Thirty-seventh Conference on Neural Information Processing Systems （NeurIPS 2023）

Modeling Human Visual Motion Processing with Trainable Motion Energy Sensing and a Self-attention Network

Zitang Sun[1] · Yen-Ju Chen [1]· Yung-Hao Yang[1] · Shin'ya Nishida[12]*

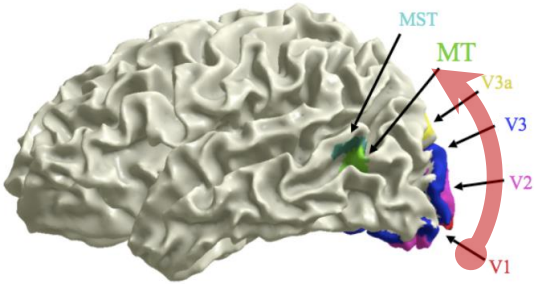
1. Cognitive Informatics Lab
Kyoto University

2. NTT Communication Science Laboratories,
Nippon Telegraph and Telephone Corporation

Problem & Motivation

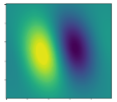
Visual Motion Processing Along the dorsal stream

- **Basic Functions:** The motion energy sensor/ spatiotemporal filters (V1) for capturing local motion.
- **More advanced functions:** MT/MST regions for motion spatial integration and segmentation.
- Existing attempts still less considered the ability to derive dense motion flows from local motion energy and thus generalize to natural videos to simulate the high-level visual function of humans.



Motivation

Using the powerful DNNs



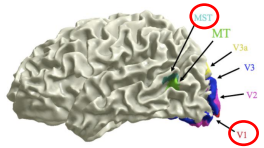
Classical energy models

Approximate and compare the motion perception functions of the human visual system

- Neurophysiologically
- psychophysically
- Engineering-wise

Problem & Motivation

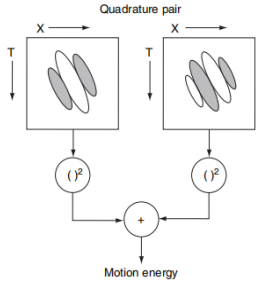
- **A two-stage process**
- **Stage I: the trainable motion energy sensor→ V1**
- **Stage II: recurrent integration based on the attention mechanism→MT**



Stage I

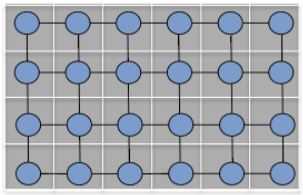
(Local motion energy capture; Simulating the function of V1):

Spatiotemporal separatable convolution

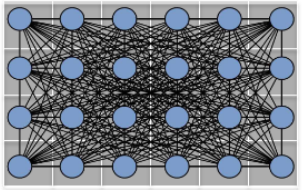


Stage II

(Global motion Integration & Segregation; Simulating the function of MT)

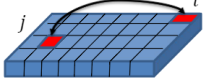


Local motion energy



Fully connected graph

♦ Choose dot-product distance to calculate relation among each nodes



$A(i, j) = A(j, i) = \theta(i) \cdot \theta^{T}(j)$

Modeling Two Stages of Motion Perception

- The first stage consists of a group of cells with **trainable** spatiotemporal frequencies tuning to capture different preferences of local motion energy.

$$x' = x\cos\theta + y\sin\theta,\ y' = -x\sin\theta + y\cos\theta$$

Spatial Filter

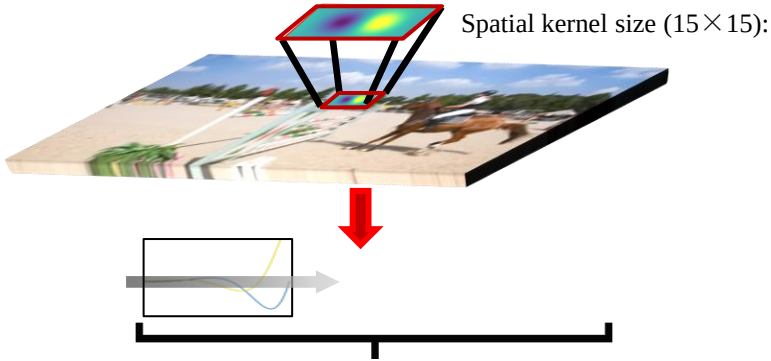
$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x^2 + \gamma^2 y^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda} + \psi\right)$$

$$H(a, \phi, \tau, \omega_t, t) = -ae^{-\tau t}\sin\left(\omega_t t + \phi\right)$$

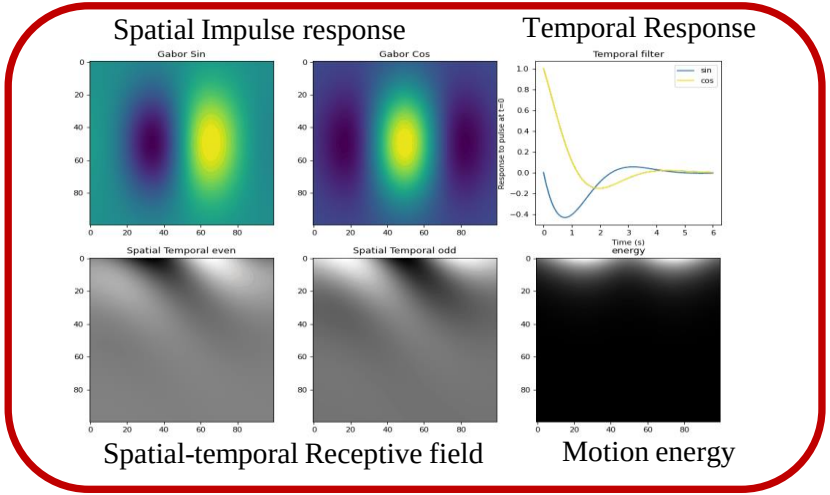
Temporal Filter

Spatiotemporal Separable Filter

$$s(x, y, t) = g(x, y; \lambda, \theta, \psi, \sigma, \gamma) \times H(a, \phi, \tau, \omega_t, t)$$



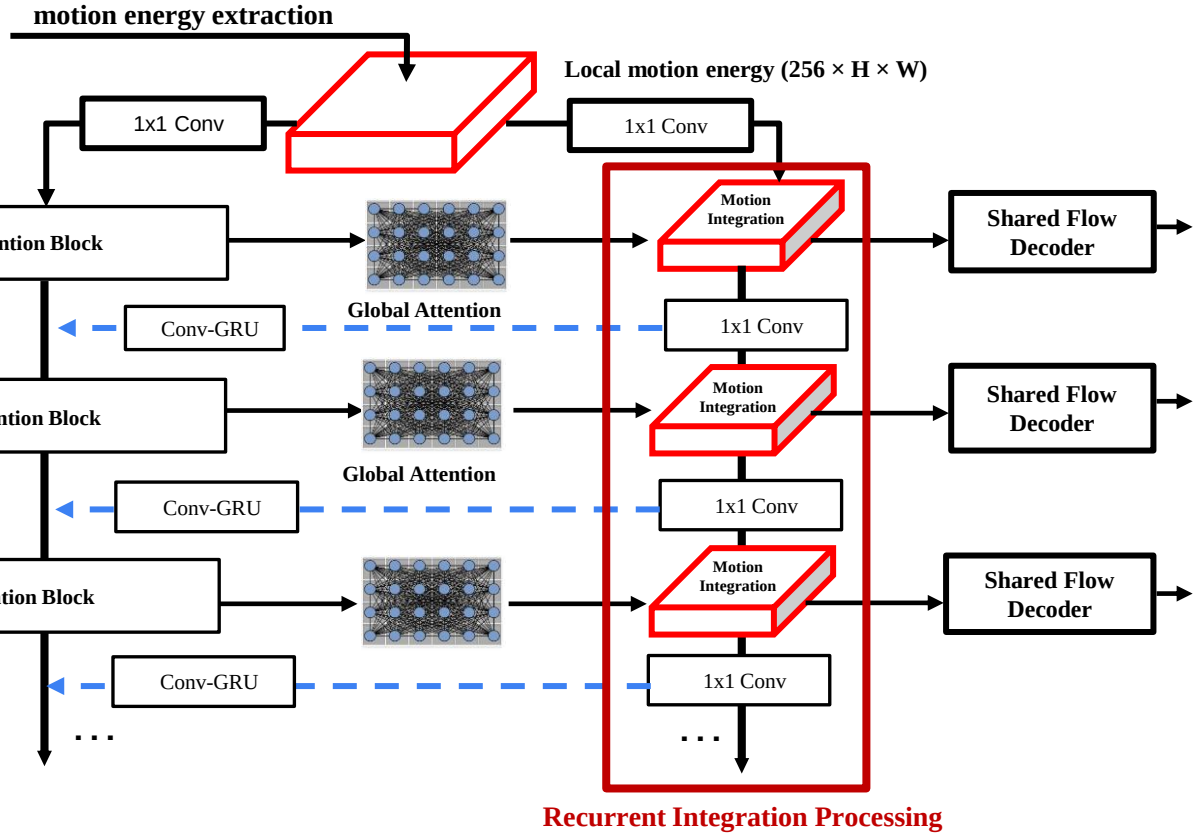
Totally 256 trainable motion energy cells with different spatiotemporal preferences (under Nyquist sampling rate constraint), orientations (0-2π), scales, temporal decays, etc.

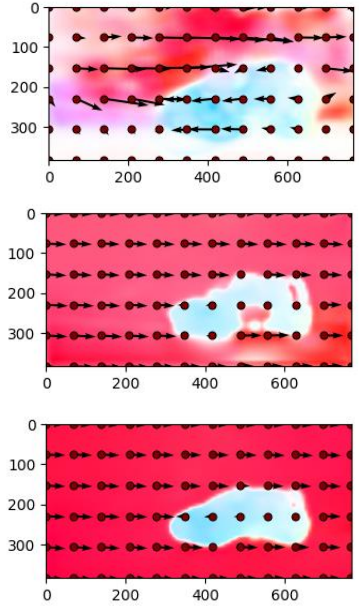


For each inference, input 11 frames across -200~+200 ms.
The temporal filter window is set to 6 frames (200 ms)

See Appendix-B for details

Modeling Two Stages of Motion Perception





Recurrent transformer block based on the gated recurrent unit for global motion integration

Training

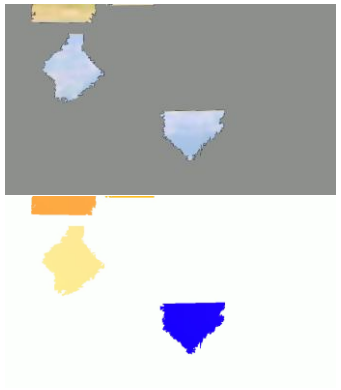
- The **end-to-end supervised training** was applied to fit the motion ground truth in digital videos, which consist of a sizeable multi-frame training set.

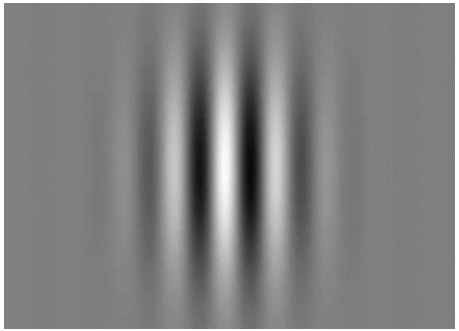
$$L2LossFunction = \sum_{i=1}^{n}(y_{true} - y_{predicted})^2$$

$Y_{predicted}$ is the optical flow from the decoder of each iteration.

Dataset: (with over 8000 samples, each containing 11 images)

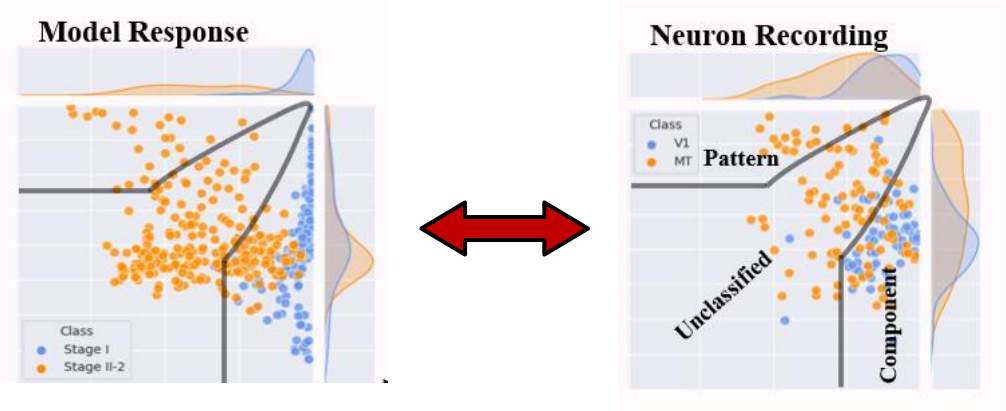
- Sintel Benchmark with ground truth + Natural images with pseudo labels (Created by DNNs)
- Self-made non-texture motion (to generalize non-texture motion widely used in vision research)
- Self-made Drifting gratings (provide prior to solving the aperture problem)

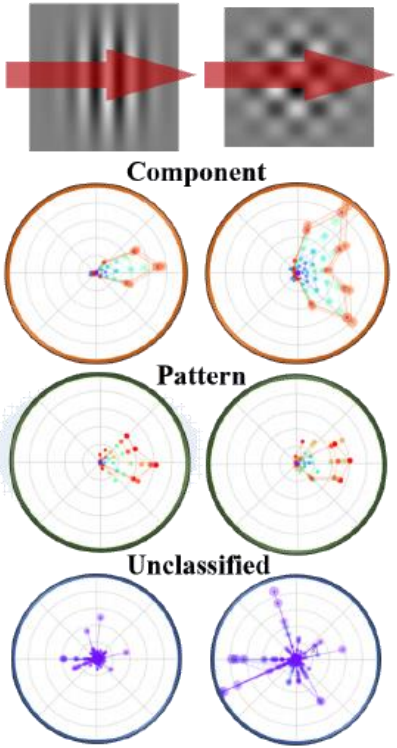




Virtual Neurophysiology-Direction Tuning

- **The distribution of all units from the model captured the tendency of neurons' distributions on partial correlation space.**





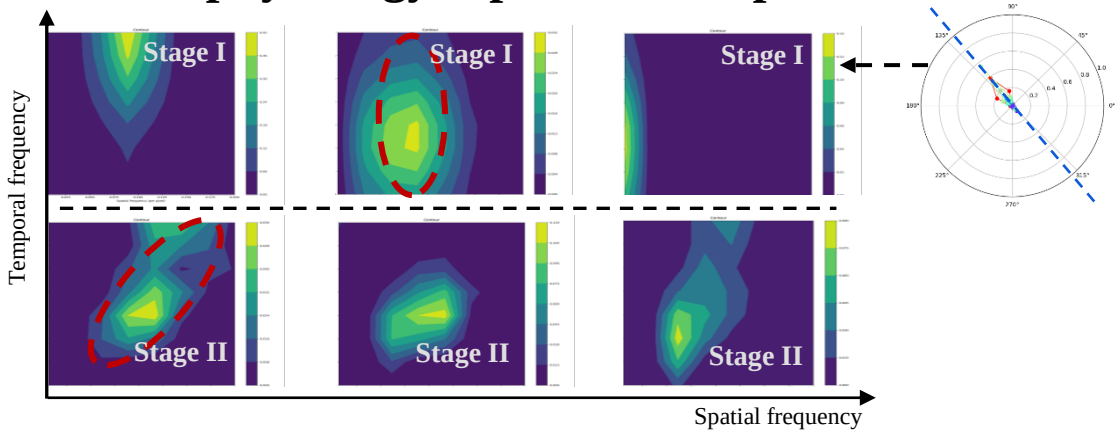
To quantify the directional tuning properties of neurons as recursive integration increases, we adopted a pair of partial correlations to judge whether a cell belongs to a plaid/component.

$$R_c = \frac{\left(r_c - r_p r_{cp}\right)}{\sqrt{\left(\left(1-r_p^2\right)\left(1-r_{cp}^2\right)\right)}} \qquad R_p = \frac{\left(r_p - r_c r_{cp}\right)}{\sqrt{\left(\left(1-r_c^2\right)\left(1-r_{cp}^2\right)\right)}}$$

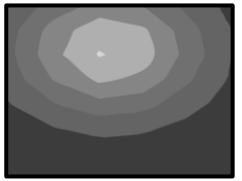
[▲] Movshon J A, Adelson E H, Gizzi M S, et al. Pattern recognition mechanisms[J]. 1985.

In silico Neurophysiology

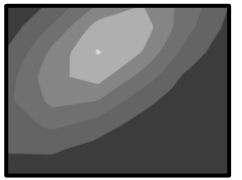
In Silico Neurophysiology: Spectral Receptive Field



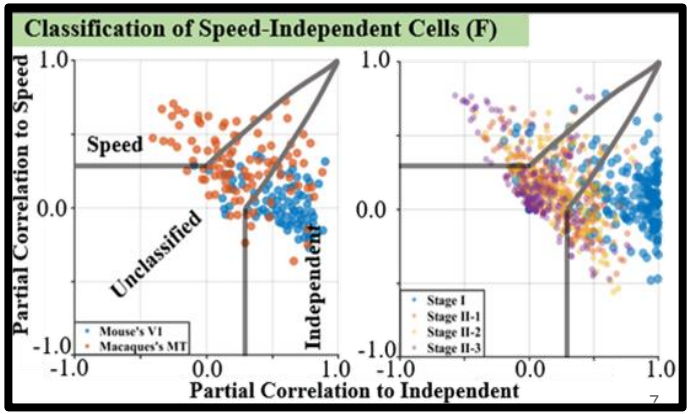
Neuron recordings of monkey's v1



Neuron recordings of monkey's MT



- In contrast to the first stage, the second stage demonstrates a slanted direction of the spectral receptive field, which implies that the cells in the second stage **have a velocity tuning capability**, consistent with the neural recordings in the MT region.



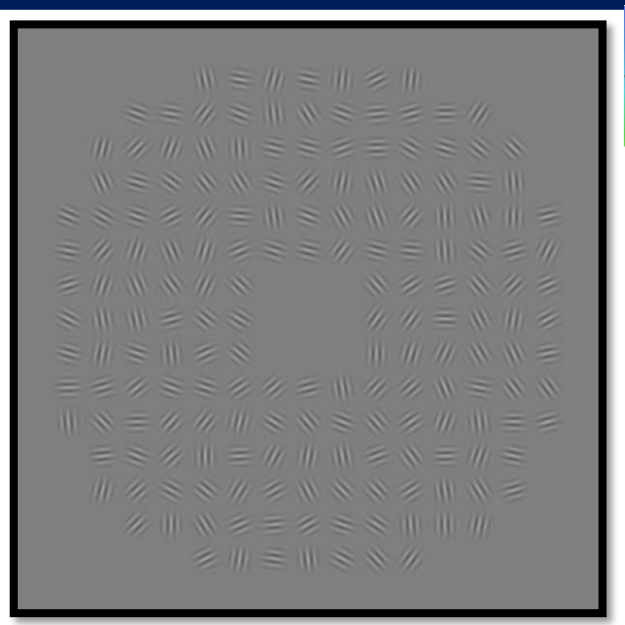
Perrone, J.A. & Thiele, A. Speed skills: measuring the visual speed analyzing properties of primate MT neurons. Nat. Neurosci. 4, 526-532

Motion Integration

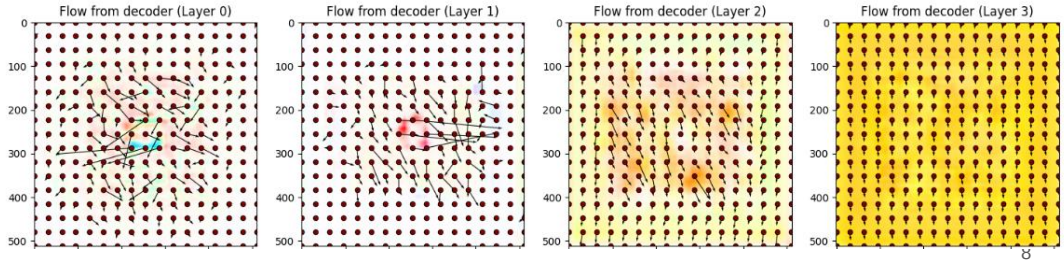
- The MT region of humans could integrate the motion signal from the V1 cells so that one could perceive a global downward motion, which is well captured by our two-stage structures.

Each local region contains different motion directions and different speeds, but globally it is easy to perceive a downward motion.

Stimuli:



Model Response:

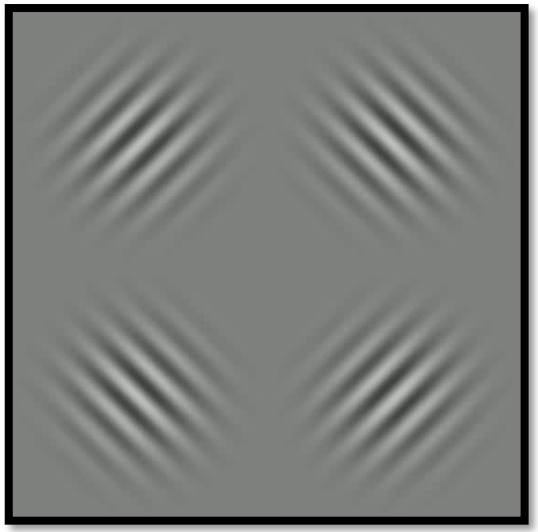


Amano K, Edwards M, Badcock D R, et al. Adaptive pooling of visual motion signals by the human visual system revealed with a novel multi-element stimulus[J]. Journal of vision, 2009, 9(3): 4-4.

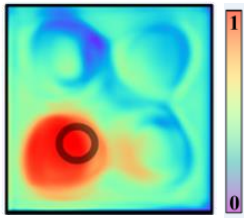
Adaptive Motion Integration

- The proposed model replicates the human's adaptive motion integration strategies: the local ambiguous motion (**Local-grating**) is easier to integrate across long-distance.

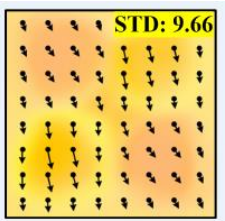
Each to integrate (with peripheral vision)



Unit's Connection Heatmap

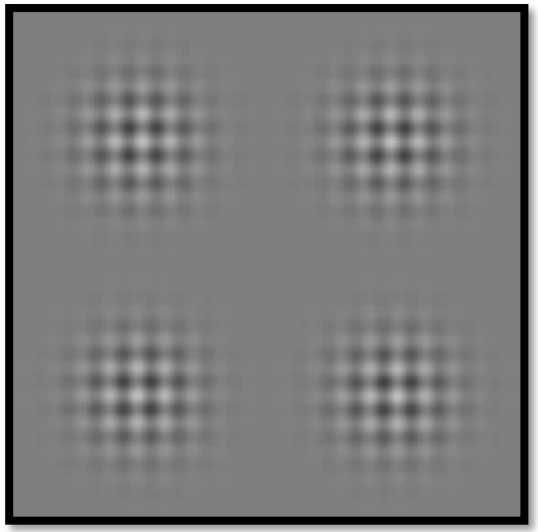


Optical flow

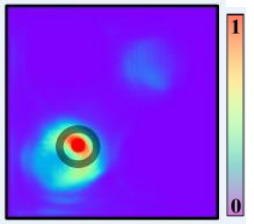


Local-grating (Global integration)

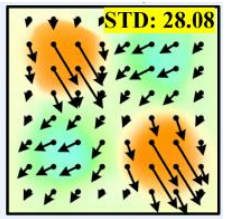
Hard to integrate



Unit's Connection Heatmap



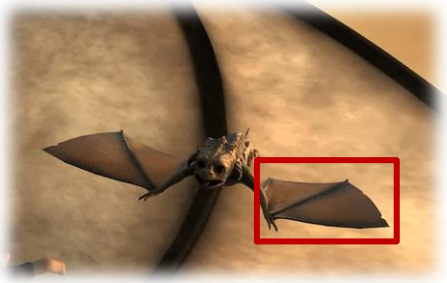
Optical flow

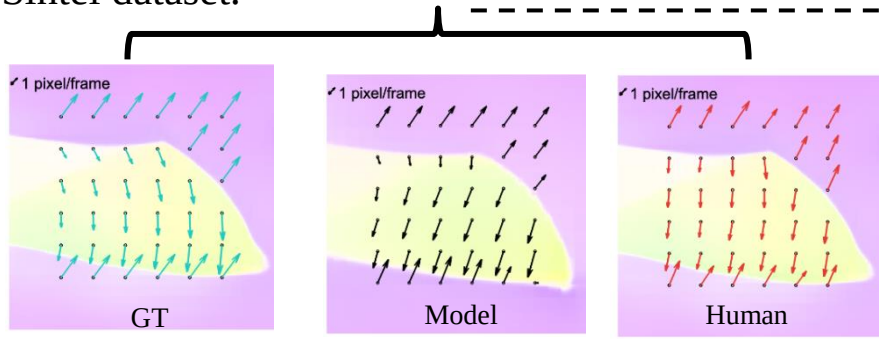


Local-plaid (local integration)

Compare to Psychophysical Human Response

- Compare to Human Response on Sintel dataset.



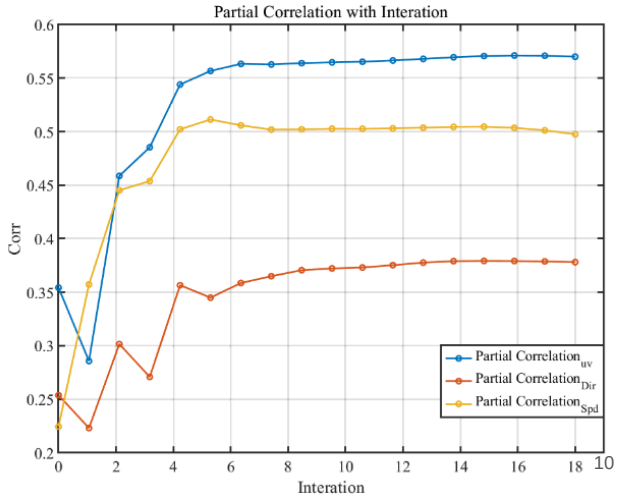


Partial Correlation:

Partial correlation for direction/speed between the model prediction and the human response with the effects of GT removed.

Table 1: ***Model v.s. Human v.s. GT.*** ρ : Partial correlation between human & model controlling GT; r: Pearson correlation coefficient; epe: vector end-point error; uv, dir, spd represent motion components in Cartesian space, direction, and speed, respectively. **3DCNN**: consists of multi-layer 3DCNNs with residual connections; **DorsalNet**: a pre-trained DorsalNet with frozen parameters, featuring a 3D convolution layer as a flow decoder, trained on natural dense optical flow datasets.

Method	ρ_{uv}	ρ_{dir}	ρ_{spd}	v.s. Human				v.s. GT			
				r_{uv}	r_{spd}	r_{dir}	epe	r_{uv}	r_{spd}	r_{dir}	epe
Farneback[58]	0.27	0.23	0.11	0.41	0.91	0.34	2.02	0.34	0.33	0.92	1.96
FlowNet2.0[11]	0.39	0.26	0.34	0.92	0.90	0.96	0.94	0.95	0.94	0.98	0.47
RAFT[14]	0.20	0.22	0.14	0.92	0.90	0.96	0.93	0.98	**0.99**	**0.99**	**0.25**
RAFT-val	0.43	0.17	0.42	0.92	0.89	0.96	1.01	0.92	0.89	0.98	0.69
AGFlow[57]	0.30	0.16	0.20	0.93	0.90	0.96	0.92	0.98	0.98	0.98	0.27
GMFlow[15]	0.34	0.32	0.17	0.91	0.84	**0.96**	1.03	0.93	0.90	0.97	0.73
FlowFormer[43]	0.36	0.14	0.32	**0.93**	**0.91**	0.95	**0.90**	**0.98**	0.97	0.98	0.42
FFV1MT[59]	0.31	0.16	0.31	0.83	0.64	0.92	1.48	0.59	0.84	0.94	1.29
3DCNN	0.27	0.29	0.42	0.83	0.86	0.95	1.31	0.83	0.86	0.96	1.14
DorsalNet[60]	0.17	0.19	-0.10	0.20	-0.08	0.86	2.35	0.20	-0.04	0.86	2.33
Ours-fixed	-0.02	0.12	0.16	0.31	0.23	0.78	2.24	0.35	0.18	0.80	2.29
Ours-Stage I	0.34	0.23	0.35	0.71	0.71	0.92	1.52	0.67	0.67	0.92	1.49
Ours-Stage II	**0.57**	**0.43**	**0.47**	0.91	0.88	0.95	0.98	0.86	0.87	0.95	1.04



Conclusion

- **Two-Stage Architecture**: We introduced a two-stage architecture that models the entire process of biological motion perception, showing good generalization across stimuli.
- **Attention-Based Motion Integration**: Our novel attention-based recurrent process aligned well with physiological and psychophysical findings, offering insights into motion integration mechanisms.
- **Bridging Human and DNN Perception**: Combining classical motion energy and deep learning technology holds promise for closing the gap between human and deep neural network motion perception systems.