We thank the reviewers for their feedback. Below we list the respective revisions that will be made (**R = Reviewer**).

Conceptual and intuitive introduction to transformation learning (R1,R2,R3) We will completely revise Section 2 which will now open with the following paragraph: "Next, we turn our attention to learning to detect transformations from pairs of consecutive video frames. We start with the observation that much of the change in pixel intensities in consecutive frames arises from a local translation of the image. For small translations pixel intensity change is given by a linear operator (or matrix) multiplying the vector of pixel intensity scaled by the magnitude of translation. Because, for a 2D image, multiple directions of translation are possible, there is a set of translation matrices with corresponding magnitudes. Our goal is to learn both the translation matrices from pairs of consecutive video frames and the magnitudes of translations for each pair. Such a learning problem will reduce to the one discussed in the previous section, but performed on an unusual feature – the outer product of pixel intensity and variation of pixel intensity vectors."

Results of learning in our model (R1,R3) We will revise Section 3. Specifically, we present PCA and K-means results because these are well understood computations that help with an intuitive understanding of our biologically plausible algorithm. PCA illustrates the learning of generators in the sign-unconstrained case and K-means illustrates the effect of constraining the sign of the output.

In the case of 1D translations and 2D rotations there is only one generator of transformation (for sign-unconstrained output), which explains the choice of $K = 2$ for signed-constrained output. In the case of 2D images undergoing both horizontal and vertical motions our model learns two different generators, left-right and up-down motion ($K = 4$ for sign-constrained output). Fig.1c-d-e-f show the filters learned by our model, each accounting for a motion in a cardinal direction. These generators were also reported in [17]. In addition, when presented with pairs of points in $\mathbb{R}^n$ transformed by the elements of group $SO(n)$, our model learns the various generators ($K > 2$).

Comparison of model predictions with the biological observations (R1,R2,R3) Our theory's predictions are consistent with experimental measurements of physiology and anatomy of the T4 circuit including phi and reverse phi optical illusions. The predicted output of our detectors, integrated over the visual field is consistent with experimental observations such as the increase with image contrast, the oscillations in the motion signal locked to the phase of the visual stimulus, non-monotonic dependence of output on motion velocity. Our reference to pixels in the context of fly vision is justified by the facet structure of the fly eye wherein photoreceptors respond to light intensity in a hexagonal grid of locations in the visual field.

Biological implementation of the algorithm (R1,R2) We will revise Section 4 to clarify the relevant biological mechanisms and make a stronger connection with the algorithm. In particular, it is true that backpropagating action potential briefly interrupts dendritic integration yet it is widely thought to underlie Hebbian-like learning [32].

Comparison of our model with other models (R1,R2) The main difference between our model and most published models (including the model in ref.[28]) is that the motion detector is learned from data using biologically plausible learning rules in an unsupervised setting. Thus, our model can generate somewhat different receptive fields for different natural image statistics such as that in ON and OFF pathways potentially accounting for minor differences reported between T4 and T5 circuits [33]. In addition, the model in [28] is architecturally different from ours as it is composed of a shared non-delay line flanked by two delay lines. Our model instead uses a temporal derivative in the middle pixel flanked with two non-shared non-delay lines. Whereas, after integration over the visual field, the outputs predicted by our model, HR and [28] are algebraically the same, the predicted output of a single motion detector in our model is different from both HR and [28].

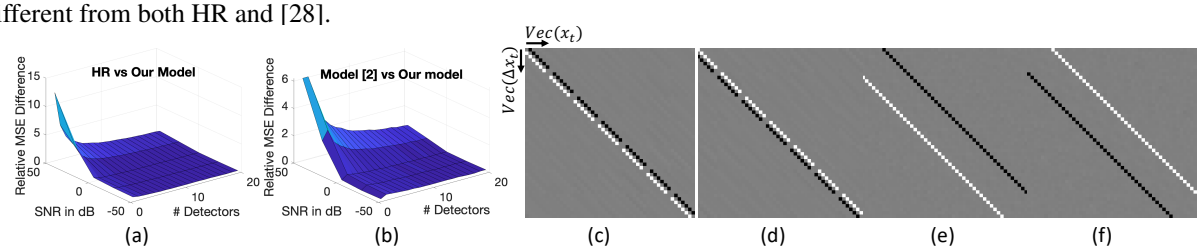

Figure 1: Robustness to noise (a) our model vs HR, (b) our model vs [28]. Learned generators on 2D images of (c-d) horizontal motions and (e-f) vertical motions.

A very recent paper [1] reported experimental measurements of direction opponency (DO) in T4 and T5 cells. They showed that the HR model cannot account for DO and proposed a biophysical model that reproduces observed DO. Our model also reproduces DO, as will be demonstrated in the revised version of our paper.

Finally, we evaluated our model against HR and [28] in terms of robustness of their output to noise. Fig.1a (resp.1b) show the relative difference in mean squared error (MSE) between our model and HR (resp.[28]), for different SNR and different number of detectors. A positive value indicates that our model is less sensitive to noise than the competition. For both low SNR (<0dB) and integration over a large number of detectors our model, HR, and [28] perform similarly. In realistic settings, however, our model is more robust to noise than the other two.

[1] Bara A. Badwan et al. Dynamic nonlinearities enable direction opponency in drosophila elementary motion detectors. *Nature neuroscience*, 22(8):1318, 2019.