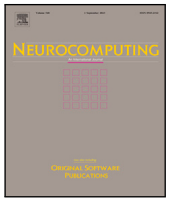
Neurocomputing 560 (2023) 126775

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Local propagation of visual stimuli in focus of attention

Lapo Faggi [a,b,*], Alessandro Betti [c], Dario Zanca [d], Stefano Melacci [b], Marco Gori [b,c]

[a] DINFO, University of Florence, Florence, Italy
[b] DIISM, University of Siena, Siena, Italy
[c] Université Côte d'Azur, Inria, CNRS, Laboratoire I3S, Maasai team, Nice, France
[d] AIBE, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

ARTICLE INFO

Communicated by J. Cao

Dataset link: https://gitlab.com/mela64/localfoa

Keywords:
Computer vision
Visual attention
Scanpath prediction
Saliency

ABSTRACT

Fast reactions to changes in the surrounding visual environment require efficient attention mechanisms to reallocate computational resources to the most relevant locations in the visual field. In this paper, we present a biologically-plausible computational model of focus of attention that exhibits spatiotemporal locality and that is very well-suited for parallel and distributed implementations. 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. The proposed model is obtained as a hyperbolic regularization of the Poisson equation to which it reduces in the limit of high speed of propagation. According to the MultiMatch algorithm for scanpaths comparison, the proposed model achieves very competitive results when considering dynamical input stimuli.

1. Introduction

Visual attention plays a central role in our daily activities. While we are playing, teaching a class or driving a vehicle, the amount of information our eyes collect is way greater than what we are able to process [1,2]. To work properly, we need a mechanism that only locates the most relevant objects, thus optimizing the computational resources [3]. Human visual attention performs this task so efficiently that, at a conscious level, it goes unnoticed. Visual attention has also been proposed as a fundamental mechanism for grouping low-level features into coherent and unitary objects [4]. The selection mechanism performed by visual attention can be driven by different factors. On the one hand, attention is naturally attracted by salient regions in the visual scenes, which are those areas that appear to stand out with respect to their neighboring parts (*bottom-up stimulus-driven* attention). On the other hand, attention is also strongly influenced by high-level and cognitive factors that drive attention to elements relevant to a behavior, such as looking for a particular object in the visual scene, making a sandwich [5] or driving a car [6] (*top-down goal driven attention*). Semantic object dependencies [7], global scene context [8], emotions and expectations can also influence attention as well. Some of these and other top-down factors also affect attention in *free-viewing conditions*, as confirmed by the experimental findings of [9,10] and [11]. Despite different neural processes drive bottom-up and top-down attentional mechanisms, it has been shown that they share a common neural apparatus, the frontoparietal network, that is essential for both [12].

Given the fundamental role of spatial attention in the human visual system, computational models of visual attention have been the subject of massive investigation, and have proven their usefulness in many applications, especially whenever they are asked to solve tasks related to human perception such as video compression, where loss of quality is not perceivable by viewers [13,14], or caption generation [15,16]. Following the seminal works by Treisman et al. [17,18] and Koch and Ullman [19], as well as the first computational implementations [20], over the last three decades scientists have presented numerous attempts to model the Focus Of Attention (FOA) mechanism [21]. The notion of *saliency map* has been introduced, which consists of a spatial map that indicates the probability of focusing on each pixel. In the last decade, deep neural networks allowed the scientific community to make strong progresses in saliency estimation, overcoming classical bottom-up saliency models [22]. Usually, established convolutional models devised for object recognition are exploited, along with some possible architectural innovations, to predict saliency. To suffice the lack of large fixation datasets, transfer learning is employed, and deep saliency models are pre-trained on large image datasets and then fine-tuned to predict accurate saliency maps. Among the many neural approaches that have been proposed so far for saliency estimation we

* Corresponding author at: DINFO, University of Florence, Florence, Italy.
E-mail addresses: lapo.faggi@unifi.it (L. Faggi), alessandro.betti@inria.fr (A. Betti), dario.zanca@fau.de (D. Zanca), mela@diism.unisi.it (S. Melacci), marco.gori@unisi.it (M. Gori).

https://doi.org/10.1016/j.neucom.2023.126775
Received 13 October 2022; Received in revised form 7 August 2023; Accepted 11 September 2023
Available online 18 September 2023

mention, for example, the Saliency Attentive Model presented in [23], which considers a neural attention mechanism based on a convolutional LSTM architecture to iteratively refine saliency predictions. Similarly, a visual attention network is proposed in [24] to capture hierarchical saliency information by combining multi-scale details originating from different convolutional layers within the same architecture. In this case, supervisions are fed into both the earlier and last layers of the network. The authors of [25] propose an adversarial framework for saliency estimation and, in order to enlarge existing saliency datasets, they present a novel one made up of data-augmented images.

However, saliency models neglect the temporal dimension of the intrinsically dynamic process of attention, since the order of the fixations is not taken into account. Discarding such an important aspect may critically lead to a poor description of the phenomenon [26]. Note that, assuming to have computed a meaningful saliency map, we can still generate shifts in visual attention through a winner-take-all mechanism [19], selecting the most relevant location in space at each time step. Some authors have tried to improve the quality of the generated scanpaths by introducing a hand-crafted human bias to choose subsequent fixations [27]. Similarly, [28] tries to formalize the idea that during visual exploration high-level cues continue to increase their importance, to the disadvantage of more perceptive low-level information. In [29], the authors propose a bio-inspired visual attention model based on the pragmatic choice of certain proto-objects and learning the order in which these are attended. In [30], a saliency map is used to stochastically generate visual shifts by means of a constrained random walk. Some authors have also proposed biologically-inspired computational models in which attentional shifts are explained and deeply interlaced with top-down object identification capabilities [31–34]. In particular, in SAIM [31] and its extensions [32,33], the visual input is mapped into the focus through a "content" network, whose activity is spatially modulated by a "selection" network. A "knowledge" network then matches the content of the focus – that is, a properly modulated region of the original input – with the given template units. The selection and identification processes are thus developed in parallel until the convergence of neural units is reached, time at which the object is fully recognized and selected. Thereafter, attentional shifts are obtained through an inhibition mechanism that, unlike standard ones, operates both at spatial and semantic level. The evolution of the various networks is driven by a set of differential equations, derived from energy minimization principles that express the desire model's output as constraints. Note that, while these models define a dynamics of the attentional process, they do so at the level of the underlying neuronal circuits rather than investigating the spatio-temporal dynamics of the focus itself over the visual scene, which is the aim of this paper. Data-driven approaches have also been proposed to estimate scanpaths. In [35], a generative network is used to generate attentional shifts, complemented by an adversarial discriminator model trained to distinguish human scanpaths from simulated ones. A recent approach [36] proposes using a dynamic priority map, influenced by both semantic content and fixation history, to guide the fixational process. Here, the inhibition of return mechanism is modeled as a convolutional LSTM network, while mixture density networks are used to predict probability distributions of fixations for each pixel. Note, that, unlike the scanpaths models we have introduced above, the model presented in this paper is based only on low-level features and does not need any training phase.

For our ensuing discussion, it is important to realize that the majority of these approaches rely on a long stack of global computations over the entire visual field before establishing the next fixation point, and this is especially true for all those models in which scanpaths are generated, through a winner-take-all mechanism or a probabilistic sampling, from an underlying saliency map. This is hardly compatible with what is done by humans where, most likely, attention modulates visual signals before they even reach the cortex [37,38] and restricts computation to a small portion of the available visual information [17,39]. Moreover, by doing so, the spatial and temporal continuity that is inherent to the attentional phenomenon is overlooked. On the contrary, Zanca et al. recently proposed an approach that is inspired by physics to directly model the process of visual attention as a continuous dynamic phenomenon [40,41]. In [41], the FOA is modeled as a point-like particle gravitationally attracted by virtual masses originating from details and motion in the visual scene. Masses due to details are determined by the magnitude of the gradient of the brightness, while masses due to motion are proportional to the norm of the optical flow. This framework can be applied to both images and videos, as long as one considers a static image as a video whose frames are repeated at each time step. Unlike the other described approaches, the prediction of the focus does not rely on a saliency map, but it acts directly on early representations of basic features organized in spatial maps. Besides the advantage in real-time applications, these models make it possible to characterize patterns of eye movements (such as *fixations*, *saccades* and *smooth pursuit*) and, despite their simplicity, they reach state-of-the-art results in scanpath prediction, also overcoming the classic winner-take-all approach [42]. However, when analyzing the gravitational model [41] from the computational perspective, one promptly realizes that determining the position of the FOA at a certain time instant requires to have access to all the visual information of the retina, in order to sum up the attraction arising from all the virtual masses. In other words, the model is not local in space and also lacks temporal coherence between consecutive frames. Moreover, being it based on Newtonian gravity, it introduces an instantaneous propagation of the visual signals, so that a sudden change in the mass density of a given pixel immediately affects the focus of attention, regardless of its location on the retina.

Still focusing on predicting human-like scanpaths, this paper proposes a paradigm shift in the computation of the gravitational potential for modeling attention, following a novel viewpoint that allows the model to exhibit spatio-temporal locality, and this could also shed light on the biological processes behind the FOA mechanism [43]. To achieve this, we introduce an explicit temporal dynamic in the static Poisson equation, so that the evolution of the potential is described through a wave-like propagation mechanism. As the velocity of the propagation goes to infinity, the gravitational model of [41] is recovered. Moreover, locality allows the computation of the potential to be carried out independently at each pixel of the retina, so that the proposed model is very well-suited for Single Instruction Multiple Data (SIMD) specialized implementations. A detailed experimental analysis confirms that the information coming from virtual masses is properly transmitted to the entire retina and, in the case of scanpath prediction in video signals, the proposed model achieves state-of-the-art results.

2. Methods

2.1. Computing the FOA trajectory in gravitational models

According to the gravitational model of [41], the FOA is modeled as a point-like particle in a gravitational potential $\varphi^0 : \mathbb{R}^2 \times [0, T] \to \mathbb{R}$. At each time instant t, such potential is determined by the input video stream, through the mass density function $\mu(x, t)$, where $x \in \mathbb{R}^2$, that is responsible for "attracting" the focus of attention. Such mass distribution involves information about the visual details and motion of the input stream, and it includes an Inhibition Of Return (IOR) mechanism. In detail, the trajectory of the focus of attention $t \in [0, T] \mapsto a(t) \in \mathbb{R}^2$, starting at $a(0) = a_0$ with velocity $\dot{a}(0) = a_1$, is the solution of the following Cauchy problem (Newton's second law with damping term):

$$\begin{cases} \ddot{a}(t) + \varpi \dot{a}(t) + \nabla \varphi^0(a(t), t) = 0; \\ a(0) = a_0; \\ \dot{a}(0) = a_1, \end{cases} \tag{1}$$

where $\varpi > 0$ and the scalar function $\varphi^0 : \mathbb{R}^2 \times [0,T] \to \mathbb{R}$ is the 2-D gravitational potential defined as

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

Here $|\cdot|$ is the Euclidean norm in $\mathbb{R}^2$ and $\mu : R \subset \mathbb{R}^2 \times [0,T] \to [0,+\infty)$ is the mass distribution at a certain temporal instant that is present on the retina and is given by

$$\mu(x,t) = \mu_1(x,t)\,(1 - I(x,t)) + \mu_2(x,t). \tag{3}$$

In particular, $\mu_1 = \alpha_1 |\nabla b|$, where $b : \mathbb{R} \times [0,T] \to \mathbb{R}$ is the brightness, while $\mu_2 = \alpha_2 |v|$, where $v : \mathbb{R} \times [0,T] \to \mathbb{R}$ is the optical flow and α_1 and α_2 are positive parameters. The term $I(x,t)$ implements the IOR mechanism by decreasing the mass distribution associated with details in those areas of the retina that have already been explored in the previous moments. This is achieved requiring $I(x,t)$ to satisfy the differential equation

$$I_t + \beta I = \beta \exp(-|x - a(t)|^2 / 2\sigma^2), \tag{4}$$

with $0 < \beta < 1$ (I_t is the time derivative of I).

Due to the integral of Eq. (2), the model of [41] is not local in space. Furthermore, the fact that φ^0 is expressed through this integral is strictly related to the fact that φ^0 satisfies the Poisson equation on $\mathbb{R}^2$: $-\nabla^2 \varphi = \mu$ where ∇^2 is the *Laplacian* in two dimensions. This can be verified by direct calculation exploiting that (see for example [44]):

1. the function $G(x) := 1/(2\pi) \log(1/|x|)$, defined for $x \in \mathbb{R}^2$, $x \neq 0$, is the fundamental solution of Laplace equation (i.e. $-\nabla^2 G = \delta$);
2. to get a solution of the Poisson equation $-\nabla^2 u(x) = f(x)$ in $\mathbb{R}^2$, when f is regular and compactly supported, it is sufficient to choose u as the convolution of G with f.

2.2. Spatio-temporal local visual attention

Since the mass density μ is time-dependent and its temporal dynamic is synced with the variations of the video, determining the FOA trajectory requires φ^0 to be found at each frame from scratch, thus ignoring any temporal relation. This is essentially related to the fact that Poisson equation is an elliptic PDE, and it does not involve any temporal dynamics. The main idea behind the formulation presented in this paper is that since we expect that small temporal changes in the source μ cause small changes in the solution φ, then it is natural to model the potential φ by dynamical equations which prescribe, for each spatial point x, how the solution must be updated depending on its spatial neighborhood at time $t - dt$. In practice, this is achieved with a novel formulation based on a hyperbolic wave-like equation governing the evolution of the potential. Elliptic PDEs are well suited to describe static equilibrium states, while hyperbolic equations regard true dynamical processes so that it appears natural to use this kind of equations to model the FOA mechanism. Unlike [41], the resulting computational scheme is both local in space and coherent in time.

We introduced an explicit temporal dynamics in the Poisson equation by considering the following regularization[1]

$$\gamma \varphi_{tt}(x,t) + \lambda \varphi_t(x,t) = c^2 \nabla^2 \varphi(x,t) + \mu(x,t) \tag{5}$$

where $c > 0$ $\lambda \geq 0$ is the drag coefficient, $\gamma \geq 0$ and φ_t (φ_{tt}) is the first (second) time derivative of φ. Such equation in one spatial dimension (and without the source term μ) is known as the telegraph equation (see [44]). More generally, it describes the propagation of a damped wave. The pure diffusion case corresponds to $\gamma = 0$, with a diffusion coefficient equals to $\alpha = c^2/\lambda$ and a source term of $\mu(\vec{x},t)/\lambda$. With $\gamma = 1$ and $\lambda = 0$ we obtain the pure wave equation instead.

[1] We refer the reader to Appendix A for an in-depth analysis of hyperbolic and parabolic regularizations of the Poisson equation.

When paired with the appropriate numerical implementation, the proposed formulation leads to a computational scheme that is local in both space and time, which is a fundamental ingredient of biological plausibility. Spatial locality also opens for significant technological improvements, since it can be exploited to build computational schemes that are well-suited for SIMD hardware implementations, allowing a fast computation of the potential. The following subsection give a detailed description of the numerical implementation of the proposed model and the stability limit of the associated (explicit) numerical schemes.

2.3. Computational schemes

FOA trajectories are determined by solving the ordinary differential equation (1), once the gravitational potential has been evaluated at the current FOA position. This can be done by introducing an auxiliary variable $z(t) = \dot{a}(t)$, thus reducing Eq. (1) to a system of first order equations that can be numerically solved through classical methods such as the Euler's one.[2]

The complete problem for the PDE (5), with boundary condition, that we used in all our experiences is

$$\begin{cases} \gamma \varphi_{tt}(x,t) + \lambda \varphi_t(x,t) = c^2 \nabla^2 \varphi(x,t) + \mu(x,t) & \text{in} \quad \mathbb{R}^2 \times (0,+\infty); \\ \varphi(x,0) = 0, \quad \varphi_t(x,0) = 0 & \text{in} \quad \mathbb{R}^2 \times \{t = 0\}, \end{cases} \tag{6}$$

The FOA model proposed in this paper is thus based on Eqs. (6) along with the inhibition of return equation expressed by (4) and the FOA equation of motion (1), see Fig. 1 for a comparison with [41]. Clearly, Eq. (6) is local in both space and time, which is a fundamental ingredient of biological plausibility. Notice that, at a first sight, Eq. (4) does not possess spatial locality. While this holds true in any computer-based retina, in nature, moving eyes rely on the principle that you can simply pre-compute $\exp(-|x - a(t)|^2/2\sigma^2)$ by an appropriate foveal structure.

This section is dedicated to devising local and stable computational schemes to evaluate the gravitational potential by solving Eq. (6). We start from the trivial consideration that, from a computational point of view, all the operations are limited to a finite region of space, the open and bounded subset $R \subseteq \mathbb{R}^2$ that we will denote as the *retina*. Thus, to determine the potential and its time evolution on R we have to impose additional boundary conditions on ∂R. We adopt Dirichlet boundary conditions, requiring the vanishing of the potential on the boundary, $\varphi(x,t) = 0$ on ∂R, $\forall t$. Then, Eq. (6) becomes

$$\begin{cases} \gamma \varphi_{tt}(x,t) + \lambda \varphi_t(x,t) = c^2 \nabla^2 \varphi(x,t) + \mu(x,t) & \text{in} \quad R \times (0,+\infty); \\ \varphi(x,t) = 0 & \text{in} \quad \partial R \times (0,+\infty); \\ \varphi(x,0) = 0, \quad \varphi_t(x,0) = 0 & \text{in} \quad R \times \{t = 0\}. \end{cases} \tag{7}$$

Spurious reflections originating from the boundary are avoided through the dumping term since, given an appropriate choice of the parameters, out-going waves are suppressed before they can reach the boundary.[3]

The first step to numerically solve the above update equation is to discretize both the retina and the time dimension. The former is discretized considering a mesh $M = \{(i,j) \in \mathbb{R}^2 : i = 0, \dots h-1, \quad j = 0, \dots, w-1\}$ of $h \times w$ points (pixels). The latter is discretized with steps of length Δt, where Δt is chosen according to the numerical analysis and the corresponding stability limits derived in this section. We adopt the so-called *finite difference method*, approximating spatial and temporal derivatives through finite differences. Considering an arbitrary pixel

[2] In our implementation, we have exploited a standard numerical routine (scipy.odeint) to perform such a computation.

[3] In the experiments, we have also considered an additional area external to R, that covers 100 additional pixels on each side. This, other than helping in avoiding spurious boundary reflections, also fosters the exploration of the entire visual scene decreasing model's bias towards the center of the input stimulus.

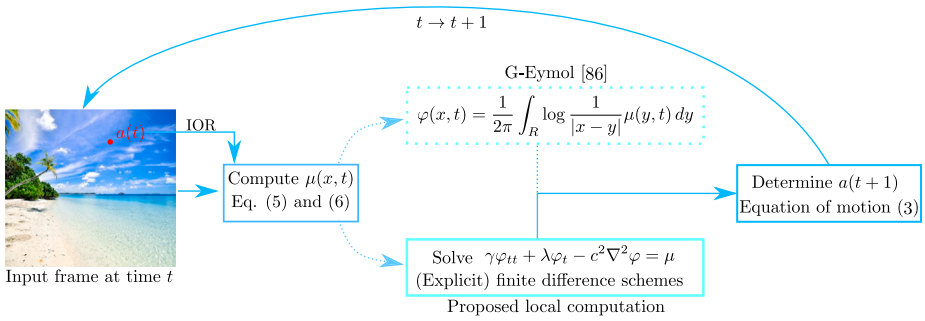


Fig. 1. Comparison between G-Eymol [41] and the proposed local computational scheme. G-Eymol includes an integral on the entire retina, without considering any temporal dependencies, while the proposed model locally updates the computation both in space and time.

(i, j) of the retina at a certain time t, the evaluation of spatial (temporal) derivatives of the potential in this point will just require the knowledge of the potential in its adjacent points in space (time). According to the chosen approximations for the derivatives, the unknown value of the potential at the following time instant is determined by a set of algebraic equations (*explicit methods*) or by a set of coupled equations (*implicit methods*). As we will show in what follows, implicit schemes are less afflicted by numerical instabilities (they are *unconditionally* stable) with respect to explicit ones. On the other hand, explicit models are computationally cheaper and better suited for a parallel or distributed implementation. Notice that any finite difference scheme devised to solve the discretized Poisson equation results in a system of coupled equations, so that the corresponding computational model would still lack spatial locality other than temporal coherence.

We consider three finite differences algorithms, where two of them (EX1 and EX2) are explicit and the last one (IMP) is implicit. In the case of EX1 we consider centered difference approximations for all the time and spatial derivatives (both of the first and second order),

$$\gamma\frac{\varphi^{n+1}_{i,j} - 2\varphi^{n}_{i,j} + \varphi^{n-1}_{i,j}}{\Delta t^2} + \lambda\frac{\varphi^{n+1}_{i,j} - \varphi^{n-1}_{i,j}}{2\Delta t} = \mu^{n}_{i,j} + c^2\left(\frac{\varphi^{n}_{i,j+1} - 2\varphi^{n}_{i,j} + \varphi^{n}_{i,j-1}}{\Delta x^2} + \frac{\varphi^{n}_{i+1,j} - 2\varphi^{n}_{i,j} + \varphi^{n}_{i-1,j}}{\Delta y^2}\right), \tag{8}$$

where $\varphi^{n}_{i,j}$ is the discrete variable of the problem and $\mu^{n}_{i,j}$ is the value of the density mass function $\mu(x,t)$ on the spatial mesh point i, j at the time-step n. In the EX2 scheme, the first order time derivative is approximated by a forward difference instead:

$$\gamma\frac{\varphi^{n+1}_{i,j} - 2\varphi^{n}_{i,j} + \varphi^{n-1}_{i,j}}{\Delta t^2} + \lambda\frac{\varphi^{n+1}_{i,j} - \varphi^{n}_{i,j}}{\Delta t} = \mu^{n}_{i,j} + c^2\left(\frac{\varphi^{n}_{i,j+1} - 2\varphi^{n}_{i,j} + \varphi^{n}_{i,j-1}}{\Delta x^2} + \frac{\varphi^{n}_{i+1,j} - 2\varphi^{n}_{i,j} + \varphi^{n}_{i-1,j}}{\Delta y^2}\right). \tag{9}$$

As for the IMP scheme, we consider backward finite difference approximations for the time derivatives (both of the first and second order) and a central one for the spatial derivatives,

$$\gamma\frac{\varphi^{n}_{i,j} - 2\varphi^{n-1}_{i,j} + \varphi^{n-2}_{i,j}}{\Delta t^2} + \lambda\frac{\varphi^{n}_{i,j} - \varphi^{n-1}_{i,j}}{\Delta t} = \mu^{n-1}_{i,j} + c^2\left(\frac{\varphi^{n}_{i,j+1} - 2\varphi^{n}_{i,j} + \varphi^{n}_{i,j-1}}{\Delta x^2} + \frac{\varphi^{n}_{i+1,j} - 2\varphi^{n}_{i,j} + \varphi^{n}_{i-1,j}}{\Delta y^2}\right). \tag{10}$$

One can easily verify that all these three numerical schemes are *consistent* with the partial differential equation (PDE) in Eq. (7). In particular, a given numerical scheme is said to be consistent with a PDE if by reducing the distances among the knots in the space mesh and the time step size, the exact solution of the PDE satisfies the corresponding discretized equation at least up to first-order in the step sizes, i.e., the so-called *truncation error* is at least a linear function of Δx, Δy and Δt, that approaches 0 as Δx, Δy and Δt become arbitrary small [45,46]. The accuracy of the EX1 case turns out to be of the second-order both in space and time (quadratic function of the Δ terms), while it is of the first-order in time and second-order in space in EX2 and IMP. It is worth noting that, in all the experiment reported in the next sections, we have always considered the EX2 potential update Eq. (9). This is because, even if it is less accurate than the EX1 scheme, it is still stable in the pure diffusion case $\gamma = 0$.

2.4. Stability analysis

We now focus on computing the *stability* bounds related to EX1, EX2 and IMP schemes, mainly following the conventions of [47]. The importance of the stability analysis is emphasized by the fact that, when paired with the already discussed *consistency*, it allows us to conclude on the *convergence* of the considered schemes, i.e., ensuring that the numerical solution of the finite difference schemes tends to a solution of problem (7) as Δx, Δy and Δt tend to zero. In particular, convergence in linear PDEs (that is our case) is an immediate consequence of the Lax–Richtmyer equivalence theorem, reported below [45,47].

Theorem (*Lax–Richtmyer Equivalence Theorem*)**.** Given a properly posed linear initial-value problem and a linear finite-difference approximation to it that satisfies the consistency condition, stability is the necessary and sufficient condition for convergence.

The core idea in defining stability is that the temporal evolution of the chosen numerical scheme should limit the amplification of all the components of the initial condition, and this also implies the boundedness of rounding errors as the computations proceeds over time. Stability is a condition related to the numerical scheme and its associated solution solely (it makes no reference to the original PDE). The boundedness of the solution is expressed in terms of its L^2 norm. In order to make the stability analysis easier, we will assume an unbounded domain, conducting a Fourier analysis of the finite difference schemes, following the so-called *Von Neumann stability analysis* approach [45–48]. Within this assumption, we are implicitly neglecting instabilities that may emerge from the numerical treatment of the true boundary conditions.[4] Finally, it is worth noting that our stability analysis will focus on the homogeneous version of Eq. (7). Due to the Duhamel principle, this will not invalidate our conclusions [47].

We will start describing our stability analysis considering a linear PDE of the first-order in time with constant coefficients and an associated numerical scheme in which time derivatives are approximated considering two different time instants, indexed by n and $n-1$ (that is, a single-step numerical scheme). Without loss of generality, we also restrict to one spatial dimension. Then, we will generalize our findings to multi-step schemes and, finally, we will include second-order time derivatives in the original PDE (matching the cases of EX1,

[4] Despite of this, the qualitative experimental analysis reported in Appendix B confirms that our predictions on the various stability limits are quite accurate and that instabilities originating from the boundaries do not occur.

EX2, IMP). Given a single-step numerical scheme associated to a linear PDE of the first-order in time with constant coefficients, we can write for the Fourier components $\hat{\varphi}^n(k) := (2\pi)^{-1/2} \sum_{j=-\infty}^{+\infty} e^{-ik\Delta x j} \varphi_j^n \Delta x$ of the solution at the time-step n (see [47]):

$$\hat{\varphi}^n(k) = g\,(k\Delta x, \Delta x, \Delta t)\,\hat{\varphi}^{n-1}(k) \tag{11}$$

that implies

$$\hat{\varphi}^n(k) = \left(g\,(k\Delta x, \Delta x, \Delta t)\right)^n \hat{\varphi}^0(k)\,, \tag{12}$$

where $k\Delta x \in [-\pi\,, \pi]$ and we are considering each Fourier component separately thanks to the linearity of the scheme. We indicated with function g what is known as the *amplification factor*, that regulates the amplification of each component. The Von Neumann stability condition – a necessary and sufficient one in this particular case – requires

$$\exists\, K > 0: \qquad |g| \leq 1 + K\Delta t, \qquad \forall\, k\Delta x \in [-\pi\,, \pi]\,. \tag{13}$$

In the case of multi-steps schemes, the analysis involves the computation of the roots of the so-called *amplification polynomial* $\Phi(g)$, which in general depends on $k\Delta x, \Delta x$ and Δt. This can be defined as the polynomial obtained substituting $\varphi_j^n = g^n e^{ik\Delta x j}$ in the homogeneous version of the selected finite difference scheme and then canceling out the common factors $g^n e^{ik\Delta x j}$ (see [47] for details), where i is the imaginary unit. In general, considering a scheme involving τ time steps (that is, $\tau + 1$ different time instances), Φ is a polynomial of order τ in g. The resulting stability criteria is analogous to Eq. (13). In particular, indicating with g_i's the roots of the amplification polynomial, the stability condition is

$$\exists\, K > 0: \quad |g_i| \leq 1 + K\Delta t, \quad \forall\, k\Delta x \in [-\pi\,, \pi]\,,\ \forall\, i. \tag{14}$$

Considering schemes associated to second-order equations, the amplification polynomial Φ has always at least two roots,[5] and the coalescence of two different roots near the unit circle is allowed [47]. In order to keep the following derivations more manageable, we drop the dependence on K in (14) by requiring a stronger stability condition[6]

$$|g_i| \leq 1, \forall\, k\Delta x \in [-\pi\,, \pi]\,,\ \forall\, i, \tag{15}$$

so that, according to the classification of [47,49], our amplification polynomials must be of the second-order Von Neumann type.[7]

All the schemes we have analyzed (EX1, EX2, IMP) consider just three different time instants. Then, the corresponding amplification polynomials are of the second order in g (with real coefficients) and the multiplicity condition on their roots is automatically satisfied, since the coalescence of two roots is allowed. Then, in the cases of our interest, the amplification polynomials always assume the form

$$\Phi\,(g) = A + Bg + Cg^2\,, \tag{16}$$

where A, B and C are real constants. In order to easily determine if their roots lie in the unit circle, we can finally exploit the following lemma [49].

Lemma 1. *The two roots g_i of a polynomial of the form (16) lie in the unit disk ($|g_i| \leq 1$), iff either:*

1 $(C - A)(C + A) > 0$ *and* $(B - A - C)(B + A + C) \leq 0$
2 $(C - A) = 0$ *and* $(B - 2A)(B + 2A) \leq 0$
3 $(C + A) = 0$ *and* $B = 0$

[5] In order to properly approximate a second order time derivative, three different time levels are required at least.

[6] In principle, one has to require this restricted condition only when the amplification polynomial Φ does not depend on Δx and Δt.

[7] Its roots g_i lie in the unit circle $|g_i| \leq 1$ and the multiplicity of those roots for which $|g_i| = 1$ is at most 2.

For the EX1 scheme one finds that (for the derivations of the results presented below, see Appendix B)

$$\begin{cases} C = \gamma + \dfrac{\lambda\Delta t}{2} \\ B = -2\gamma + 4\left(C_x^2 \sin^2\left(\dfrac{k_x\Delta x}{2}\right) + C_y^2 \sin^2\left(\dfrac{k_y\Delta y}{2}\right)\right) \\ A = \gamma - \dfrac{\lambda\Delta t}{2} \end{cases} \tag{17}$$

where $C_x = c\Delta t/\Delta x$, $C_y = c\Delta t/\Delta y$ and (k_x, k_y) are the components of the two-dimensional wave vector. The scheme is *conditionally stable* and, according to the Von Neumann stability condition (15) and Lemma 1, we have to require $C_x^2 + C_y^2 \leq \gamma$ for $\gamma > 0, \lambda > 0$ (point 1 of Lemma 1, $(C - A)(C + A) = (\lambda\Delta t/2)(2\gamma) > 0$), $C_x^2 + C_y^2 \leq \gamma$ for $\gamma > 0, \lambda = 0$ (point 2, $C - A = 0$), while for $\gamma = 0$ and $\lambda > 0$ the scheme is unstable (point 3, $(C + A) = 0$ but $B > 0$ for a general wave vector (k_x, k_y)). The EX2 is conditionally stable as well. The amplification polynomial is determined by

$$\begin{cases} C = \gamma + \lambda\Delta t \\ B = -2\gamma - \lambda\Delta t + 4\left(C_x^2 \sin^2\left(\dfrac{k_x\Delta x}{2}\right) + C_y^2 \sin^2\left(\dfrac{k_y\Delta y}{2}\right)\right) \\ A = \gamma \end{cases} \tag{18}$$

and we have to require the stability bound $C_x^2 + C_y^2 \leq \gamma + \lambda\Delta t/2$ for $\gamma \geq 0\,, \lambda \geq 0$ with $(\gamma, \lambda) \neq (0, 0)$ (point 1, $(C - A)(C + A) = (\lambda\Delta t)(2\gamma + \lambda\Delta t) > 0$ if $\lambda \neq 0$ and point 2 if $\lambda = 0$). Finally, the IMP scheme is unconditionally stable. The amplification polynomial is given by

$$\begin{cases} C = \gamma + \lambda\Delta t + 4\left(C_x^2 \sin^2\left(\dfrac{k_x\Delta x}{2}\right) + C_y^2 \sin^2\left(\dfrac{k_y\Delta y}{2}\right)\right) \\ B = -2\gamma - \lambda\Delta t \\ A = \gamma \end{cases} \tag{19}$$

and the stability condition (15) is always satisfied for $\gamma \geq 0\,, \lambda \geq 0$ with $(\gamma, \lambda) \neq (0, 0)$ (point 1, $(C - A)(C + A) > 0$ and $(B - A - C)(B + A + C) \leq 0$ $\forall\, (k_x\Delta x, k_y\Delta y) \in (-\pi/2, \pi/2)$). We confirmed the validity of the reported stability bounds by means of a concrete experimental activity, reported in Appendix B.

3. Experimental setup

We evaluated the proposed model in the scanpath prediction task and, for completeness, also in saliency prediction, comparing it with several state-of-the-art models. In this Section, we describe all the details involved in our experimental activity.

3.1. Datasets

We considered different human eye-tracking datasets, that represent well-established benchmarks for the evaluation of computational models of visual attention. In particular, in the scanpath prediction task we used four datasets, that are MIT1003 [50], TORONTO [51], KOOTSTRA [52], SIENA12 [53], for a total of 1234 input stimuli. All these datasets provide temporal information of human visual explorations collected in free-viewing conditions. Human subjects were exposed to the stimulus from 3 up to 5 s. In the case of dynamical input stimuli, we have considered the COUTROT database 1 [54], which contains 60 different videos. The temporal resolution of each video is 25 frames per second and their average length is 17 seconds. Also in this case eye-tracking data were collected in free-viewing conditions, and 72 subjects were exposed to 4 different auditory conditions. To evaluate the performance of the model in the saliency prediction task, we have considered the CAT2000 [55] and MIT300 [56] datasets, provided by the MIT Saliency Team [57,58]. The CAT2000 test set includes 2000 input stimuli, grouped into 20 different semantic categories, while the MIT300 dataset consists of 300 stimuli.

3.2. Experimental details, parameter selection and sensitivity analysis

In each experiment, all the images and video frames have been resized to 224 × 224 pixels and converted to gray-scale. We have also included 100 additional pixels on each side of the image to avoid spurious boundary reflections and to foster the exploration of the entire visual scene. For each tested model, 10 different visual explorations have been generated for each stimulus numerically integrating Eq. (1), initializing the system with various randomly generated initial conditions. In all the experiments, the potential has been always computed through the EX2 update Eq. (9). Different initializations result in different FOA trajectories and slightly differing dynamics of the potential, due to the IOR mechanism of Eq. (4). Therefore, in the proposed model, randomness is injected in the initial conditions and not in the dynamic of the FOA itself, as, for example, is done in [30]. Since our goal is to emulate human scanpaths, some sort of randomness is necessary to take into account the stochasticity exhibited by different human subjects when looking at the same scene, or even by the same subject when inspecting an identical stimulus more times. Since human subjects during free-viewing are usually asked to look at a target point in the center of the screen before a visual stimulus is presented, we have initialized the FOA position within a range of 2.5° of visual angle from the center of the stimulus, with a random initial velocity close to zero. Then, each scanpath has been obtained numerically integrating Eq. (1). Finally, whenever we needed to extract fixations, we did it by processing the continuous simulated scanpaths according to the following criteria. Given a spatial threshold $r_{\max}$ and a temporal one $t_{\min}$, a fixation is detected in a certain position if the FOA lies in the corresponding surrounding area (determined by $r_{\max}$) for at least $t_{\min}$. To this end, we have exploited the Python PYGAZE package [59] that extracts fixations from raw data of eye-tracker devices. Threshold values have been set with default ones, that are designed to extract human-like fixations ($r_{\max} = 25$ pixels, $t_{\min} = 100$ ms). See Fig. 4 for some examples of fixation lists generated by our model.

In order to determine the optimal values for the hyper-parameters of our model, we performed several qualitative preliminary experiences, considering video streams acquired from webcams and multiple images taken from the web, testing different configurations. We considered the EX2 scheme, setting the frame-rate Δt^{-1} to 142 s^{-1}, to remain within the stability limit of the explicit scheme. Without loss of generality, we set $\gamma = 1$ in Eq. (6)—it is always possible to rescale c, λ and μ (through α_1 and α_2) accordingly. Considering the parameters that belong to the main equations of our model, i.e., Eqs. (1), (3), (4), (6), we identified in $\varpi = 1$, $\alpha_1 = 0.5$, $\alpha_2 = 20$, $\beta = 0.1\Delta t$, $\sigma = 3650$, $c = 100$, $\lambda = 10$ a configuration that performed in an appropriate manner during our initial experiences. The force acting on the FOA particle has been also rescaled by a constant multiplicative factor, i.e., $-\nabla\varphi \rightarrow -z\nabla\varphi$ and z has been set to $3{\cdot}10^6$ after the preliminary tests. Then, we analyzed the sensitivity of the model to changes in $c, \lambda, \alpha_1, \alpha_2$, that we considered to be the key hyper-parameters. Fig. 2 illustrates the outcome of scanpath prediction experiments, using three exemplar videos from the COUTROT dataset and checking the MultiMatch indices (see Section 3.3). The top row of Fig. 2 shows how doubling or halving the initial guess of $\lambda = 10$ triggers evident changes in the results, while doing the same to the guessed $c = 100$ is less critical. From the bottom row of Fig. 2 we see that the sensitivity to changes of α_1 and α_2 (halving or doubling them, almost) is pretty limited when increasing or decreasing the previously described guesses, since the range of values in the vertical axis is very small. For these reasons, in the scanpath prediction experiments we cross-validated λ in $\{0.5, 1, 10, 50, 100\}$, while we kept fixed the values of the other parameters. In the case of saliency prediction, hyper-parameters cb and σ_{blur} (see Section 3.4) are associated with sets of values, since they are subject to cross-validation jointly with λ. The validation data consist of one tenth of the CAT2000 training data (saliency), the MIT1003 (scanpath-images/saliency) datasets, and a few left-out videos from COUTROUT (scanpath-video), respectively. Table 1 summarizes the values of all the hyper-parameters used in the experiments of this paper or the sets of values from which they are selected for cross-validation.

Table 1
Parameters involved in different parts of the experimental setup and their associated values (see the main text for details). When a set of values is reported, the optimal value of the associated hyper-parameter is obtained by cross-validation, comparing values taken from the set.

Description	Hyper-parameters
Frame rate	$\Delta t^{-1} = 142$ s^{-1}
FOA trajectory Eq. (1)	dissipation $\varpi = 1$
Mass distribution Eq. (3)	details $\alpha_1 = 0.5$, motion $\alpha_2 = 20$
IOR Eq. (4)	$\beta = 0.1/\Delta t$, $\sigma = 3650$
Grav. potential Eq. (6)	$\gamma = 1$, velocity $c = 100$, drag $\lambda \in \{0.5, 1, 10, 50, 100\}$
Force multiplicative factor	$z = 3 \cdot 10^6$
FOA initialization	2.5° of visual angle from the center
Fixations - PYGAZE [59]	$r_{\max} = 25$ pixels (default value), $t_{\min} = 100$ ms (default value)
Saliency prediction	$k_{\text{size}} = (501, 501)$, $cb \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, $\sigma_{\text{blur}} \in \{10, 25, 50, 75, 100, 125, 150, 175, 200, 250, 300\}$

3.3. Metrics for scanpath evaluation

As explained in the nice reviews of [60,61], various alternatives exist to evaluate the quality of simulated list of fixations with respect to the ground truth ones, each one having its own pros and cons. In our experiments, we have considered the well-established *MultiMatch algorithm* [62,63] for gaze path comparison, exploiting the Python implementation of [64]. This is a vector-based metric that evaluates five different properties of the simulated scanpaths: *shape*, *length*, *direction*, *position* and *duration*. Scanpaths are described as the ordered sequences of saccadic vectors connecting consecutive fixations. Input sequences may be simplified by merging close fixations according to the saccades' amplitude and subsequent fixations along the same direction. This reduces the computational burden and makes irrelevant differences between scanpaths not affect the final result. On the other hand, this simplification makes it not clear how robust each measure is to variations of the input scanpaths [60]. For this reason, we have chosen to not simplify the input sequences in our experiments. Nevertheless, we have also verified that the final results were almost the same. The eventually simplified scanpaths are then temporally aligned according to their shape through the Dijkstra algorithm [65], so that aligned saccadic vectors pairs (u_i, v_j) are finally compared on the basis of the above five different features. The *shape* score computes the average difference between aligned vectors $|u_i - v_j|$, normalized by 2ξ, where ξ is the retina diagonal. The *length* score directly compares the average difference between the amplitude of these vectors $|u_i| - |v_j|$, normalized by ξ. The *direction* score considers the average angular difference between u_i and v_j and it is normalized by π. The *position* score is the average Euclidean distance between aligned fixations, normalized by ξ. Finally, the *duration* score considers the average difference in the temporal duration of aligned fixations and it is normalized by the maximum duration between the two. The resulting similarity metrics lie in the range $[0, 1]$, where 0 (1) stands for maximal dissimilarity (similarity) with respect to the given measure.

3.4. Saliency maps from continuous scanpaths

Even if the proposed model is designed to simulate human-like scanpaths, saliency maps can be still obtained as by-products following the procedure here presented. For each stimulus, 10 different scanpaths have been generated through the numerical integration of Eq. (1) and fixations have been extracted from those continuous scanpaths. For the saliency prediction task, each scanpath has been generated in 10 seconds of exploration. Moreover, we again underline that each

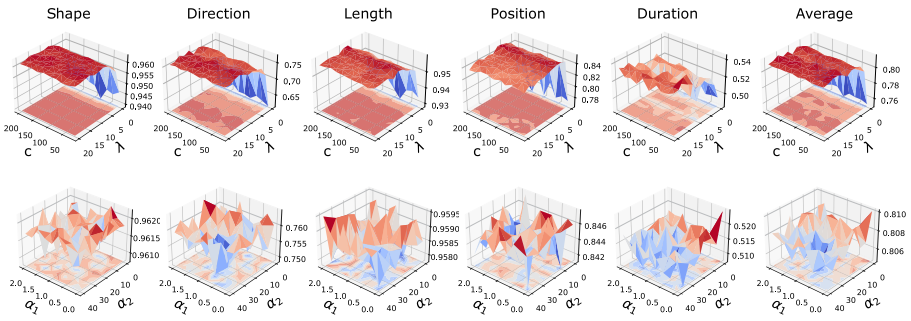


Fig. 2. Sensitivity of the proposed model to changes in the key parameters. We have considered the scanpath prediction task on three exemplar videos from the COUTROT dataset. The last column depicts the average of the five multimatch indices. In the first row (α_1, α_2) have been fixed to (0.5, 20), while in the second row (c, λ) = (100, 10).

different visual exploration corresponds to a different initialization of the FOA position and velocity. Saliency maps are then obtained by accumulating the extracted fixations in the so-called fixation maps. Next, we have applied a Gaussian smoothing (defined by the hyper-parameters σ_{blur} and the dimension of the Gaussian filter k_{size}) to each fixation map to obtain some intermediate saliency maps. Finally, in order to improve the performance of the model [21,66,67], we have combined these intermediate results (S_{map}) with a center bias prior (S_{cb}) according to the following equation

$$(1 - cb)\, S_{\text{map}} + cb\, S_{\text{cb}}\,, \qquad cb \in [0,\, 1]\ , \tag{20}$$

to get the final saliency maps.

3.5. Metrics for saliency evaluation

Different metrics exist to evaluate the generated saliency maps with respect to the ground-truth fixations, see [66] for a review. We have considered the ones of the MIT saliency benchmark [57]. Some of these metrics (AUC, NSS) directly take into account human fixations, while others (SIM, EMD, CC and KL) compare the simulated saliency maps with the ground-truth ones that are obtained applying a Gaussian blur to the corresponding human fixation maps. In detail:

- *Area Under the ROC Curve [50] (AUC)*. The saliency map is treated as a binary classifier to distinguish between pixels being fixated or not. The Receiver Operating Characteristic (ROC) curve is assessed, considering the true and false positive rates at various thresholds. The AUC metric measures the area under the ROC curve. The true positive rate is evaluated as the proportion of fixations above the specific threshold over the total number of fixations, while the evaluation of the false positive rate depends on the specific implementation. In the AUC-Judd one, it corresponds to the proportion of non-fixated pixels in the thresholded saliency map over the total number of non-fixated pixels. The AUC-Borji implementation [68] considers instead a uniform random sample of h pixels and the false positive ratio is equal to the number of those random pixels above the selected threshold over h. In particular, h is fixed to be equal to the corresponding number of fixations. Finally, in the shuffled-AUC (sAUC) version [69], false positives are sampled considering fixations from other images and the false positive ratio corresponds to the number of fixated pixels over the total number of fixations. This, in turn, mitigates the performance of all those models predicting fixations with a strong center bias.
- *Normalized Scanpath Saliency [70] (NSS)*. It measures the mean saliency value at fixated locations of the saliency map, normalized with zero mean and unit variance.
- Similarity metric (SIM). It measures the similarity between simulated and ground-truth saliency maps, viewed as histograms. Having normalized the input saliency maps, it is computed as the sum of the minimum saliency value at each pixel. The resulting score lies in (0, 1) interval, where a null score indicates no overlap between the input maps while a score = 1 means that they are the same distribution.
- Pearson's Correlation Coefficient (CC). The input saliency maps are considered as two random variables. Their linear correlation coefficient is given by their covariance over the product of their standard deviations.
- Earth Mover's Distance (EMD). Also known as Wasserstein metric, it measures the distance between two probability distributions evaluating the minimum cost that must be paid to turn one distribution into the other. Intuitively, we can consider one distribution as a pile of dirt spread in a certain region, and the other as a collection of holes. The EMD then measures the least amount of work that is needed to fill the holes of one distribution with the dirt of the other. An EMD score = 0 means that the two input distributions are the same. The EMD measure is practically computed by solving an instance of the well-known transportation problem [71].
- KL divergence (KL) It is the standard Kullback–Leibler divergence and measures the loss of information when the simulated saliency map is used to approximate the ground-true one. The KL divergence is 0 if and only if the input distributions are the same.

4. Results

4.1. Running times of local and non-local models

To highlight the benefits of the spatially local and temporal coherent computation of the potential characterizing the proposed model, we compared it with its mostly related competitor, G-Eymol [41]. Models based on a saliency maps may also be exploited to generate scanpaths through the winner-take-all mechanism [19] or probabilistic samplings. Nevertheless, since they assume a completely different computational structure, they are not included in this comparison. Moreover, note that within the saliency approach, the spatial and temporal continuity inherent to the phenomenon of overt spatial attention is lost. We considered a PyTorch-based implementation of the proposed model[8] and the original implementation of G-Eymol made available by the authors. We have selected 3 images from MIT1003 data [50] and 3 videos from COUTROT data [54] (cumulative length of 63 seconds at 25 frames per second), running the algorithms with the best parameters found in the scanpath prediction experiments (see the parameter selection procedure illustrated in Section 3.2), both on an Intel i9 CPU 2.3 GHz with 8 cores and on a GPU NVIDIA GeForce 1080 Ti with 3584 CUDA cores. While the former has limited parallel computational capabilities, the latter allows us to exploit a stronger level of parallelism in performing operations with tensors. G-Eymol processed visual streams at 25 frames per second, while our local model was provided with streams at 142

[8] Freely downloadable at https://gitlab.com/mela64/localfoa

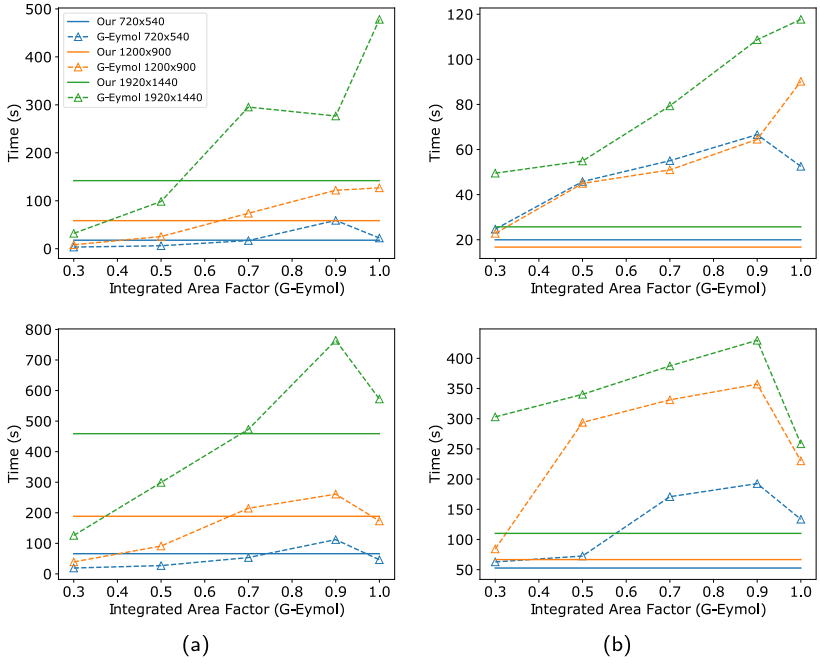


Fig. 3. Time (in seconds – average) to process image data (top row) or video data (bottom row) for different input resolutions. (a) CPU; (b) GPU. The x-axis is the factor that down-scales the size of the area on which the integral of Eq. (2) is computed – it only affects G-Eymol. As far as our model is concerned, no sub-selection of the visual scene is done at all. In fact, in order to consistently propagate visual information along the time dimension without constraining the focus of attention to a given area, the value of the potential is evaluated, for each time step, on each pixel of the retina. The GPU case also includes the time it takes to transfer data to the GPU memory.

frames per second to guarantee numerical stability (the original video frames are artificially repeated to match such a rate). In the case of images, streams consist of the same images repeated for 10 s at 25 frames per second, while videos are processed up to their natural end.

Fig. 3(a) and 3(b) report the average computational times in the case of CPU and GPU, respectively (top: images; bottom: videos), for high-resolution visual streams, in which the benefits of parallel computational schemes are more evident. We also considered the case in which we restrict the computation of the potential in G-Eymol to a squared area of edge $q \cdot w$ around the coordinates of the focus of attention, being w the width of each image/frame and q the value reported in the horizontal axis of Fig. 3 (when $q = 1$ the full image/frame is considered). Notice that the GPU case also includes the time it takes to transfer data to the GPU memory, and in both CPU and GPU cases we precomputed the optical flow. Results show that the proposed local scheme strongly benefits from parallel hardware, while G-Eymol, due to its non-local nature, struggles with larger resolutions. In G-Eymol, when $q = 1$, no data sub-selections are performed, making it sometimes faster than when using smaller q. In the case of GPU both the algorithms are faster, but it is evident that our model performs in a similar manner when increasing the resolution, confirming the crucial role of the local computational scheme. Reducing the area in which gravitational potential is integrated allows G-Eymol to gain speed, but it will not be able to consider details out of the restricted area, thus limiting the capabilities of the attention mechanism. Differently, the proposed model propagates the local information over the whole frame without any restrictions, but it requires processing streams with a larger temporal resolution to make the propagation effective. However, the total computational time is still smaller than the one of G-Eymol. Going beyond the parallelism introduced by the tensor-related operations in PyTorch, the parallelism of the pixel-wise local computational units could have been even more evident when using, for example, specific implementations for custom hardware.

4.2. Scanpath prediction

The scanpath prediction task consists of predicting the sequence of fixations that a human subject performs in free-viewing conditions when presented with a certain stimulus. Not only the spatial location of fixations is evaluated, but also their temporal order so that the FOA temporal dynamic plays a crucial role in this task. Considering static inputs, the proposed model is compared with other three unsupervised models designed to predict scanpaths (G-Eymol [41], Eymol [40] and CLE [30]). We also considered the classical Itti's model [20], two different baselines (Random and Center) and two state-of-the-art supervised saliency maps models (SAM-Resnet [23] and Deep Gaze II [72]), as a reference. Whenever possible, we used the authors' original implementation to generate fixation sequences. The Random and Center baselines sample fixations from the entire visual scene either uniformly or considering a Gaussian prior [50], respectively. Parameters of G-Eymol and Eymol have been set to their optimal values as reported in the corresponding original papers, according to various scanpath metrics. SAM-ResNet and Deep Gaze II, being them supervised models, were both trained on the SALICON dataset [73], and Deep Gaze II was additionally fine-tuned on MIT1003 [50]. Itti's saliency maps have been generated according to the original procedure presented in [20] and these have also been the starting point to generate CLE's fixations, using default parameters. Considering the strictly saliency-oriented models SAM-Resnet and Itti, the sequences of fixations have been generated applying the winner-take-all algorithm instead [19]. For Deep Gaze II, the same procedure has been adopted, with the only difference that fixations have been sampled from the resulting saliency map, considered as a probability density distribution. It is worth noting that, for each model, human sequences of fixations have been compared with simulated list of fixations of the same length.

We chose the Multimatch metrics [62,63] for the quantitative evaluation of the proposed model since, with respect to other evaluation indices and as explained in Section 3.3, they provide information about different specific attributes of the simulated scanpaths (shape, direction, length, position and duration). Results are reported in Table 2 (top portion), where the Multimatch-duration metric is shown only for those models that predict the temporal length of each fixation. As suggested by [74], results are shown in terms of *mean* and *best* prediction scores. In the case of *mean*, scores are averaged over all subjects in the dataset. In the case of *best* we consider, for each

Table 2
Scanpath prediction scores on the collection of image datasets MIT1003 [50], **TORONTO** [51], **KOOTSTRA** [52], **SIENA12** [53], **and on videos of the COUTROT dataset** [54]. Human sequences of fixations have been compared with simulated list of fixations of the same length. In the case of videos, fixations lists have been extracted from scanpaths whose length was equal to the entire length of the input videos (in Eymol the number of extracted fixations was always almost null, so lists of fixations have been defined by sampling the gaze position every half second). The "*mean* score" is the average of the various results considering, for each stimulus, each possible combination of simulated and human scanpaths. On the other hand, the "*best* score" considers, for each stimulus and simulated scanpaths, only the subject that best matches with that simulated scanpath, according to the metric under analysis. The standard deviation of these scores is also reported in brackets. The SPV columns specifies supervised models—best in bold among unsupervised ones.

	Model	SPV	MULTIMATCH									
			Shape		Direction		Length		Position		Duration	
			mean	*best*	*mean*	*best*	*mean*	*best*	*mean*	*best*	*mean*	*best*
COLLECTION OF IMAGES	OUR ($\lambda = 100$)	No	**0.93 (0.03)**	**0.97 (0.01)**	0.59 (0.16)	0.82 (0.08)	**0.90 (0.05)**	**0.97 (0.02)**	0.80 (0.08)	**0.90 (0.04)**	0.53 (0.13)	0.75 (0.08)
	G-Eymol [41]	No	0.90 (0.04)	0.94 (0.02)	**0.68 (0.15)**	**0.86 (0.06)**	0.89 (0.06)	0.95 (0.03)	**0.82 (0.06)**	**0.90 (0.04)**	**0.55 (0.14)**	**0.76 (0.08)**
	Eymol [40]	No	0.88 (0.07)	0.93 (0.05)	0.62 (0.16)	0.83 (0.07)	0.84 (0.12)	0.92 (0.08)	0.67 (0.11)	0.77 (0.10)	0.45 (0.20)	0.67 (0.19)
	CLE [30]	No	0.92 (0.04)	0.96 (0.02)	0.64 (0.15)	0.84 (0.06)	**0.90 (0.05)**	0.96 (0.02)	0.79 (0.09)	0.89 (0.05)	–	–
	Itti [20]	No	0.90 (0.06)	0.95 (0.04)	0.65 (0.15)	0.85 (0.07)	0.87 (0.09)	0.95 (0.05)	0.73 (0.11)	0.82 (0.09)	–	–
	Center	No	0.86 (0.04)	0.90 (0.03)	0.65 (0.15)	0.84 (0.06)	0.82 (0.08)	0.91 (0.05)	0.75 (0.06)	0.83 (0.04)	–	–
	Random	No	0.84 (0.05)	0.89 (0.03)	0.65 (0.15)	0.84 (0.06)	0.79 (0.09)	0.89 (0.06)	0.72 (0.07)	0.81 (0.05)	–	–
	SAM-ResNet [23]	Yes	0.94 (0.03)	0.98 (0.01)	0.64 (0.17)	0.85 (0.07)	0.90 (0.06)	0.97 (0.02)	0.84 (0.09)	0.94 (0.04)	–	–
	Deep Gaze II [72]	Yes	0.90 (0.04)	0.94 (0.02)	0.67 (0.15)	0.86 (0.05)	0.88 (0.06)	0.95 (0.03)	0.81 (0.07)	0.89 (0.04)	–	–
COUTROT VIDEOS	OUR ($\lambda = 10$)	No	**0.94 (0.02)**	**0.97 (0.01)**	**0.73 (0.07)**	**0.86 (0.04)**	**0.93 (0.04)**	**0.97 (0.01)**	0.83 (0.05)	0.90 (0.03)	**0.52 (0.12)**	**0.72 (0.05)**
	G-Eymol [41]	No	**0.94 (0.02)**	**0.97 (0.01)**	0.71 (0.10)	0.85 (0.05)	**0.93 (0.04)**	**0.97 (0.01)**	0.83 (0.05)	**0.91 (0.04)**	0.51 (0.12)	**0.72 (0.06)**
	Eymol [40]	No	**0.94 (0.01)**	0.96 (0.01)	0.72 (0.06)	0.83 (0.03)	**0.93 (0.02)**	0.96 (0.01)	**0.84 (0.05)**	0.90 (0.03)	–	–
	Center	No	0.87 (0.02)	0.91 (0.02)	0.72 (0.06)	0.84 (0.03)	0.84 (0.05)	0.91 (0.02)	0.75 (0.03)	0.82 (0.03)	–	–
	Random	No	0.86 (0.03)	0.90 (0.02)	0.71 (0.06)	0.83 (0.03)	0.81 (0.06)	0.89 (0.03)	0.73 (0.04)	0.80 (0.03)	–	–

simulated scanpath, only the human scanpath that best matches it. At a first glance, the supervised SAM-ResNet [23] outperforms all the others, due to the possibility of learning semantic characteristics of the scenes in a data-driven approach, which highly correlates with human visual attention [75,76]. However, the proposed model is one of the best among unsupervised models, with performance comparable with the other supervised saliency model we have tested, Deep Gaze II [72].

As far as videos are concerned, saliency models are not included in the comparison since the procedure to extract fixations from saliency maps is defined for static images only. For the same reason, the CLE model is no longer considered either. Concerning G-Eymol, Eymol and the model here proposed scanpath simulations last the total length of the corresponding input video.[9] The random and center models are still evaluated comparing list of fixations of the same length instead. Results are summarized in Table 2 (bottom part). We can observe how the proposed model reaches, with a fully spatio-temporal local implementation, state-of-the-art results equaling (or in some cases overcoming) the performance of G-Eymol [41].

4.3. Saliency prediction

The saliency prediction task consists of generating saliency maps to predict the probability of each pixel to be attended by a human subject during free-viewing [77]. When analyzing still images, the temporal dynamic of the visual exploration is no longer considered, and only the spatial location of the various fixations is meaningful. For this reason, as already remarked, the proposed model is not designed to directly compute saliency maps, even if they can be obtained as by-product, that is what we experimentally evaluated. As already described in Section 3.4, the procedure we have exploited to obtain saliency maps from continuous scanpaths essentially consists of collecting fixations from different scanpaths and applying a Gaussian smoothing with an additional center bias prior. Considering one tenth of the CAT2000 [55] training and the MIT1003 [50] datasets, we have performed a grid search on the $(\lambda, \sigma_{\text{blur}}, cb)$ parameters. The dimension of the Gaussian filter has been set to $(501, 501)$. We found that $\left(\lambda = 0.5,\ \sigma_{\text{blur}} = 100,\ cb = 0.2\right)$ maximizes the saliency performance on both the training datasets with respect to the AUC-Judd metric. Qualitative results are shown in Fig. 4 together with the associated ground-truth maps.

We compared our model with competitors already considered in the scanpath prediction task[10] and some others, that are the bottom-up unsupervised saliency models AIM [51] and GBVS [78] and the data-driven ones DeepFix [79] and MSI-Net [80]. The results for all these competitors are computed and provided by the MIT saliency benchmark [57]. Results on the CAT2000 and MIT300 [56] test datasets are summarized in Table 3 (top and bottom part, respectively). We notice that supervised models maintain the state of the art in the estimation of saliency. This is again due to the possibility of learning semantic properties of the input image in a data driven-way. However,

[9] In the case of Eymol, since the number of extracted fixations was always almost null, lists of fixations have been defined by sampling the gaze position every half second and have been compared with the ground true ones characterized in the same way.

[10] The CLE [30] model is not included and the results of Deep Gaze II [72] are available only for the MIT300 dataset.

Fig. 4. Some input stimuli along with the fixations coming from one exemplar scanpath (first column). The first (last) fixation is represented by the darkest (brightest) dot. Ground-truth saliency maps are on the second column, while the simulated ones ($(\lambda = 0.5, \sigma_{\text{blur}} = 100, cb = 0.2)$) are on the right. Images come from the MIT1003 dataset.

Table 3
Saliency prediction scores, CAT2000 dataset [55] **and MIT300 dataset** [56]. The results for other competitors may be found at http://saliency.mit.edu/results_cat2000.html (CAT2000) and http://saliency.mit.edu/results_mit300.html (MIT300). In our method, the chosen hyper-parameters are $\lambda = 0.5$, $\sigma_{\text{blur}} = 100$ and $cb = 0.2$. Best in bold. ↓ (↑) means that lower (higher) results are better. The SPV column specifies supervised models.

	Model	SPV	AUC-Judd(↑)	SIM(↑)	EMD(↓)	AUC-Borji(↑)	sAUC(↑)	CC(↑)	NSS(↑)	KL(↓)
CAT2000	OUR	No	0.82	0.51	3.09	0.81	0.52	0.58	1.39	0.79
	G-Eymol [41]	No	0.81	0.50	2.61	0.65	0.53	0.54	1.38	3.65
	Eymol [40]	No	0.83	0.61	1.91	0.76	0.51	0.72	1.78	1.67
	GBVS [78]	No	0.80	0.51	2.99	0.79	0.58	0.50	1.23	0.80
	AIM [51]	No	0.76	0.44	3.69	0.75	**0.60**	0.36	0.89	1.13
	Itti [20]	No	0.56	0.34	4.66	0.53	0.52	0.09	0.25	6.71
	CenterBias	No	0.83	0.42	4.31	0.81	0.50	0.46	1.06	1.13
	Random	No	0.50	0.32	5.30	0.50	0.50	0.00	0.00	2.00
	SAM-ResNet [23]	Yes	**0.88**	**0.77**	**1.04**	0.80	0.58	**0.89**	**2.38**	0.56
	DGII [72]	Yes	–	–	–	–	–	–	–	–
	DeepFix [79]	Yes	0.87	0.74	1.15	0.81	0.58	0.87	2.28	0.37
	MSI-Net [80]	Yes	**0.88**	0.75	1.07	**0.82**	0.59	0.87	2.30	**0.36**
MIT300	OUR	No	0.78	0.46	3.88	0.77	0.54	0.44	1.07	1.01
	G-Eymol [41]	No	0.76	0.39	4.45	0.60	0.54	0.35	0.91	5.61
	Eymol [40]	No	0.77	0.46	3.64	0.72	0.51	0.43	1.06	1.53
	GBVS [78]	No	0.81	0.48	3.51	0.80	0.63	0.48	1.24	0.87
	AIM [51]	No	0.77	0.40	4.73	0.75	0.66	0.31	0.79	1.18
	Itti [20]	No	0.60	0.20	5.17	0.54	0.53	0.14	0.43	2.30
	CenterBias	No	0.78	0.45	3.72	0.77	0.51	0.38	0.92	1.24
	Random	No	0.50	0.33	6.35	0.50	0.50	0.00	0.00	2.09
	SAM-ResNet [23]	Yes	0.87	**0.68**	2.15	0.78	0.70	0.78	**2.34**	1.27
	DGII [72]	Yes	**0.88**	0.46	3.98	**0.86**	**0.72**	0.52	1.29	0.96
	DeepFix [79]	Yes	0.87	0.67	2.04	0.80	0.71	0.78	2.26	**0.63**
	MSI-Net [80]	Yes	0.87	**0.68**	**1.99**	0.82	**0.72**	**0.79**	2.27	0.66

the proposed model competes very well with them and with the other unsupervised models, which is a remarkable result given that what we propose is not designed to directly estimate saliency maps. Moreover, it is interesting to see that the results we obtained are in-line or even better than the directly related non-local model G-Eymol [41]. This experience confirms the capability of the model to propagate the information through the whole retina, despite the local nature of the computations.

5. Discussion and conclusions

This paper presented a novel computational model for emulating visual overt attention. Attention is a crucial component of the human visual system and has been identified as the fundamental means for grouping low-level features into coherent object representations [4]. We contend that replicating this mechanism in the artificial realm is of crucial importance, as also evidenced by recent studies that have

integrated visual attention into neural architectures (foveated convolutional layers) to foster the development of object-centered visual features [81].

The existing research in this area typically employs saliency maps to describe attention as a collective phenomenon characterized by collections of unordered fixations, thus ignoring its intrinsic dynamics. Saliency maps can be learnt in a data-driven fashion yielding impressive results, and can also be extended to dynamical visual scenes [35, 82]. Attentional shifts are then obtained as a by-product through a winner-take-all mechanism [19] or a probabilistic sampling, thus not preserving the spatial and temporal continuity that is inherent to the phenomenon. Conversely, as in the gravitational model of [41], we proposed to directly model the dynamics of the focus of attention that is determined by low-level features characterizing the visual scene and without the need for any prior training on large collections of fixations. The resulting framework naturally applies to both images and videos and potentially describes other types of eye movements such as saccades and smooth pursuits.

Moreover, we grounded the model on spatially local and temporal coherent computations, which we argue is of the utmost importance. According to the framework proposed by David Marr in the early 1980s [83], there are three levels of understanding for information processing systems, such as the human brain. The most abstract level is the computational level, which involves specifying the goal that the system must achieve. In our case, as already remarked, this goal is the emulation of visual spatial attention. The second level is the algorithmic level, in which we explain how the computational problem is solved. This includes defining the representation of inputs and outputs, as well as the algorithm for transforming inputs into outputs. This is the level we considered in Section 2 of this manuscript. The final level is the implementation level, which involves understanding how the task is actually performed in a biological or computational system through the interaction of basic elements. Since any biological process must rely on spatially local interactions and evolves continuously in time, we believe that developing algorithms that adhere to these principles is crucial, as this may also provide insights into the biological mechanisms behind the focus of attention.

The implementation of a spatially local and temporally coherent computation entails some tradeoffs. With respect to the baseline competitor G-Eymol [41], the devised explicit algorithms for evaluating the potential come with intrinsic stability limits that we expressed in terms of the temporal resolution $(\Delta t)^{-1}$ of the incoming visual stream. Out of these bounds, the dynamics of the potential and the corresponding scanpaths are meaningless (see also Appendix B). In any case, as shown in Section 4.1, despite requiring higher frames per second to remain within the stability limits with respect to the G-Eymol case, the overall computational time needed to process an incoming stream of the same duration is shorter due to the parallel capabilities of our spatially local and temporal coherent implementation. We also underline that the tuning of other parameters, such as the drag coefficient λ or the multiplicative factor z rescaling the force acting on the focus, is also required to obtain scanpaths that match human ones (see Section 3.2). We have evaluated the performance of the proposed model on both images and videos and have found that it reaches very competitive results in the latter case. However, when considering static images, in which saliency models can be applied, SAM-ResNet [23] outperforms our model. This is due to the fact that our model generates scanpaths by only considering the gradient of the brightness of the input stimulus, whereas SAM-ResNet learns saliency maps in a data-driven way based on a large number of labeled samples. In this respect, to enhance the quality of the generated scanpaths, it may be beneficial to incorporate top-down information into the model by introducing appropriate density masses, following a similar approach to the one analyzed in [81] in the case of G-Eymol. In this context, other than free-viewing conditions, we could also consider goal-directed settings, in which test subjects are asked to search for a specific visual target.

In addition to the aforementioned research line, several other directions can be pursued to extend the current understanding of visual attention. One such direction is the incorporation of multi-modal sensory information that influences visual attention, in addition to visual stimuli. Recent works, such as [84–86] have extended saliency map estimation by considering the specific case of audio information. A similar approach to [84,85] can then be adopted to explore the interconnection between visual and auditory modes, by first spatially localizing sound sources and then introducing additional sound density masses in Eq. (3). Of course, this approach is viable under the assumption that the sound is generated from elements in the visual scene under observation. Another research direction is the consideration of visual attention in augmented [87] and virtual [88–91] reality settings. For augmented reality, modeling visual attention can assist in adjusting virtual contents to match human expectations. Also in this case, we may introduce virtual density masses and fine-tune the weighting parameters of Eq. (3) to best replicate human scanpaths in these augmented settings. Modeling visual attention is essential in virtual reality, for optimizing the computational resources required to render high-resolution 360° input images or videos. In this case, other than ocular movements, to which our model can be directly applied, there is also the need to model head movements. Similarly to what has been done in this paper, we could think of extending the recent work of [91] that is also based on a classical gravitational framework (similar to the one of G-Eymol) by considering its hyperbolic regularization. In future work we also plan to exploit the proposed model as one of the main components of an artificial agent that learns visual features over time in a lifelong manner.

CRediT authorship contribution statement

Lapo Faggi: Conceptualization, Methodology, Investigation, Software, Formal analysis, Writing – original draft. **Alessandro Betti:** Conceptualization, Formal analysis, Writing – review & editing. **Dario Zanca:** Software, Conceptualization, Writing – review & editing. **Stefano Melacci:** Software, Methodology, Writing – review & editing, Supervision. **Marco Gori:** Conceptualization, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

An implementation of the proposed model is freely downloadable at https://gitlab.com/mela64/localfoa.

Appendix A. Regularizations of the Poisson equation

In this appendix, we will clarify in which sense Eq. (5) can be considered as a regularization of the Poisson equation. In particular, for the sake of simplicity of the derivations, in the following we will consider the equations

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

Problem H is a Cauchy problem for the heat equation with source $\mu(x,t)$, whereas problem W is a Cauchy problem for a wave equation. The term c in H represents the diffusivity constant, whereas the constant c in problem W can be regarded as the wave propagation velocity. The reason why we can consider problem H and W as *temporal regularizations* of Poisson equation with source μ is due to the following fundamental result.

Theorem 1. *Let φ^0 be the solution of $-\nabla^2\varphi^0 = \mu$. Then, the gradients $\nabla\varphi_H$ and $\nabla\varphi_W$ of the solutions φ_H and φ_W to problems H and W in Eq. (A.1) (at least pointwise) converge to $\nabla\varphi^0$ as $c \to +\infty$.*

Notice that the convergence result expressed by Theorem 1 is given on the gradients of the potentials and not on the potentials themselves. The interpretation of this result is quite straightforward. For problem H it means that the solution of the heat equation in a substances with high diffusivity c, instantly converges to its stationary value which is given by Poisson equation. For problem W, Theorem 1 turns out to be the two dimensional analogue of the infinite-speed-of-light limit in electrodynamics and in particular it expresses the fact that the retarded potential (see [92]), which in three spatial dimensions are the solutions of problem W, converges to the electrostatic potential as the speed of propagation of the wave goes to infinity ($c \to +\infty$).[11] Although both temporal regularization H and W achieve the goal of transforming the Poisson equation into an initial value problem in time from which all subsequent states can be evolved from, the different nature of the two PDE determines, for finite c, qualitative differences in the FOA trajectories.

In what follows we will indicate with $B_r(x)$ the ball of radius r and center x, with $S_r(x)$ the sphere with the same radius and the same center. We will furthermore indicate with $d\mathcal{H}^{n-1}$ the $n-1$ dimensional Hausdorff measure in $\mathbb{R}^n$ (for surface integrals). Let Ω be an open set in $\mathbb{R}^n$ and D a domain with $\overline{D} \subset \Omega$. For any regular function $f : \overline{D} \to \mathbb{R}$ we define

$$\fint_D f(x)\,dx := \frac{1}{m_n(D)}\int_D f(x)\,dx, \qquad m_n(D) := \int_D 1\,dx.$$

The next Appendix A.1 contains a brief summary of the Duhamel's Principle (see also [44]), while the proof of Theorem 1 is given in Appendix A.2.

A.1. Duhamel's principle

Consider the following abstract form of the Cauchy problem for an evolution equation of the first order:

$$P : \begin{cases} u'(t) = Au(t) + f(t) & (t > 0); \\ u(0) = \varphi, \end{cases}$$

where the spatial dependence of u is not explicitly written and A is spatial differential equation (Au indeed can be regarded again as a function of time). Now consider for every fixed $s \in [0,t]$ the additional homogeneous problem

$$P' : \begin{cases} v'(t) = Av(t) & (t > 0); \\ v(0) = f(s) + A\varphi, \end{cases}$$

and let us indicate with $v_s(t)$ the solution to this problem. Then we can prove that the function

$$u(t) = \varphi + \int_0^t v_s(t-s)\,ds,$$

solves problem P. Indeed we have $u(0) = 0$ and

$$u'(t) = v_t(0) + \int_0^t v_s'(t-s)\,ds = f(t) + Au(t).$$

In a similar way we can treat the second order problem

$$Q : \begin{cases} u''(t) = Au(t) + f(t) & (t > 0); \\ u(0) = \varphi; \\ u'(0) = \psi. \end{cases}$$

This time for any $s \in [0,t]$ we consider the solution v_s to the problem

$$Q' : \begin{cases} v''(t) = Au(t) & (t > 0); \\ v(0) = 0; \\ v'(0) = f(s) + A\varphi + sA\psi. \end{cases}$$

Then one can verify that the function

$$u(t) = \varphi + t\psi + \int_0^t v_s(t-s)\,ds$$

solves problem Q.

A.2. Proof of Theorem 1

Heat equation. Consider the solution to the following problem

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

We start by studying the problem

$$\begin{cases} u_t = c\nabla^2 u & \text{in} \quad \mathbb{R}^2 \times (0,+\infty); \\ u(x,0) = u_0(x), & \text{in} \quad \mathbb{R}^2 \times \{0\}. \end{cases} \tag{A.3}$$

In this case the constant c can be absorbed entirely by a time rescaling $t \to ct$. So that $u(x,t) = v(x,ct)$ where v solves

$$\begin{cases} v_t = \nabla^2 v & \text{in} \quad \mathbb{R}^2 \times (0,+\infty); \\ v(x,0) = u_0(x), & \text{in} \quad \mathbb{R}^2 \times \{0\}. \end{cases} \tag{A.4}$$

Let us define

$$U(x,t) := \frac{1}{4\pi t} e^{-|x|^2/4t}. \tag{A.5}$$

Therefore the solution of (A.4) is $v(x,t) = (U(\cdot,t) * u_0)(x)$, so that the solution to (A.3) is

$$u(x,t) = \frac{1}{4\pi ct}\int_{\mathbb{R}^2} e^{-|x-y|^2/4ct} u_0(y)\,dy.$$

Using again Duhamel principle we find that

$$\varphi(x,t) = \int_0^t \frac{1}{4\pi(t-s)}\int_{\mathbb{R}^2} e^{-|x-y|^2/4c(t-s)}\mu(y,s)\,dy\,ds.$$

Now let $c(t-s) = \tau$, then

$$\varphi(x,t) = \frac{1}{4\pi}\int_0^{ct}\int_{\mathbb{R}^2} \frac{e^{-|x-y|^2/4\tau}}{\tau}\mu(y,t-\tau/c)\,dy\,d\tau.$$

Then its gradient is:

$$\nabla\varphi(x,t) = -\frac{1}{8\pi}\int_0^{ct}\int_{\mathbb{R}^2} \frac{e^{-|x-y|^2/4\tau}}{\tau^2}(x-y)\mu(y,t-\tau/c)\,dy\,d\tau.$$

Taking the formal limit as $c \to +\infty$

$$\nabla\varphi(x,t) = -\frac{1}{8\pi}\int_{\mathbb{R}^2}\left(\int_0^{+\infty} \frac{e^{-|x-y|^2/4\tau}}{\tau^2}\,d\tau\right)(x-y)\mu(y,t)\,dy.$$

Because $\int_0^\infty e^{-a^2/\tau}/\tau^2\,d\tau = a^{-2}$, we have

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

which is indeed the gradient of the potential that solves Poisson equation with source μ. Notice however that performing the formal limit $c \to \infty$ directly into the expression for the potential would lead to a divergent limit since $\int_0^\infty e^{-a^2/\tau}/\tau\,d\tau$ is divergent.

Wave equation. Let us begin to analyze the solution of the non-homogeneous wave equation

$$\begin{cases} \varphi_{tt} = c^2(\nabla^2\varphi + \mu) & \text{in} \quad \mathbb{R}^2 \times (0,+\infty); \\ \varphi(x,0) = 0, \quad \varphi_t(x,0) = 0 & \text{in} \quad \mathbb{R}^2 \times \{0\}. \end{cases} \tag{A.6}$$

[11] It is worth mentioning that while this kind of regularization is well-known in three dimensions, the same properly has not been formally stated in two dimensions. A formal proof of the property in the case of two dimensions is given in the paper.

In order to find an explicit solution to this problem, as usual (see [44]), we start by considering the related problem

$$\begin{cases} u_{tt} - c^2\nabla^2 u = 0 & \text{in} \quad \mathbb{R}^2 \times (0,+\infty); \\ u(x,0) = f(x), \quad u_t(x,0) = g(x) & \text{in} \quad \mathbb{R}^2 \times \{0\}. \end{cases} \tag{A.7}$$

The solution of such problem is $u(x,t) = v(x,ct)$, where v solves

$$\begin{cases} v_{tt} - \nabla^2 v = 0 & \text{in} \quad \mathbb{R}^2 \times (0,+\infty); \\ v(x,0) = f(x), \quad v_t(x,0) = g(x)/c & \text{in} \quad \mathbb{R}^2 \times \{0\}. \end{cases} \tag{A.8}$$

Since the solution of (A.8) is given by the Poisson's formula in two dimensions

$$v(x,t) = \frac{1}{2}\fint_{B_t(x)} \frac{tf(y) + c^{-1}t^2 g(y) + t\nabla f(y)\cdot(y-x)}{(t^2 - |y-x|^2)^{1/2}}\, dy,$$

we have

$$u(x,t) = \frac{1}{2}\fint_{B_{ct}(x)} \frac{ctf(y) + ct^2 g(y) + ct\nabla f(y)\cdot(y-x)}{(c^2t^2 - |y-x|^2)^{1/2}}\, dy. \tag{A.9}$$

The solution to Eq. (A.6) can be obtained from the solution of Eq (A.8) via the Duhamel's principle (see Appendix A.1). In this case we have that

$$\varphi(x,t) = \int_0^t w_s(x,t-s)\, ds, \tag{A.10}$$

where w_s solves

$$\begin{cases} w_{tt} - c^2\nabla^2 w = 0 & \text{in} \quad \mathbb{R}^2 \times (0,+\infty); \\ w(x,0) = 0, \quad w_t(x,0) = c^2\mu(x,s) & \text{in} \quad \mathbb{R}^2 \times \{0\}. \end{cases}$$

Thus from Eq. (A.9) and (A.10) we immediately have:

$$\varphi(x,t) = \frac{1}{2\pi}\int_0^t \int_{B_{c(t-s)}(x)} \frac{c\mu(y,s)}{(c^2(t-s)^2 - |y-x|^2)^{1/2}}\, dy ds.$$

Now let us make the change of variables $c(t-s) = \tau$ in the integral over s; we thus obtain:

$$\varphi(x,t) = \frac{1}{2\pi}\int_0^{ct} \int_{B_\tau(x)} \frac{\mu(y,t-\tau/c)}{(\tau^2 - |y-x|^2)^{1/2}}\, dy d\tau. \tag{A.11}$$

Example 1. In order to understand Eq. (A.11) let us consider the case of a unit mass fixed at the origin: $\mu(x,t) = \delta_x$. In this case

$$\begin{aligned} \varphi(x,t) &= \frac{1}{2\pi}\int_0^{ct} \int_{B_\tau(x)} \frac{\delta_y}{(\tau^2 - |y-x|^2)^{1/2}}\, dy d\tau \\ &= \frac{1}{2\pi}\int_{|x|}^{ct} \frac{1}{(\tau^2 - |x|^2)^{1/2}}\, d\tau \\ &= \frac{1}{2\pi}\left[\log\left\|\tau + \sqrt{\tau^2 - |x|^2}\right\|\right]_{|x|}^{ct} \\ &= \frac{1}{2\pi}\log\left(ct + \sqrt{(ct)^2 - |x|^2}\right) + \frac{1}{2\pi}\log\frac{1}{|x|}. \end{aligned} \tag{A.12}$$

Then

$$\begin{aligned} \varphi(x,t) =& \frac{1}{2\pi}\log(ct) + \frac{1}{2\pi}\log\frac{1}{|x|} \\ &+ \frac{1}{2\pi}\log\left(1 + \sqrt{1 - (|x|/ct)^2}\right). \end{aligned} \tag{A.13}$$

Notice that in the last formula as $c \to \infty$ we have a divergent part plus a finite part which is indeed our initial guess for this limit; moreover the divergent part has a vanishing spatial gradient meaning that it does not effect the force which is entirely given by gradient of $\log(1/|x|)$ which is indeed the force that ones derive from Poisson equation on $\mathbb{R}^2$.

This example suggests to look for the convergence of $\nabla\varphi$ rather than that of φ itself that can give rise to divergences.

In general let us now come back to Eq. (A.11). This integral is performed over a cilinder in the space $y - \tau$. A little thinking shows that such integration can be rearranged as follows:

$$\varphi(x,t) = \frac{1}{2\pi}\int_{B_{ct}(x)} \left(\int_{|y-x|}^{ct} \frac{\mu(y,t-\tau/c)}{(\tau^2 - |y-x|^2)^{1/2}}\, d\tau\right) dy. \tag{A.14}$$

Or, equivalently performing the change of variables $z = y - x$

$$\varphi(x,t) = \frac{1}{2\pi}\int_{B_{ct}(0)} \left(\int_{|z|}^{ct} \frac{\mu(z+x,t-\tau/c)}{(\tau^2 - |z|^2)^{1/2}}\, d\tau\right) dz. \tag{A.15}$$

Since we are interested in the limit $c \to \infty$ we can expand $\mu(z+x,t-\tau/c)$ in powers of $1/c$ around zero:

$$\begin{aligned} \mu(z+x,t-\tau/c) =\ & \mu(z+x,t) - \mu_t(z+x,t)\tau\frac{1}{c} \\ &+ \mu_{tt}(z+x,t)\tau^2\frac{1}{c^2} + o(1/c^2). \end{aligned} \tag{A.16}$$

At order zero in $1/c$ we have

$$\begin{aligned} \varphi(x,t) &= \frac{1}{2\pi}\int_{B_{ct}(0)} \left(\int_{|z|}^{ct} \frac{1}{(\tau^2 - |z|^2)^{1/2}}\, d\tau\right)\mu(z+x,t)\, dz \\ &= \frac{1}{2\pi}\int_{B_{ct}(0)} \mu(z+x,t)\left(\log(ct) + \log\frac{1}{|z|}\right. \\ &\quad \left. + \log\left(1 + \sqrt{1 - |z^2|/(ct)^2}\right)\right) dz. \end{aligned} \tag{A.17}$$

The gradient of such expression is

$$\begin{aligned} \nabla\varphi(x,t) =& \frac{1}{2\pi}\int_{B_{ct}(0)} \nabla\mu(z+x,t)\left(\log(ct) + \log\frac{1}{|z|}\right. \\ &\left. + \log\left(1 + \sqrt{1 - |z^2|/(ct)^2}\right)\right) dz. \end{aligned} \tag{A.18}$$

Now we can use the following version of the divergence theorem

$$\int_\Omega f\nabla g\, dx = \int_{\partial\Omega} fg\nu\, d\mathcal{H}^{n-1} - \int_\Omega \nabla f\, g\, dx, \tag{A.19}$$

where ν is the normal to $\partial\Omega$. In order to prove this start from the divergence theorem for vector fields:

$$\int_\Omega \operatorname{div} v\, dx = \int_{\partial\Omega} v\cdot\nu\, d\mathcal{H}^{n-1},$$

then choose $v_k = fg\delta_{ki}$, therefore

$$\begin{aligned} \int_\Omega f\nabla_i g\, dx + \int_\Omega \nabla_i f\, g\, dx &= \int_\Omega \operatorname{div} v\, dx \\ &= \int_{\partial\Omega} fg\delta_{ki}\nu_k\, d\mathcal{H}^{n-1} \\ &= \int_{\partial\Omega} fg\nu_i\, d\mathcal{H}^{n-1}, \end{aligned}$$

which gives the wanted formula. If we apply such expression to Eq. (A.18) we get

$$\begin{aligned} \nabla\varphi(x,t) =& \frac{1}{2\pi}\log(1)\int_{S_{ct}(0)} \mu(z+x,t)\nu(z)\, d\mathcal{H}^1(z) \\ &- \frac{1}{2\pi}\int_{B_{ct}(0)} \mu(z+x,t)\left(\nabla\log\frac{1}{|z|}\right. \\ &\left. + \nabla\log\left(1 + \sqrt{1 - |z|^2/(ct)^2}\right)\right) dz. \end{aligned} \tag{A.20}$$

Notice that the surface term must indeed be zero since it comes from an integral $\int_{|z|}^{ct}$ so when $|z| = ct$ the whole term is vanishing.

Expanding the gradient

$$\begin{aligned} \nabla\varphi(x,t) = \frac{1}{2\pi}\int_{B_{ct}(0)} \mu(z+x,t)\left(\frac{z}{|z|^2}\right. \\ \left. + \frac{1}{c^2}\frac{z}{1 - |z|^2/(ct)^2 + \sqrt{1 - |z|^2/(ct)^2}}\right) dz. \end{aligned} \tag{A.21}$$

As we take the formal limit $c \to +\infty$, we get

$$\begin{aligned} \nabla\varphi(x,t) &= \frac{1}{2\pi}\int_{\mathbb{R}^2} \frac{z}{|z|^2}\mu(z+x,t)\, dz \\ &= -\frac{1}{2\pi}\int_{\mathbb{R}^2} \frac{x-y}{|x-y|^2}\mu(y,t)\, dy, \end{aligned} \tag{A.22}$$

which gives the expected limit.

We will now consider the generic term in (A.16) to show that indeed it was correct, in view of the limiting procedure on c, to approximate $\mu(z+x,t-\tau/c)$ up to the lowest order in $1/c$.

Let us define

$$I_n^c(\xi) := \frac{1}{c^n}\int_\xi^{ct} \frac{\tau^n}{(\tau^2-\xi^2)^{1/2}}\,d\tau.$$

with the change of variables $s=\tau-\xi$

$$I_n^c(\xi) = \frac{1}{c^n}\int_0^{ct-\xi} \frac{(s+\xi)^n}{(s^2+2s\xi)^{1/2}}\,ds.$$

And

$$I_n^{c\prime}(\xi) = -\frac{t^n}{\sqrt{(ct)^2-\xi^2}} + \frac{1}{c^n}\int_0^{ct-\xi}\left(\frac{n(s+\xi)^{n-1}}{(s^2+2s\xi)^{1/2}} - \frac{s(s+\xi)^n}{(s^2+2s\xi)^{3/2}}\right)ds$$

In the last integral let us perform the change of variable $s=cr$

$$I_n^{c\prime}(\xi) = -\frac{t^n}{\sqrt{(ct)^2-\xi^2}} + \frac{1}{c}\int_0^{t-\xi/c}\left(\frac{n(r+\xi/c)^{n-1}}{(r^2+2r\xi/c)^{1/2}} - \frac{r(r+\xi/c)^n}{(r^2+2r\xi/c)^{3/2}}\right)dr$$

As we formally let $c\to\infty$ we have that the integral converges to $(n-1)\int_0^t r^{n-2}$ so that for $n\geq 2$ it is immediate to check that

$$I_n^{c\prime}(\xi)\to 0 \quad \text{as } c\to\infty;$$

for $n=1$ this property can be checked by direct calculations; indeed

$$I_1^c(\xi) = \frac{1}{c}\sqrt{(ct)^2-\xi^2}, \qquad I_1^{c\prime}(\xi) = -\frac{1}{c}\frac{\xi}{\sqrt{(ct)^2-\xi^2}}.$$

With this notation, we have that the nth term in the expansion in powers of $1/c$ of the gradient of the potential would be

$$\begin{aligned}&\frac{1}{2\pi}\int_{B_{ct}(0)}\left(\frac{1}{c^n}\int_{|z|}^{ct}\frac{\tau^n}{(\tau^2-|z|^2)^{1/2}}\,d\tau\right)\nabla\partial_t^n\mu(z+x,t)\,dz\\ &= \frac{1}{2\pi}\int_{B_{ct}(0)} I_n^c(|z|)\nabla\partial_t^n\mu(z+x,t)\,dz.\end{aligned} \tag{A.23}$$

Since $I_n^c(ct)=0$ using (A.19) Eq. (A.23) becomes

$$-\frac{1}{2\pi}\int_{B_{ct}(0)} I_n^{c\prime}(|z|)\frac{z}{|z|}\partial_t^n\mu(z+x,t)\,dz,$$

and this quantity for $n\geq 1$ goes to 0 as $c\to\infty$.

Appendix B. Stability analysis (derivations)

This section collects all the derivations that are about the stability results presented in the main paper, considering all the proposed schemes (EX1, EX2, IMP). Before going into further details, we report the results of part of the qualitative experimental analysis we carried out to confirm the validity of the derived stability bounds. In particular, Figs. B.1 and B.2 show the dynamical evolution of the potential for the EX1 and EX2 schemes in the case of a toy example ($\Delta x=\Delta y=1$, $c=100$, $\gamma=1$ and $\lambda=10$) considering a static input. The considered test image, as already described in the main section of the paper, is surrounded by 100 additional pixels on each side to avoid spurious boundary reflections and to foster the exploration of the entire visual scene. The frame rate $(\Delta t)^{-1}$ is chosen to be above (left column of both the figures) or below (right column of the figures) the corresponding stability bounds. From the analysis of the Methods section of the main paper, in the case of the EX1 algorithm we have to require $(\Delta t)^{-1}\geq \sqrt{2}c/\sqrt{\gamma}\simeq 141.42$ to obtain stability, while in the EX2 case $(\Delta t)^{-1}\geq 4c^2\left(\lambda/2+\sqrt{\lambda^2/4+8\gamma c^2}\right)^{-1}\simeq 138.94$. Both Figs. B.1 and B.2 then confirm the accuracy of these bounds. In the IMP case, the scheme is absolutely stable and we have not observed any instability in our practical experimentation indeed.

B.1. The EX1 scheme

The first step required to evaluate the stability limit of a given multi-step scheme is to compute the associate amplification polynomial Φ. As described in the main paper, we have to substitute $\varphi_{m,j}^n = g^n e^{i(k_x\Delta x j + k_y\Delta y m)}$ in the homogeneous version of the selected finite difference scheme and cancel out the common factors $\varphi_{m,j}^n = g^n e^{i(k_x\Delta x j+k_y\Delta y m)}$. Here, the i outside the parenthesis refers to the standard imaginary unit. Moreover, following the conventions of [47], $k_x\Delta x$ and $k_y\Delta y$ lie in the range $[-\pi,\pi]$. Substituting $\varphi_{m,j}^n = g^n e^{i(k_x\Delta x j+k_y\Delta y m)}$ in the homogeneous version of the EX1 scheme (Eq. (8) of the main text) we obtain:

$$\begin{aligned}&\Big[\Big(\gamma+\frac{\lambda\Delta t}{2}\Big)g^{n+2}\\ &\quad -2\big(\gamma-2\left(C_x^2\sin^2\theta_x + C_y^2\sin^2\theta_y\right)\big)g^{n+1}\\ &\quad +\Big(\gamma-\frac{\lambda\Delta t}{2}\Big)g^n\Big]\; e^{i(k_x\Delta x j+k_y\Delta y m)}\end{aligned} \tag{B.1}$$

where $C_x=c\Delta t/\Delta x$, $C_y=c\Delta t/\Delta y$, $\theta_x=k_x\Delta x/2$ and $\theta_y=k_y\Delta y/2$. Then, the amplification polynomial Φ is equal to

$$\Phi(g) = \Big(\gamma+\frac{\lambda\Delta t}{2}\Big)g^2 - 2\big(\gamma-2(C_x^2\sin^2\theta_x + C_y^2\sin^2\theta_y)\big)g + \gamma - \frac{\lambda\Delta t}{2}, \tag{B.2}$$

that is a second order polynomial in g with real coefficients $\Phi(g)=Cg^2+Bg+A$, where:

$$\begin{cases} C=\gamma+\dfrac{\lambda\Delta t}{2}\\ B=-2\gamma+4(C_x^2\sin^2\theta_x+C_y^2\sin^2\theta_y)\\ A=\gamma-\dfrac{\lambda\Delta t}{2}\end{cases} \tag{B.3}$$

According to the restricted Von Neumann criteria, the two roots of this polynomial must lie in the unit disk for all the admissible wave vectors (k_x,k_y) in order to obtain stability. Thus, we can consider Lemma 1 of the main text to evaluate the corresponding stability limit, if any. In particular, we have

$$\begin{cases} C-A=\lambda\Delta t;\\ C+A=2\gamma,\end{cases} \tag{B.4}$$

so that we distinguish the following cases:

- For $\gamma,\lambda>0$ we are in the first case of Lemma 1, since $(C-A)(C+A)=2\gamma\lambda\Delta t$, and we have to require

  $$\left[(C_x^2\sin^2\theta_x+C_y^2\sin^2\theta_y)-\gamma\right]\left(C_x^2\sin^2\theta_x+C_y^2\sin^2\theta_y\right)\leq 0$$

  $\forall\theta_x,\theta_y$ in $[-\pi/2,\pi/2]$. In this range, $C_x^2\sin^2\theta_x+C_y^2\sin^2\theta_y$ is maximized by $C_x^2+C_y^2$, so that stability is reached for $C_x^2+C_y^2\leq\gamma$.
- For $\gamma>0$ and $\lambda=0$ (the pure wave case) we are in the second case of the above lemma, since $C+A=0$. Then, we have again

  $$\left[(C_x^2\sin^2\theta_x+C_y^2\sin^2\theta_y)-\gamma\right]\left(C_x^2\sin^2\theta_x+C_y^2\sin^2\theta_y\right)\leq 0$$

  $\forall\theta_x,\theta_y$, and this is again equivalent to $C_x^2+C_y^2\leq\gamma$.
- For $\gamma=0$ and $\lambda>0$ (the pure diffusion case) $C+A=0$ and we are in the third case. We have to require:

  $$C_x^2\sin^2\theta_x+C_y^2\sin^2\theta_y=0$$

  $\forall\theta_x,\theta_y$, so that the EX1 scheme is unstable in this case.

We can conclude that the EX1 scheme is conditionally stable. In particular, for $\gamma>0$ and $\lambda\geq 0$ the stability condition is $C_x^2+C_y^2\leq\gamma$.

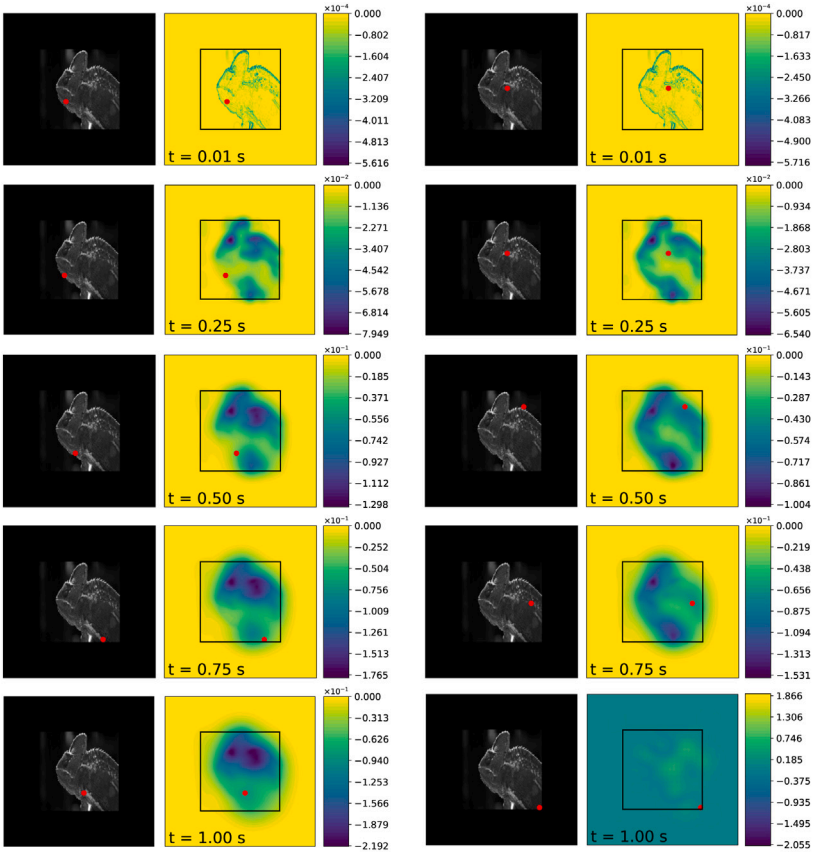


Fig. B.1. Dynamical evolution of the potential in the EX1 scheme in the case of a test image. Stability on the left ($(\Delta t)^{-1} = 142$) and instability on the right ($(\Delta t)^{-1} = 141$). With the set of parameters $\Delta x = \Delta y = 1$, $c = 100$, $\gamma = 1$ and $\lambda = 10$, the stability limit is $(\Delta t)^{-1} \simeq 141,42$. The red dots represent the position of the FOA at the selected time-step. Instability manifests itself in the positive, inconsistent and out-of-range values of the potential in the right column.

B.2. The EX2 scheme

Let us now consider the EX2 scheme, see Eq. (9) of the main text. As for the EX1 scheme, substituting $\varphi^n_{m,j} = g^n e^{i(k_x \Delta x j + k_y \Delta y m)}$ in its homogeneous version we obtain

$$\Big[(\gamma + \lambda\Delta t)g^{n+2} - 2\left(\gamma + \frac{\lambda\Delta t}{2} - 2\left(C_x^2 \sin^2\theta_x + C_y^2 \sin^2\theta_y\right)\right) g^{n+1} + \gamma g^n \Big] e^{i(k_x \Delta x j + k_y \Delta y m)} \tag{B.5}$$

so that the amplification polynomial $\Phi(g)$ is equal to

$$\Phi(g) = (\gamma + \lambda\Delta t)g^2 - 2\left(\gamma + \frac{\lambda\Delta t}{2} - 2\left(C_x^2 \sin^2\theta_x + C_y^2 \sin^2\theta_y\right)\right) g + \gamma\,. \tag{B.6}$$

Then, in this case, we obtain

$$\begin{cases} C = \gamma + \lambda\Delta t; \\ B = -2\gamma - \lambda\Delta t + 4(C_x^2 \sin^2\theta_x + C_y^2 \sin^2\theta_y); \\ A = \gamma, \end{cases} \tag{B.7}$$

so that

$$\begin{cases} C - A = \lambda\Delta t; \\ C + A = 2\gamma + \lambda\Delta t. \end{cases} \tag{B.8}$$

According to Lemma 1, we have to distinguish the following cases:

- For γ, $\lambda > 0$ or $\gamma = 0$, $\lambda > 0$ we have that $(C - A)(C + A) > 0$ and we are in the first case. Then, we have to require:

  $$\left[2\gamma + \lambda\Delta t - 2\left(C_x^2 \sin^2\theta_x + C_y^2 \sin^2\theta_y\right)\right]\left(C_x^2 \sin^2\theta_x + C_y^2 \sin^2\theta_y\right) \geq 0$$

  $\forall\,\theta_x, \theta_y$ in $[-\pi/2, \pi/2]$ so that the stability condition is $C_x^2 + C_y^2 \leq \gamma + \lambda\Delta t/2$.
- For $\gamma > 0$, $\lambda = 0$ we find that $C - A = 0$ and we are in the second case. We have now to impose the condition:

  $$\left[(C_x^2 \sin^2\theta_x + C_y^2 \sin^2\theta_y) - \gamma\right]\left(C_x^2 \sin^2\theta_x + C_y^2 \sin^2\theta_y\right) \leq 0$$

  and again this is equivalent to requiring $C_x^2 + C_y^2 \leq \gamma$.

Thus, the EX2 scheme is conditionally stable as well, and the corresponding the stability condition is $C_x^2 + C_y^2 \leq \gamma + \lambda\Delta t/2$ for $\gamma \geq 0\,, \lambda \geq 0$ with $(\gamma, \lambda) \neq (0, 0)$.

B.3. The IMP scheme

Substituting $\varphi^n_{m,j} = g^n e^{i(k_x \Delta x j + k_y \Delta y m)}$ in the IMP scheme (Eq. (10) of the main text) we obtain

$$\Big[\left(\gamma + \lambda\Delta t + 4(C_x^2 \sin^2\theta_x + C_y^2 \sin^2\theta_y)\right) g^{n+2} - 2\left(\gamma + \frac{\lambda\Delta t}{2}\right) g^{n+1} + \gamma g^n \Big] e^{i(k_x \Delta x j + k_y \Delta y i)} \tag{B.9}$$

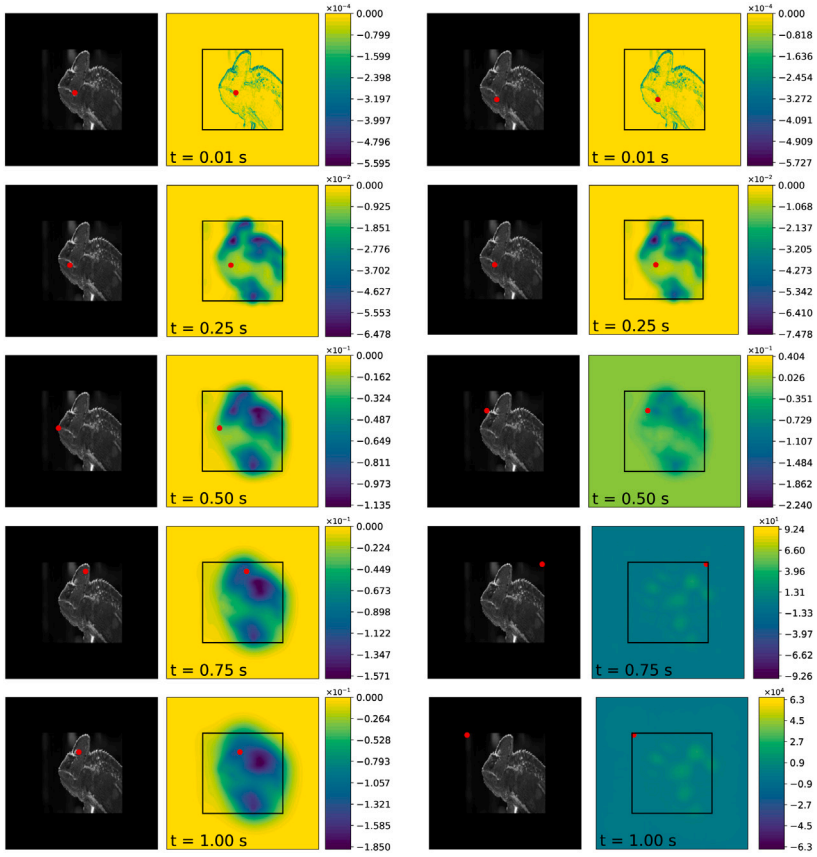


Fig. B.2. Dynamical evolution of the potential in the EX2 scheme in the case of a test image. Stability on the left ($(\Delta t)^{-1} = 140$) and instability on the right ($(\Delta t)^{-1} = 138$). With the set of parameters $\Delta x = \Delta y = 1$, $c = 100$, $\gamma = 1$ and $\lambda = 10$, the stability limit is $(\Delta t)^{-1} \simeq 138,94$. The red dots represent the position of the FOA at the selected time-step. Instability manifests itself in the positive, inconsistent and out-of-range values of the potential in the right column.

and the amplification polynomial is

$$\Phi(g) = \left[\gamma + \lambda\Delta t + 4\left(C_x^2 \sin^2\theta_x + C_y^2 \sin^2\theta_y\right)\right] g^2 - 2\left(\gamma + \frac{\lambda\Delta t}{2}\right) g + \gamma\,. \tag{B.10}$$

Then, the A, B and C real coefficients are equal to

$$\begin{cases} C = \gamma + \lambda\Delta t + 4\left(C_x^2 \sin^2\theta_x + C_y^2 \sin^2\theta_y\right) \\ B = -2\gamma - \lambda\Delta t \\ A = \gamma \end{cases} \tag{B.11}$$

and

$$\begin{cases} C - A = \lambda\Delta t + 4\left(C_x^2 \sin^2\theta_x + C_y^2 \sin^2\theta_y\right) \\ C + A = 2\gamma + \lambda\Delta t + 4\left(C_x^2 \sin^2\theta_x + C_y^2 \sin^2\theta_y\right) \end{cases} \tag{B.12}$$

In this case, for $\gamma \geq 0$, $\lambda \geq 0$ and $(\gamma, \lambda) \neq (0, 0)$, we are always in the first case of Lemma 1 and, to gain stability, we have to require that $\forall\, \theta_x, \theta_y$ in $[-\pi/2, \pi/2]$

$$\left[2\gamma + \lambda\Delta t + 2\left(C_x^2 \sin^2\theta_x + C_y^2 \sin^2\theta_y\right)\right]\left(C_x^2 \sin^2\theta_x + C_y^2 \sin^2\theta_y\right) \geq 0$$

but this condition is always satisfied so that the IMP scheme turns out to be unconditionally stable.

References

[1] Allen Allport, Visual attention, in: Foundations of Cognitive Science, The MIT Press, 1989, pp. 631–682.

[2] Kristin Koch, Judith McLean, Ronen Segev, Michael A. Freed, Michael J. Berry I.I., Vijay Balasubramanian, Peter Sterling, How much the eye tells the brain, Curr. Biol. 16 (14) (2006) 1428–1434.

[3] Sabine Kastner Ungerleider, Leslie G., Mechanisms of visual attention in the human cortex, Annu. Rev. Neurosci. 23 (1) (2000) 315–341.

[4] Pieter R. Roelfsema, Roos Houtkamp, Incremental grouping of image elements in vision, Attent. Percept. Psychophys. 73 (8) (2011) 2542–2572.

[5] Mary Hayhoe, Vision using routines: A functional account of vision, Visual Cognit. 7 (1–3) (2000) 43–64.

[6] Michael F. Land, David N. Lee, Where we look when we steer, Nature 369 (6483) (1994) 742–744.

[7] Alex D. Hwang, Hsueh-Cheng Wang, Marc Pomplun, Semantic guidance of eye movements in real-world scenes, Vis. Res. 51 (10) (2011) 1192–1205.

[8] Antonio Torralba, Aude Oliva, Monica S. Castelhano, John M. Henderson, Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search, Psychol. Rev. 113 (4) (2006) 766.

[9] Sabira K. Mannan, Christopher Kennard, Masud Husain, The role of visual salience in directing eye movements in visual object agnosia, Curr. Biol. 19 (6) (2009) R247–R248.

[10] John M. Henderson, Andrew Hollingworth, High-level scene perception, Annu. Rev. Psychol. 50 (1) (1999) 243–271.

[11] Laurent Itti, Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes, Vis. Cogn. 12 (6) (2005) 1093–1123.

[12] Fumi Katsuki, Christos Constantinidis, Bottom-up and top-down attention: different processes and overlapping neural systems, Neuroscientist 20 (5) (2014) 509–521.

[13] Laurent Itti, Automatic foveation for video compression using a neurobiological model of visual attention, IEEE Trans. Image Process. 13 (10) (2004) 1304–1318.

[14] Hadi Hadizadeh, Ivan V. Bajić, Saliency-aware video compression, IEEE Trans. Image Process. 23 (1) (2013) 19–33.

[15] Chenxi Liu, Junhua Mao, Fei Sha, Alan Yuille, Attention correctness in neural image captioning, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[16] Shi Chen, Qi Zhao, Boosted attention: Leveraging human attention for image captioning, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 68–84.

[17] Anne M. Treisman, Garry Gelade, A feature-integration theory of attention, Cogn. Psychol. 12 (1) (1980) 97–136.
[18] Anne M. Treisman, Strategies and models of selective attention, Psychol. Rev. 76 (3) (1969) 282.
[19] Christof Koch, Shimon Ullman, Shifts in selective visual attention: Towards the underlying neural circuitry, in: Matters of Intelligence, Springer, 1987, pp. 115–141.
[20] Laurent Itti, Christof Koch, Ernst Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. (1998) 1254–1259.
[21] Ali Borji, Laurent Itti, State-of-the-art in visual attention modeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2012) 185–207.
[22] A. Borji, Saliency prediction in the deep learning era: Successes and limitations, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 679–700.
[23] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, Rita Cucchiara, Predicting human eye fixations via an lstm-based saliency attentive model, IEEE Trans. Image Process. 27 (10) (2018) 5142–5154.
[24] Wenguan Wang, Jianbing Shen, Deep visual attention prediction, IEEE Trans. Image Process. 27 (5) (2017) 2368–2378.
[25] Zhaohui Che, Ali Borji, Guangtao Zhai, Xiongkuo Min, Guodong Guo, Patrick Le Callet, How is gaze influenced by image transformations? dataset and model, IEEE Trans. Image Process. 29 (2019) 2287–2300.
[26] Giuseppe Boccignone, Vittorio Cuculo, Alessandro D'Amelio, Problems with saliency maps, in: International Conference on Image Analysis and Processing, Springer, 2019, pp. 35–46.
[27] Olivier Le Meur, Zhi Liu, Saccadic model of eye movements for free-viewing condition, Vis. Res. 116 (2015) 152–164.
[28] Ming Jiang, Xavier Boix, Gemma Roig, Juan Xu, Luc Van Gool, Qi Zhao, Learning to predict sequences of human visual fixations, IEEE Trans. Neural Netw. Learn. Syst. 27 (6) (2016) 1241–1252.
[29] Deepak Khosla, Christopher K. Moore, David Huber, Suhas Chelian, Bio-inspired visual attention and object recognition, in: Intelligent Computing: Theory and Applications V, Vol. 6560, International Society for Optics and Photonics, 2007, 656003.
[30] Giuseppe Boccignone, Mario Ferraro, Modelling gaze shift as a constrained random walk, Physica A 331 (1–2) (2004) 207–218.
[31] Dietmar Heinke, Glyn W. Humphreys, Attention, spatial representation, and visual neglect: Simulating emergent attention and spatial memory in the selective attention for identification model (SAIM), Psychol. Rev. 110 (1) (2003) 29.
[32] Dietmar Heinke, Andreas Backhaus, Modelling visual search with the selective attention for identification model (VS-SAIM): A novel explanation for visual search asymmetries, Cogn. Comput. 3 (2011) 185–205.
[33] Alireza Khatoon Abadi, Keyvan Yahya, Massoud Amini, Karl Friston, Dietmar Heinke, Excitatory versus inhibitory feedback in Bayesian formulations of scene construction, J. R. Soc. Interface 16 (154) (2019) 20180344.
[34] Leo Schwinn, Doina Precup, Bjoern Eskofier, Dario Zanca, Behind the machine's gaze: Neural networks with biologically-inspired constraints exhibit human-like visual attention, Trans. Mach. Learn. Res. (ISSN: 2835-8856) (2022).
[35] Marc Assens Reina, Xavier Giro-i Nieto, Kevin McGuinness, Noel E. O'Connor, Saltinet: Scan-path prediction on 360 degree images using saliency volumes, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 2331–2338.
[36] Ryan Anthony Jalova de Belen, Tomasz Bednarz, Arcot Sowmya, ScanpathNet: A recurrent mixture density network for scanpath prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5010–5020.
[37] Farran Briggs, W. Martin Usrey, A fast, reciprocal pathway between the lateral geniculate nucleus and visual cortex in the macaque monkey, J. Neurosci. 27 (20) (2007) 5431–5436.
[38] Kerry McAlonan, James Cavanaugh, Robert H. Wurtz, Guarding the gateway to cortex with attention in visual thalamus, Nature 456 (7220) (2008) 391–394.
[39] K.-H. Schlingensiepen, F.W. Campbell, G.E. Legge, T.D. Walker, The importance of eye movements in the analysis of simple patterns, Vis. Res. 26 (7) (1986) 1111–1117.
[40] Dario Zanca, Marco Gori, Variational laws of visual attention for dynamic scenes, in: Advances in Neural Information Processing Systems, 2017, pp. 3823–3832.
[41] Dario Zanca, Stefano Melacci, Marco Gori, Gravitational laws of focus of attention, IEEE Trans. Pattern Anal. Mach. Intell. 42 (12) (2020) 2983–2995, http://dx.doi.org/10.1109/TPAMI.2019.2920636.
[42] Dario Zanca, Marco Gori, Stefano Melacci, Alessandra Rufa, Gravitational models explain shifts on human visual attention, Sci. Rep. 10 (1) (2020) 1–9.
[43] David Marr, Tomaso Poggio, From understanding computation to understanding neural circuitry, 1976.
[44] Lawrence C. Evans, Partial Differential Equations, Vol. 19, American Mathematical Soc, 2010.
[45] R.D. Richtmyer, K.W. Morton, Difference Methods for Initial Value Problems, Interscience, 1967.
[46] Gordon D. Smith, Numerical Solution of Partial Differential Equations: Finite Difference Methods, Oxford University Press, 1985.
[47] John C. Strikwerda, Finite Difference Schemes and Partial Differential Equations, SIAM, 2004.
[48] P.D. Lax, R.D. Richtmyer, Survey of the stability of linear finite difference equations, Comm. Pure Appl. Math. 9 (2) (1956) 267–293, http://dx.doi.org/10.1002/cpa.3160090206.
[49] John J.H. Miller, On the location of zeros of certain classes of polynomials with applications to numerical analysis, IMA J. Appl. Math. 8 (3) (1971) 397–406.
[50] Tilke Judd, Krista Ehinger, Frédo Durand, Antonio Torralba, Learning to predict where humans look, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE, 2009, pp. 2106–2113.
[51] Neil Bruce, John Tsotsos, Attention based on information maximization, J. Vis. 7 (9) (2007) 950.
[52] Gert Kootstra, Bart de Boer, Lambert R.B. Schomaker, Predicting eye fixations on complex visual stimuli using local symmetry, Cogn. Comput. 3 (1) (2011) 223–240.
[53] Dario Zanca, Valeria Serchi, Pietro Piu, Francesca Rosini, Alessandra Rufa, FixaTons: A collection of human fixations datasets and metrics for scanpath similarity, 2018, arXiv preprint arXiv:1802.02534.
[54] Antoine Coutrot, Nathalie Guyader, Toward the introduction of auditory information in dynamic visual attention models, in: 2013 14th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS, IEEE, 2013, pp. 1–4.
[55] Ali Borji, Laurent Itti, Cat2000: A large scale fixation dataset for boosting saliency research, 2015, arXiv preprint arXiv:1505.03581.
[56] Tilke Judd, Frédo Durand, Antonio Torralba, A Benchmark of Computational Models of Saliency to Predict Human Fixations, MIT Computer Science and Artificial Intelligence Laboratory Technical Report, 2012.
[57] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, Antonio Torralba, MIT saliency benchmark, http://saliency.mit.edu/.
[58] Matthias Kümmerer, Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, Antonio Torralba, MIT/Tübingen saliency benchmark, 2023.
[59] Edwin S. Dalmaijer, Sebastiaan Mathôt, Stefan Van der Stigchel, PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments, Behav Res Methods 46 (4) (2014) 913–921.
[60] Nicola C. Anderson, Fraser Anderson, Alan Kingstone, Walter F. Bischof, A comparison of scanpath comparison methods, Behav. Res. Methods 47 (4) (2015) 1377–1392.
[61] Ramin Fahimi, Neil D.B. Bruce, On metrics for measuring scanpath similarity, Behav. Res. Methods (2020) 1–20.
[62] Halszka Jarodzka, Kenneth Holmqvist, Marcus Nyström, A vector-based, multidimensional scanpath similarity measure, in: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, 2010, pp. 211–218.
[63] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, Kenneth Holmqvist, It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach, Behav. Res. Methods 44 (4) (2012) 1079–1100.
[64] Adina S. Wagner, Yaroslav O. Halchenko, Michael Hanke, Multimatch-gaze: The MultiMatch algorithm for gaze path comparison in python, J. Open Source Softw. 4 (40) (2019) 1525.
[65] Edsger W. Dijkstra, et al., A note on two problems in connexion with graphs, Numer. Math. 1 (1) (1959) 269–271.
[66] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, F. Durand, What do different evaluation metrics tell us about saliency models? IEEE Trans. Pattern Anal. Mach. Intell. 41 (3) (2019) 740–757, http://dx.doi.org/10.1109/TPAMI.2018.2815601.
[67] Matthias Kümmerer, Thomas S.A. Wallis, Matthias Bethge, Information-theoretic model comparison unifies saliency metrics, Proc. Natl. Acad. Sci. 112 (52) (2015) 16054–16059.
[68] A. Borji, Dicky N. Sihite, L. Itti, Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study, IEEE Trans. Image Process. 22 (2013) 55–69.
[69] Benjamin W. Tatler, Roland J. Baddeley, Iain D. Gilchrist, Visual correlates of fixation selection: Effects of scale and time, Vis. Res. 45 (5) (2005) 643–659.
[70] Robert J. Peters, Asha Iyer, Laurent Itti, Christof Koch, Components of bottom-up gaze allocation in natural images, Vis. Res. 45 (18) (2005) 2397–2416.
[71] George B. Dantzig, Application of the simplex method to a transportation problem, in: Activity Analysis and Production and Allocation, Wiley, 1951.
[72] Matthias Kummerer, Thomas S.A. Wallis, Leon A. Gatys, Matthias Bethge, Understanding low-and high-level contributions to fixation prediction, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4789–4798.
[73] Ming Jiang, Shengsheng Huang, Juanyong Duan, Qi Zhao, SALICON: Saliency in context, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 1072–1080.
[74] Dario Zanca, Stefano Melacci, Marco Gori, Toward improving the evaluation of visual attention models: a crowdsourcing approach, 2020, arXiv preprint arXiv:2002.04407.
[75] Jan Theeuwes, Paul Atchley, Arthur F. Kramer, On the time course of top-down and bottom-up control of visual attention, in: Control of cognitive processes: Attention and performance XVIII, 2000, pp. 105–124.

[76] Charles E. Connor, Howard E. Egeth, Steven Yantis, Visual attention: Bottom-up versus top-down, Curr. Biol. 14 (19) (2004) R850–R852.
[77] Ali Borji, Hamed R Tavakoli, Dicky N Sihite, Laurent Itti, Analysis of scores, datasets, and models in visual saliency prediction, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 921–928.
[78] Jonathan Harel, Christof Koch, Pietro Perona, Graph-based visual saliency, Adv. Neural Inf. Process. Syst. 19 (2006).
[79] Srinivas S.S. Kruthiventi, Kumar Ayush, R. Venkatesh Babu, Deepfix: A fully convolutional neural network for predicting human eye fixations, IEEE Trans. Image Process. 26 (9) (2017) 4446–4456.
[80] Alexander Kroner, Mario Senden, Kurt Driessens, Rainer Goebel, Contextual encoder–decoder network for visual saliency prediction, Neural Netw. 129 (2020) 261–270.
[81] Matteo Tiezzi, Simone Marullo, Alessandro Betti, Enrico Meloni, Lapo Faggi, Marco Gori, Stefano Melacci, Foveated neural computation, in: 23rd European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML–PKD, 2022.
[82] Wenguan Wang, Jianbing Shen, Jianwen Xie, Ming-Ming Cheng, Haibin Ling, Ali Borji, Revisiting video saliency prediction in the deep learning era, IEEE Trans. Pattern Anal. Mach. Intell. 43 (1) (2019) 220–237.
[83] David Marr, Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information, MIT Press, 2010.
[84] Xiongkuo Min, Guangtao Zhai, Ke Gu, Xiaokang Yang, Fixation prediction through multimodal analysis, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 13 (1) (2016) 1–23.
[85] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Xiao-Ping Zhang, Xiaokang Yang, Xinping Guan, A multimodal saliency model for videos with high audio-visual correspondence, IEEE Trans. Image Process. 29 (2020) 3805–3819.
[86] Shunyu Yao, Xiongkuo Min, Guangtao Zhai, Deep audio-visual fusion neural network for saliency estimation, in: 2021 IEEE International Conference on Image Processing, ICIP, IEEE, 2021, pp. 1604–1608.
[87] Huiyu Duan, Wei Shen, Xiongkuo Min, Danyang Tu, Jing Li, Guangtao Zhai, Saliency in augmented reality, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 6549–6558.
[88] Yucheng Zhu, Guangtao Zhai, Xiongkuo Min, The prediction of head and eye movement for 360 degree images, Signal Process., Image Commun. 69 (2018) 15–25.
[89] Yucheng Zhu, Guangtao Zhai, Xiongkuo Min, Jiantao Zhou, The prediction of saliency map for head and eye movements in 360 degree images, IEEE Trans. Multimed. 22 (9) (2019) 2331–2344.
[90] Yucheng Zhu, Guangtao Zhai, Yiwei Yang, Huiyu Duan, Xiongkuo Min, Xiaokang Yang, Viewing behavior supported visual saliency predictor for 360 degree videos, IEEE Trans. Circuits Syst. Video Technol. 32 (7) (2021) 4188–4201.
[91] Miguel Fabian Romero Rondon, Dario Zanca, Stefano Melacci, Marco Gori, Lucile Sassatelli, Hemog: A white-box model to unveil the connection between saliency information and human head motion in virtual reality, in: 2021 IEEE International Conference on Artificial Intelligence and Virtual Reality, AIVR, IEEE, 2021, pp. 10–18.
[92] John David Jackson, Classical Electrodynamics, John Wiley & Sons, 2007.

Lapo Faggi received the Ph.D. degree in Computer Science (Smart Computing) from the University of Florence, Pisa, and Siena in July 2023. In 2019 he graduated in Theoretical Physics from the University of Florence. His primary research interests revolve around theoretical aspects of machine learning and computer vision, with a particular emphasis on the incorporation of temporal dynamics into modern neural approaches.

Alessandro Betti received the M.S. Degree in Theoretical Physics in 2016 form the University of Pisa, Italy and the Ph.D. degree in Computer Science (Smart Computing) awarded jointly from the Universities of Florence, Pisa, and Siena in 2020. He is currently a postdoctoral researcher of the 3iA Côte d'Azur at Centre Inria d'Université Côte d'Azur in the Maasai team. His main research interests are in Machine Learning and specifically in the formulation of a class of learning problems that poses a natural temporal embedding using the formalism of Calculus of Variations, with applications to Online Learning, Continual Learning and Computer Vision.

Dario Zanca received the B.Sc. and M.Sc. Degrees in Mathematics from the University of Palermo, Italy. He received the Ph.D. in Smart Computing from the University of Florence, Italy, working partly at the California Institute of Technology (Caltech), Pasadena, United States. He worked as postdoc researcher at the Department of Medicine, Surgery and Neuroscience at the University of Siena, Italy. He is currently postdoc researcher at the Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany. His research interests broadly fall into the areas of computer vision and machine learning with emphasis on computational biologically-inspired self-learning systems, human attention, and eye-tracking.

Stefano Melacci received the M.S. Degree in Computer Engineering (cum laude) and the Ph.D. degree in Computer Science (Information Engineering) from the University of Siena, Italy, in 2006 and 2010, respectively. He worked as Visiting Scientist at the Computer Science and Engineering Department of the Ohio State University, Columbus, USA, and he is currently Associate Professor of the Department of Information Engineering and Mathematics, University of Siena. His research interests include machine learning and pattern recognition, mainly focused on Neural Networks and Kernel Machines, with applications to Computer Vision and Natural Language Processing tasks. He served as Associate Editor of the IEEE Transactions on Neural Network and Learning Systems.

Marco Gori received the Ph.D. degree in 1990 from Università di Bologna, Italy, working partly at the School of Computer Science (McGill University, Montreal). He is currently full professor of computer science at the University of Siena, where he is leading the Siena Artificial Intelligence Lab (SAILAB) http://sailab.diism.unisi.it. Professor Gori is primarily interests in machine learning with applications to pattern recognition, Web mining, and game playing. He has recently published the monograph "Machine Learning: A constraint-based approach," (MK, 560 pp., 2018), which contains a unified view of his approach to machine learning. His pioneering role in neural networks has been emerging especially from the recent interest in Graph Neural Networks, that he contributed to introduce in the seminal paper "Graph Neural Networks," IEEE-TNN, 2009, which received nearly 1000 citations last year. Professor Gori is a Fellow of IEEE, EurAI, IAPR, and ELLIS. Dr. Gori is currently holding an International 3IA Chair position at the Université Côte d'Azur.