backward process begins the improvement of the focus of attention. A reinforcement loop is generated which is finalized to optimize the final purpose of the agent in its own learning environment.

BOX 8 | What drives the focus of attention is definitely a crucial issue, simply because of its already discussed fundamental role. We conjecture that this driving process must undergo a developmental process, where we begins with details and optical flow and proceed with the fundamental feedback from the environment which is clearly defined by the specific purpose of the agent.

4.9 Why is Baby Vision Blurred?

There are surprising results that come from developmental psychology on what a newborns see. Basically, their visual acuity grows gradually in early months of life. Interestingly, Charles Darwin had already noticed this very interesting phenomenon. In his words:

> It was surprising how slowly he acquired the power of following with his eyes an object if swinging at all rapidly; for he could not do this well when seven and a half months old.

At the end of the seventies, this early remark was given a technically sound basis (see, e.g., Dobson and Teller 1978). In the paper, three techniques, — optokinetic nystagmus (OKN), preferential looking (PL), and the visually evoked potential (VEP)— were used to assess visual acuity in infants between birth and 6 months of age. More recently (Braddick and Atkinson, 2011), provides an in-depth discussion on the state of the art in the field. It is clearly stated that for newborns to gain adult visual acuity, depending on the specific visual test, several months are required. The following question naturally arises:

> **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?*

This brings up the discussion on the "protection" of the learning agent from information overloading, which might be

of fundamental importance also in computer vision. The blurring of the video at an early stage of learning is compatible with a broader view of learning that, in addition to the involvement of the classic synaptic connections, this provides a direct "simplification of the input." Regardless of this specific input modification, the underlying idea is that the process of learning consists of properly filtering the input with the purpose of gradually acquiring the information. The development of any computation model that adheres to this view is based on modifying the connections along with an appropriate input filtering so that the learning agent always operates at an equilibrium point (Betti et al., 2021).

BOX 9 | When promoting the role of time, the arising pre-algorithmic framework suggests extending the learning process to an appropriate modification of the input, that is finalized to achieve the expected "visual acuity" at the end of the process of learning. In doing so, one can think of generalization processes that are based on the convergence to fixed values of the weight connections, but also on an opportune "small perturbation" of the input.

4.10 What is the Interplay With Language?

The interplay of vision and language is definitely one of the most challenging issues for an in-depth understanding of human vision. Along with the associated successes, the indisputable adoption of the supervised learning protocol in most challenging object recognition problems caused the losing of motivations for an in-depth understanding of the way linguistic information is synchronized with visual clues. In particular, the way humans learn the name of objects is far away from the current formal supervised protocol. This can likely be better grasped when we begin considering that top level visual skills can be found in many animals (e.g., birds and primates), which clearly indicates that their acquisition is independent of language.

Hence, as we clarify the interplay of vision and language we will likely address also the first question on how to overcome the need for "intensive artificial supervision." Since first linguistic skills arise in children when their visual acuity is already very well developed, there is a good chance that early simple associations between objects and their names can easily be obtained by "a few supervisions" because of the very rich internal representation that has already been gained of those objects. It is in fact only a true independent hidden representation of objects which makes possible their subsequent association with a label! The capability of learning motion-invariance features is a fundamental information-based principle regardless of biology, which might somehow drives the development of "what" neurons.

The interplay of language and vision has been very well addressed in a survey by Lupyan (2012). It is claimed that performance on tasks that have been presumed to be non-verbal is rapidly modulated by language, thus rejecting the distinction between verbal and non-verbal representations. While we subscribe to the importance of sophisticated interactions, we also reinforce the claim that capturing the identity of single objects is mostly a visual issue. However, when we move towards the acquisition of abstract notions of objects than the interaction with language is likely to be very important. One needs to separate single objects coming in visual contexts with their own identity with respect to abstract notions of objects. We can see a specific chair, but it's a different story to recognize that we have a chair in front of us.

In **Section 4.7** we addressed the issue of motion invariance by claiming that it must properly be considered in the unified framework of focus of attention. We experience eye movements either on still images or during motion. In the first case we can see micro-saccadic movements, whereas moving objects are properly tracked. While the enforcement of visual feature invariance makes sense in both cases, there is a fundamental difference from an information-based viewpoint: The object tracking does provide information on the object movement, so that one can propagate the label to all the pixels that are connected with the pixel where we focus attention by a non-vanishing optical flow. This conveys an enormous amount of labelled information on the moving object and on its related segmentation. There is more! While during micro-saccadic movements many invariant features can be developed and it is not clear which one–if any–refers to an explicit object as a whole, during smooth pursuit, thanks to the optical flow, the moving object, with its own label, provides an internal representation gained under this motion invariance that is likely to be the secret for bridging the linguistic attachment of labels to objects.

The opportune exploitation of optical flow in visual information is of paramount importance for the evolution of theories of vision. In the last few years we have also seen a number of contributions in egocentric vision, where the assumption is that also the camera is moving. Interestingly, any sophisticated filtering of such an external movement might neglect the importance of undergoing developmental steps just like those that are fundamental for capturing the interplay with language. Clearly, if you have already gained good visual skills in object recognition, it is quite easy to check whether you yourself are moving!

Once again, the discussion carried out so far promotes the idea that for a visual agent to efficiently obtain the capabilities of recognizing objects from a few supervisions, it must undergo some developmental steps aimed at developing invariant representations of objects, so the actual linguistic supervision takes place only after the development of those representations. But, when should we enable a visual agent to begin with the linguistic interaction? While one might address this question when attacking the specific computational model under investigation, a more natural and interesting way to face this problem is to re-formulate the question as:[4]

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

[4]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

This is done in a spectacular way in nature! Like vision, language development requires a lot of time. Interestingly, it looks like it requires more than vision. The discussion in **Section 4.9** indicates that the gradual growth of visual acuity is a possible clue to begin with language synchronization. The discussed filtering process offers a protection from visual information overloading that likely holds for language as well. As the visual acuity gradually increases, one immediately realizes that the mentioned visual-language synchronization has a spatiotemporal structure. At a certain time, we need to inform the agent about what we see at a certain position in the retina. An explicit implementation of such an association can be favored by an active learning process: the agent can ask itself what is located at (x, t). However, what if you cannot rely on such a precious active interactions? For example, a linguistic description of the visual environment is generally very sophisticated and mentions objects located in different positions of the retina, without providing specific spatiotemporal information. Basically, this is a sort of *weak supervision* that is more difficult to grasp. However, once again, developmental learning schemes can significantly help. At early stage of learning the agent's tasks can be facilitated by providing spatiotemporal information. For example, naming the object located where the agent is currently focussing attention conveys information by a sort of human-like communication protocol. As time goes by, the agent gains gradually the capability of recognizing a few objects. What really matters is the confidence that is gained in such a task. When such a developmental stage is reached, linguistic descriptions and any sort of natural language based visual communication can be conveniently used to reinforce the agent recognition confidence. Basically, these weak supervisions turn out to be very useful since they can profitably be attached where the agent came up with a prediction that matches the supervision.

BOX 10 | Computer vision and natural language processing have been mostly evolving independently one each other. While this makes sense, the time has come to explore the interplay between vision and language with the main purpose of going beyond the protocol of supervised learning for attaching labels to objects. Interestingly challenges arises in scene interpretation when we begin considering the developmental stages of vision that suggest gaining strong object invariance before the attachment of linguistic labels.

5 THE "EN PLAIN AIR" PERSPECTIVE

Posing the right questions is the first fundamental step to gain knowledge and solve problems. The intent of this paper is to provide insights and to contribute to a shift in the direction in which computer vision is presently being practiced in the deep learning community. However, one might wonder what could be the most concrete action for promoting studies on the posed questions. So far, computer vision has strongly benefited from the massive diffusion of benchmarks which, by and large, are regarded as fundamental tools for performance evaluation. However, it is clear that they are very well-suited to support the statistical machine learning approach based on huge collections of labelled images. This paper, however, opens the doors to explore a different framework for performance evaluation. The emphasis on video instead of images does not leads us to think of huge collection of video, but to adopt a different approach in which no collection at all is accumulated! Just like humans, machines are expected to live in their own visual environment. What should be the scientific framework for evaluating the performance and understand when a theory carries out important new results? Benchmarking bears some resemblance to the influential testing movement in psychology which has its roots in the turn-of-the-century work of Alfred Binet on IQ tests (Binet and Simon, 1916). Both cases consist of attempts to provide a rigorous way of assessing the performance or the aptitude of a (biological or artificial) system, by agreeing on a set of standardized tests which, from that moment onward, become the ultimate criterion for validity. On the other hand, it is clear that the skills of any visual agent can be quickly evaluated and promptly judged by humans, simply by observing its behavior. Thus, we could definitely rely on a *crowdsourcing performance evaluation scheme* where registered people can inspect and assess the performance of software agents (Gori et al., 2015). We use the *term en plein air* to mimic the French Impressionist painters of the 19th-century and, more generally, the act of painting outdoors. This term suggests that visual agents should be evaluated by allowing people to see them in action, virtually opening the doors of research labs. The en plein air proposal allows others to test our algorithms and to contribute to this evaluation method by providing their own data, their own results, or the comparisons with their own algorithms.

While the idea of shifting computer vision challenges into the wild will deserves attention one cannot neglect the difficulties that arise from the lack of a truly lab-like environment for supporting the experiments. The impressive progress in computer graphics, however, offers a very attractive alternative that can dramatically facilitate the developments of approaches to computer vision that are based on the on-line treatment of the video (see, e.g., Meloni et al., 2020).

Needless to say, computer vision has been fueled by the availability of huge labelled image collections, which clearly shows the fundamental role played by pioneer projects in this direction (see, e.g., Deng et al., 2009). The ten questions posed in this paper will likely be better addressed only when scientists will put more emphasis on the en plein air environment. In the meantime, the major claim of this paper is that the experimental setting needs to move to virtual visual environments. Their photorealistic level along with the explosion of the generative capabilities makes these environments better suited to new performance evaluation of computer vision.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Files, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Aloimonos, J., Weiss, I., and Bandyopadhyay, A. (1988). Active Vision. *Int. J. Comput. Vis.* 1, 333–356. doi:10.1007/bf00133571

Baker, S., Scharstein, D., Lewis, J. P., Roth, S., Black, M. J., and Szeliski, R. (2011). A Database and Evaluation Methodology for Optical Flow. *Int. J. Comput. Vis.* 92, 1–31. doi:10.1007/s11263-010-0390-2

Baker, N., Erlikhman, G., Kellman, P. J., and Lu, H. (2018). "Deep Convolutional Networks Do Not Perceive Illusory Contours," in Proceedings of the 40th Annual Meeting of the Cognitive Science Society, CogSci 2018, Madison, WI, USA, July 25-28, 2018. Editors C. Kalish, M. A. Rau, X. J. Zhu, and T. T. Rogers. cognitivesciencesociety.org.

Ballard, D. H. (1991). Animate Vision. *Artif. Intell.* 48, 57–86. doi:10.1016/0004-3702(91)90080-4

Beltran, W. A., Cideciyan, A. V., Guziewicz, K. E., Iwabe, S., Swider, M., Scott, E. M., et al. (2014). Canine Retina Has a Primate Fovea-like Bouquet of Cone Photoreceptors Which Is Affected by Inherited Macular Degenerations. *PLOS ONE* 9, 1–10. doi:10.1371/journal.pone.0090390

Benjamin, W. T., Mary, M. H., Michael, F. L., and Dana, H. B. (2011). Eye Guidance in Natural Vision: Reinterpreting Salience. *J. Vis.* 11, 1–23. doi:10.1167/11.5.5

Bertasius, G., Wang, H., and Torresani, L. (2021). *Is Space-Time Attention All You Need for Video Understanding?* arXiv:2102.05095

Betti, A., and Gori, M. (2016). The Principle of Least Cognitive Action. *Theor. Comput. Sci.* 633, 83–99. doi:10.1016/j.tcs.2015.06.042

Betti, A., and Gori, M. (2018). *Convolutional Networks in Visual Environments*. Arxiv preprint arXiv:1801.07110v1.

Betti, A., and Gori, M. (2019). *Backprop Diffusion Is Biologically Plausible*. CoRR abs/1912.04635.

Betti, A., Gori, M., and Melacci, S. (2018). *Cognitive Action Laws: The Case of Visual Features*. CoRR abs/1808.09162

Betti, A., Gori, M., and Melacci, S. (2021). *Learning and Visual Blurring*. Technical Report. SAILab.

Binet, A., and Simon, T. (1916). *The Development of Intelligence in Children: The Binet?Simon Scale*. Williams & Wilkins.

Borenstein, E., and Ullman, S. (2002). "Class-specific, top-down segmentation," in *Computer Vision - ECCV 2002, 7th European Conference on Computer Vision, Copenhagen, Denmark, May 28-31, 2002, Proceedings, Part II. Lecture Notes in Computer Science*. Editors A. Heyden, G. Sparr, M. Nielsen, and P. Johansen (Springer), 2351, 109–124. doi:10.1007/3-540-47967-8_8

Braddick, O., and Atkinson, J. (2011). Development of Human Visual Function. *Vis. Res.* 51, 1588–1609. doi:10.1016/j.visres.2011.02.018

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: A Large-Scale Hierarchical Image Database," in CVPR09, Miami, FL, June 22–24, 2009. doi:10.1109/cvpr.2009.5206848

Dobson, V., and Teller, D. Y. (1978). Visual Acuity in Human Infants: A Review and Comparison of Behavioral and Electrophysiological Studies. *Vis. Res.* 18, 1469–1483. doi:10.1016/0042-6989(78)90001-9

Faggi, L., Betti, A., Zanca, D., Melacci, S., and Gori, M. (2020). *Wave Propagation of Visual Stimuli in Focus of Attention*. CoRR abs/2006.11035.

Gerkema, M., Davies, W., Foster, R., Menaker, M., and Hut, R. (2013). The Nocturnal Bottleneck and the Evolution of Activity Patterns in Mammals. *Proc. R. Soc. Lond. Ser. B, Biol. Sci.* 280, 20130508. doi:10.1098/rspb.2013.0508

Gibson, J. J., and Boston (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Gibson, J. J. (1950). *The Perception of the Visual World*. Houghton Mifflin.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.

Goodale, M. A., and Keith Humphrey, G. (1998). The Objects of Action and Perception. *Cognition* 67, 181–207. doi:10.1016/s0010-0277(98)00017-1

Goodale, M. A., and Milner, A. D. (1992). Separate Visual Pathways for Perception and Action. *Trends Neurosci.* 15, 20–25. doi:10.1016/0166-2236(92)90344-8

Gori, M., Lippi, M., Maggini, M., Melacci, S., and Pelillo, M. (2015). "En plein air visual agents," in *Image Analysis and Processing - ICIAP 2015 - 18th International Conference, Genoa, Italy, September 7-11, 2015, Proceedings, Part II. Lecture Notes in Computer Science*. Editors V. Murino and E. Puppo (Springer), 9280, 697–709. doi:10.1007/978-3-319-23234-8_64

Gori, M., Lippi, M., Maggini, M., and Melacci, S. (2016). Semantic Video Labeling by Developmental Visual Agents. *Computer Vis. Image Understanding* 146, 9–26. doi:10.1016/j.cviu.2016.02.011

Goroshin, R., Bruna, J., Tompson, J., Eigen, D., and LeCun, Y. (2015). "Unsupervised Learning of Spatiotemporally Coherent Metrics," in 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, 4086–4093. doi:10.1109/ICCV.2015.465

Horn, B. K., and Schunck, B. (1981). Determining Optical Flow. *Artif. Intell.* 17, 185–203. doi:10.1016/0004-3702(81)90024-2

Hubel, D., and Wiesel, T. (1962). Receptive fields, Binocular Interaction, and Functional Architecture in the Cat's Visual Cortex. *J. Physiol. (London)* 160, 106–154. doi:10.1113/jphysiol.1962.sp006837

Kim, B., Reif, E., Wattenberg, M., and Bengio, S. (2019). *Do Neural Networks Show Gestalt Phenomena? an Exploration of the Law of Closure*. CoRR abs/1903.01069.

Kingstone, A., Daniel, S., and John, D. E. (2010). Cognitive Ethology: A New Approach for Studying Human Cognition. *Br. J. Psychol.* 99, 317–340. doi:10.1348/000712607x251243

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep Learning. *Nature* 521, 436–444. doi:10.1038/nature14539

Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009). "Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations," in Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09 (New York, NY, USA: ACM), 609–616. doi:10.1145/1553374.1553453

Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., and Pitts, W. H. (1959). What the Frog's Eye Tells the Frog's Brain. *Proc. IRE* 47, 1940–1951. doi:10.1109/jrproc.1959.287207

Lupyan, G. (2012). Linguistically Modulated Perception and Cognition: The Label-Feedback Hypothesis. *Front. Psychol.* 3, 54. doi:10.3389/fpsyg.2012.00054

Marinai, S., Gori, M., and Soda, G. (2005). Artificial Neural Networks for Document Analysis and Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 23–35. doi:10.1109/tpami.2005.4

Marr, D. (1982). *Vision*. San Francisco: Freeman. Partially reprinted in Anderson and Rosenfeld (1988).

Matin, E. (1974). Saccadic Suppression: A Review and an Analysis. *Psychol. Bull.* 81, 899–917. doi:10.1037/h0037368

Meloni, E., Pasqualini, L., Tiezzi, M., Gori, M., and Melacci, S. (2020). *Sailenv: Learning in Virtual Visual Environments Made Simple*. CoRR abs/2007.08224.

Ott, J., Linstead, E., LaHaye, N., and Baldi, P. (2020). Learning in the Machine: To Share or Not to Share? *Neural Networks* 126, 235–249. doi:10.1016/j.neunet.2020.03.016

Pan, S., and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi:10.1109/tkde.2009.191

Poggio, T. A., and Anselmi, F. (2016). *Visual Cortex and Deep Networks: Learning Invariant Representations*. 1st edn. The MIT Press.

Ranzato, M., Huang, F. J., Boureau, Y., and LeCun, Y. (2007). "Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition," in 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), Minneapolis, Minnesota, USA, 18-23 June 2007. doi:10.1109/CVPR.2007.383157

Rao, R. P. N., and Ballard, D. H. (1999). Predictive Coding in the Visual Cortex: a Functional Interpretation of Some Extra-Classical Receptive-Field Effects. *Nat. Neurosci.* 2, 79–87. doi:10.1038/4580

Ronneberger, O., Fischer, P., and Brox, T. (2015). *U-net: Convolutional Networks for Biomedical Image Segmentation*. CoRR abs/1505.04597. doi:10.1007/978-3-319-24574-4_28

Ross, C. F. (2004). *The Tarsier Fovea: Functionless Vestige or Nocturnal Adaptation?* Boston, MA: Springer US, 477–537. doi:10.1007/978-1-4419-8873-7_19

Sohn, E. (2019). The Eyes of Mammals Reveal a Dark Past. *Nature*. doi:10.1038/d41586-019-01109-6

Tavanaei, A., Masquelier, T., and Maida, A. S. (2016). *Acquisition of Visual Features through Probabilistic Spike-timing-dependent Plasticity*. CoRR abs/1606.01102. doi:10.1109/ijcnn.2016.7727213

Ullman, S. (1979). *The Interpretation of Visual Motion/Shimon Ullman*. The MIT press series in artificial intelligence (The MIT press.

Walls, G. L. (1942). *The Vertebrate Eye and its Adaptive Radiation*.

Watanabe, S. (1985). *Pattern Recognition: Human and Mechanical*. USA: John Wiley & Sons.

Wood, J. N., and Wood, S. M. (2020). One-shot Learning of View-Invariant Object Representations in Newborn Chicks. *Cognition* 199, 104192. doi:10.1016/j.cognition.2020.104192

Wood, J. N. (2016). A Smoothness Constraint on the Development of Object Recognition. *Cognition* 153, 140–145. doi:10.1016/j.cognition.2016.04.013

Zanca, D., Melacci, S., and Gori, M. (2020). Gravitational Laws of Focus of Attention. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 2983–2995. doi:10.1109/TPAMI.2019.2920636

Zitnick, C. L., and Dollár, P. (2014). "Edge Boxes: Locating Object Proposals from Edges," in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, 391–405. doi:10.1007/978-3-319-10602-1_26

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.