Paper Title: Focus of Attention Improves Information Transfer in Visual Features (Paper ID 9698)

We thank the Reviewers for their comments and their positive feedback both on the problem we decided to tackle and on the approach we followed! We did our best to answer the provided questions and to reply to some comments.

Reviewer #1. "*Are the resulting features better able to perform some task of interest?*" We considered the problem of maximizing the Mutual Information (MI) index in order to avoid focusing on a specific task, thus having an unsupervised learning criterion. Of course, the resulting features are capturing precious information about the context around each pixel, so they might help in some downstream tasks. We tried to preliminary face a semantic labeling task, and the MI-based features were helping in case of small/sparse supervision.

"*The paper introduces fourth order dynamics, and spends considerable time simplifying this to 2nd order dynamics. It is unclear what this adds to the presentation [...]"* We aim at providing the precise theoretical groundings from which what we implemented comes from. The 2nd order dynamics of Eq. (6) draws much of its motivation from Theorem 1, which is based on a particular limit of the 4th order ODEs of Eq. (5) – stationary points of functional (4). Notice that, as far as we know, there is indeed no direct way to derive Eq. (6) from a simpler functional than (4) while retaining only the correct initial conditions (because of the inherent difficulty of deriving hyperbolic problems from variational principles). In turn, functional (4) has strong analogies with functional (2), already studied in related work, that is the one from which we start the presentation, creating a clean connection to the existing literature.

"*The experiments chose hyperparameters to maximize the mutual information extracted by each algorithm, and it is not clear from the text whether this optimization was performed on a validation set or not*"

We selected the parameters that were leading to the largest MI index at the end of the learning stage, Eq. (8), and not on the test data (thanks for this comment – we were not explicitly mentioning what data were used).

"*The evaluation videos could be better motivated. They are quite specific (exactly three videos), and the paper would be improved by discussing the significance of looping the videos. How many loops were necessary [...]"* We selected videos that naturally represent repetitive contents, so that they could be looped without introducing evident scene changes (a video call from a laptop, a camera recording cars in a parking lot, static digits). Contents and events are heterogeneous among the videos, and we believe that they are able to resemble the continual life-long flow of information hitting the human eyes. The CARPARK and CALL videos are composed of 1259 and 2386 frames, respectively – we looped them ≈ 83 and ≈ 44 times, respectively.

Reviewer #2. "*The authors need to contrast their work with that of Friston and point out the novel contributions that are being made [...]. In the experiment [...] fixation/attention locations generated by other state-of-the-art video attention algorithms*" We thank the Reviewer for providing as references the work by Frinston, that we will certainly discuss in case of acceptance! Shortly, while definitely sharing several connections, we believe that is mostly the context of this work (mutual information in video streams/visual features with deep nets paired with human-like focus of attention) that represent a novel contribution, aimed at improving time-oriented machine learning systems exploiting a model of human attention. Regarding the attention models suggested by the Reviewer, we will consider them as well, even if here we decided to focus on a FOA model that is specifically designed to generate scanpaths, hence temporal sequences of fixation points, from a SOTA ODE-based formulation and not a single saliency map (such as in Cornia et al. and other approaches in literature), making it well paired with our learning scheme. We experimented different FOA model parameters, generating different (valid) scanpaths, getting similar results. In the seminal bottom-up model by Bruce and Tsotsos attention emerges maximizing the Self-Information of each local image patch. Conversely, in our case an independent gravitational process guides the visual gaze, which we show to favour the information transfer.

Reviewer #4. "*(Thm 1) and its proof in the supplementary seem correct, but no learning algorithms are given [...]*" Instead of the classical gradient-descent-based update rule $\dot{w} = -\nabla U(w(t), t)$ our model exploits the second order equation of Eq. (6). We are integrating the equation with the Euler method (we have to specify this in the experimental section, thanks for this comment!).

"*Couldn't understand how exactly the trajectory density estimation, on which the mutual information estimation is based, is actually calculated. Solving FP like equation in high dimensions is nuitoriously difficult without some parametric assumptions.*" We compared different choices for the function g which defines the probability measure μ. As we briefly discussed towards the end of Section 3 we considered the case of uniform density over the frame, and the case in which g is a peaked function on the trajectory $t \mapsto a(t)$ which in turn is estimated using the approach proposed in [29].

"*What is actiually assumed about the unknown potential U?How is it learned from the data?*" In this paper, the potential U is equivalent to the negative mutual information (see the details in Section (2)), that depends on the output of a Deep Network composed of convolutional layers – whose filters are learned from data.

Reviewer #5. "*I could not follow the discussion in the current manuscript at all, partly due to the lack of my knowledge for the previous methods [...]*" The paper is about information transfer in vision problems, merging in a clean formulation SOTA focus of attention models and learning over time. We do agree that it covers several topics and we provided precise references to get more details about all of them. Notice that we are also sharing the code to help reproducibility.