bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.451701; this version posted October 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Your head is there to move you around: Goal-driven models of the primate dorsal pathway

Patrick J Mineault
patrick.mineault@gmail.com

Shahab Bakhtiari
Mila, McGill University
bakhtias@mila.quebec

Blake A Richards
Mila, Montreal Neurological Institute, McGill University
blake.richards@mila.quebec

Christopher C Pack
Montreal Neurological Institute, McGill University
christopher.pack@mcgill.ca

Abstract

Neurons in the dorsal visual pathway of the mammalian brain are selective for motion stimuli, with the complexity of stimulus representations increasing along the hierarchy. This progression is similar to that of the ventral visual pathway, which is well characterized by artificial neural networks (ANNs) optimized for object recognition. In contrast, there are no image-computable models of the dorsal stream with comparable explanatory power. We hypothesized that the properties of dorsal stream neurons could be explained by a simple learning objective: the need for an organism to orient itself during self-motion. To test this hypothesis, we trained a 3D ResNet to predict an agent's self-motion parameters from visual stimuli in a simulated environment. We found that the responses in this network accounted well for the selectivity of neurons in a large database of single-neuron recordings from the dorsal visual stream of non-human primates. In contrast, ANNs trained on an action recognition dataset through supervised or self-supervised learning could not explain responses in the dorsal stream, despite also being trained on naturalistic videos with moving objects. These results demonstrate that an ecologically relevant cost function can account for dorsal stream properties in the primate brain.

1 Introduction

The mammalian visual cortex is organized into two processing streams [1]: the ventral stream, where neurons are selective for object class and identity; and the dorsal stream, where neurons are selective for motion. Neurons in the ventral stream exhibit selectivity for increasingly complex stimulus features at successive stages, from oriented lines in V1, to textures in V2, curved lines in V4, and culminating in representations of natural objects in the inferotemporal (IT) cortex [2, 3, 4, 5]. The myriad response properties within and across stages have been difficult to understand computationally [6].

However, in recent years, a large body of work [7, 8, 9, 10, 11, 12, 13] has found that modern convolutional neural networks trained on image classification develop representations that match those found in the ventral stream. Early CNN layers match primary visual cortex (V1), while higher-level layers better match higher-level ventral stream areas, both in terms of qualitative preferred

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.451701; this version posted October 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

features [12] and quantitative predictions of responses to arbitrary stimuli [13, 14]. Moreover, models which perform well on ImageNet image classification tend to explain a larger proportion of the variance in ventral stream area IT [8]. It has recently been found that high-performing networks can emerge through more biologically plausible self-supervised training [15, 16]. These results make it possible to interpret the sometimes baffling data about neural responses in the ventral stream in terms of a biologically plausible distributed learning algorithm whose goal is to develop invariant representations that can support object recognition behavior [9, 17].

Although this approach has been similarly fruitful in other domains (e.g., audition [18, 19]), it has not yet been applied to the dorsal visual pathway. From physiological recordings in dorsal stream areas like MT and MST [20, 21, 22], we know that neurons in this pathway are exquisitely selective for motion and increase in receptive field size and complexity along their hierarchy. These properties have inspired different conceptions of dorsal pathway function, including action recognition [23, 24], prediction of image sequences [25], and tracking of object motion [26], to name just a few. At present, there is no way to know which, if any, of these proposals is correct.

We hypothesized that dorsal pathway representations emerge from a simple objective: the need for the organism to orient itself during self-motion. As animals move through the world, they must estimate the parameters of their own motion, in order to avoid collisions, to plan trajectories, and to stabilize their gaze on objects of interest; the latter is critical for maintaining visual acuity. We suggest that this can be accomplished by learning, in a self-supervised way, the relationship between retinal images and self-motion parameters inferred from oculomotor and vestibular signals that exist in the brain [27] [28]. To test this hypothesis, we trained a 3D ResNet to predict the parameters of simulated self-motion - walking speed and head rotation – in short sequences of motion through simulated environments. We found that this network, dubbed *DorsalNet*, learned motion representations that were qualitatively similar to those found in the dorsal visual stream. Specifically, units were tuned for local motion direction in the earliest layers, object motion in intermediate layers, and complex optic flow in the highest layers [29].

To test our hypothesis quantitatively, we built a database of neural recordings from different regions of the dorsal visual pathway in non-human primates [14]. We then compared the ability of different networks to explain responses in areas V1, MT, and MST. We found that DorsalNet consistently outperformed 3D ResNets trained on action recognition in a supervised manner. Both the self-motion estimation objective and the training stimulus seemed to be critical, since 3D ResNets trained with a predictive objective [CPC; 30] or supervised on action sequences showed weaker performance. Thus, we demonstrate that the diverse neural response properties in the dorsal stream can be captured by a network that has the goal of estimating self-motion parameters from natural image sequences, both elucidating the functional role of the dorsal stream and creating a best-in-class, in-silico model of the dorsal stream.

2 Background and related work

Dorsal stream processing The dorsal stream - also known as the *where* pathway - is a network of cortical areas that are selective for visual motion (Figure 1A). It originates in primary visual cortex (area V1) with a subpopulation of neurons that respond selectively to oriented edges moving in a particular direction. These cells project to areas MT/V5 [27, 31], where most neurons respond selectively to motion direction, even for relatively complex stimuli comprised of multiple edges or features [21, 32]. These neurons in turn project to area MST, where many neurons are selective for the kinds of complex motion patterns that arise during locomotion [22]. MST is considered the terminal stage of the dorsal stream, with subsequent areas integrating information from other senses to support diverse roles in action recognition [23], decision-making [33], and spatial memory [34].

Models of the dorsal stream Previous models of the dorsal stream have emphasized different possible functions. Giese and Poggio [23] have argued that the progression of selectivity along the pathway is well-suited to the recognition of biological movements, and this is consistent with studies showing that the ability to identify shapes from motion patterns is disrupted by lesions to area MT [35]. Other models have posited a role for the dorsal pathway in segmenting moving objects [36, 37], predicting future image frames [25], or supporting reaching movements [38]. Finally, a body of computational [39, 40] and experimental [41] work has analyzed the potential role of dorsal stream

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.451701; this version posted October 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

neurons in the perception of heading or path [42]. None of these models has been quantitatively compared to the detailed properties of neural responses in the dorsal stream.

Other models have made this kind of comparison, but they have been based on shallow architectures and fit directly to the data from dorsal stream areas, including V1 [43], MT [44, 45, 46] and MST [47]. Although these models shed light on the mechanisms by which neurons attain their stimulus selectivity, they do not relate in any clear way to the functional hypotheses mentioned above.

We have therefore attempted to link the properties of dorsal stream neurons, obtained from a database of recordings in non-human primate cortex, to specific functional objectives hypothesized in previous work. In this sense our work is in line with the goals of BrainScore [14], which seeks to benchmark ANNs by their ability to explain ventral stream neurons, and to recent work examining self-supervised networks' fits to ventral areas [16, 15].

3 Methods

Training network for self-motion We generated a dataset consisting of short videos (10 frames) of self-motion in AirSim, a package for drone and land vehicle simulations in Unreal Engine [48]. These videos simulated walking along linear trajectories with constant head rotations in two environments (Figure 1B), starting at random positions, varying environmental conditions, hour of day, starting head pose speed and walking speed (Table S1 in the Appendix). Sequences that led to collisions with the environment were removed.

We trained a 6-layer 3D ResNet (layer definitions in table S1) to predict 2 of the components of head rotation (yaw and pitch rotation speed; roll was not simulated) and the 3 components of linear velocity (parametrized as yaw and pitch heading and speed). We chose a 3D ResNet architecture over alternatives for its stable training, wide use in video tasks [49, 50], and the high performance of 2D resnets in modeling the ventral stream [14]. We discretized each component into 72 bins and trained the network with a cross-entropy objective for each of the 5 components. We used the Adam optimizer with a step size of 0.003, batch norm, and trained for 100 epochs.

Neural datasets Datasets are listed in Table 1. All experiments were conducted in non-human primates (macaca fascicularis and macaca mulatta) and were approved by the governing IRB; detailed experimental procedures are available in the corresponding publications. Data are used under the license terms listed on crcns.org or by permission from the authors [47]. Methods varied from dataset to dataset, but generally, non-human primates were instructed to fixate on a small target while a contiguous image sequence was presented for several minutes. In some cases, parts of the image sequence were repeated when fixation was lost. Image sequences consisted of color movies, black-and-white movies, static pictures with simulated motions, and random dot kinematograms. Data was collected using single electrodes or multi-electrode arrays. Where available, we used previously published sorted spikes; when spikes were unsorted, we used multi-unit activity.

We split each dataset into a train and test set [51]; when only a subset of these stimuli were repeated several times, or a dataset had a designated test subset, we used this subset as the test set; in other cases, we split the data into 6-second blocks, and concatenated every 10th block to form a test set. We kept the sampling rate of the image sequence at its natural rate and resampled neural activity at the same rate, indicated in Table 1. We resampled all stimuli spatially to 112x112. In control analyses, we resampled the input to 74x74 or 168x168 to measure the sensitivity of the results to scale. In the case of [47, 44], the seeds originally used to determine the exact location of dots in the random dot kinematograms were lost, hence we regenerated stimuli with dots in different locations; it should be noted that this could limit the maximal performance of networks [52]. All data used in this paper has been previously published; we release preprocessing scripts and PyTorch loaders to facilitate replication.

Aligning ANNs and neural activity We computed latent representations at different layers of the target ANNs, listed in Table 2, in windows of 10 image frames preceding neural activity. We cropped the first and last latent activity frame and downsampled the activity 2-fold temporally to obtain 4 frames of latent representations preceding the neural activity, and spatially averaged and downsampled each layer output to 8 by 8. Following [14], we kept the first 500 PCs of the intermediate representation and used ridge regression to find mappings from latent space to experimental neural

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.451701; this version posted October 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

activity. We selected the ridge parameter using 5-fold cross-validation within the train set. Where test sets with 5 or more disaggregated repeats were available, we report an R score normalized against the maximum attainable R score [53]; otherwise, we report the raw R score.

In control analyses, we replaced ridge regression with a sparse regression estimated through boosting. To fit all the intermediate representations in memory and fit a boosted regression model, we used two different strategies to reduce the memory footprint: downsampling layer outputs (with spatial averaging as in the linear regression; table S4 in the appendix) or subsampling (without spatial averaging; Table S4). We selected the number of boosting iterations using 5-fold cross-validation within the train set. We fit these models on a commodity GPUs including P5000 and 1080Ti locally, in Paperspace and in AWS for a total of $\sim$ 1000 single-GPU-hours. Model weights and code are available [1] under an MIT license.

Contrastive Predictive Coding (CPC) CPC is a self-supervised learning algorithm that learns to predict the next latent state of a sequence (e.g. a video sequence) given its present and past states. The details of the CPC algorithm can be found in [30] and [54], but we summarize it briefly here. A sequence of video frames (x_t) are passed as input to a 3D CNN. The CNN output (z_t), which is a latent representation of the video sequence, is fed to a recurrent neural net (RNN). The RNN aggregates past and present latent states (i.e. CNN output) and generates a context variable as its output (c_t). The context variable is then passed to a single layer MLP which predicts the future latent state of the video. The predicted latent state and the true latent state (positive pairs), along with some incorrect examples of the next state (negative pairs) are given to a contrastive loss function. Minimizing the contrastive loss maximizes the similarity of the predicted and the true next states, and minimizes the similarity of the predicted and the false next states.

4 Results

4.1 3D resnets trained for self-motion learn dorsal-like representations

We hypothesized that learning to estimate self-motion from visual inputs would lead to dorsal stream-like representations. As in the ventral stream, these representations begin in V1 with receptive fields that encode simple, local features of stimuli. Through subsequent recombinations at different layers, more complex and ecologically relevant encoding emerges. To test this hypothesis, we generated self-motion videos in a simulation environment, and trained a 3D ResNet to predict its self-motion parameters, namely head rotation and linear locomotion (see Methods for details).

Qualitative matches to the dorsal stream The 6-layer 3D ResNet trained in this way learned representations similar to single units in the primate dorsal stream. We focus our attention here on layers 1, 2 and 3 of the network. Preferred features of layer 1 contained many spatiotemporally slanted filters (Figure 1C), which are the building blocks of motion selectivity in primate V1 [55]. We quantified this slant with the separability index $\sigma_1^2 / \sum_i \sigma_i^2$ from the singular values of the grayscale filters σ_i; this matched values reported in the literature for V1 [20] [.72 +/-.16 for trained network, .71 +/- .15 in real V1 neurons; figure S1 in the appendix].

To gain insight into the stimulus selectivity of these representations, we generated optimal stimuli for individual units in intermediate layers of the network by optimization [12]; we present static images of the intermediate preferred frame here, while animations can be visualized on the companion website[2]. Probed in this fashion, many intermediate features in layer 1 preferred what looked like drifting gratings (examples in Figure 1D), consistent with the selectivity of V1 cells [55, 56]. Hence, to further probe the selectivity of these units, we used full contrast, drifting gratings of different spatial and temporal frequencies, placed in the center of the visual field. Tuning curves in layer 1 (samples in Figure 1) tended to have a bias towards direction selectivity, with a mean circular variance at the preferred spatial and temporal frequency of 0.75 and a median direction selectivity index - defined as $1 - r_{pref}/r_{antipref}$ on the centered tuning curves - of 0.98. This is somewhat higher than is typically found in V1 [57], likely due to lack of noise, but it is close to the selectivity of the V1 neurons that actually project to higher levels of the dorsal visual pathway [58].

[1] https://github.com/patrickmineault/your-head-is-there-to-move-you-around

[2] https://your-head-is-there-to-move-you-around.netlify.app

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.451701; this version posted October 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

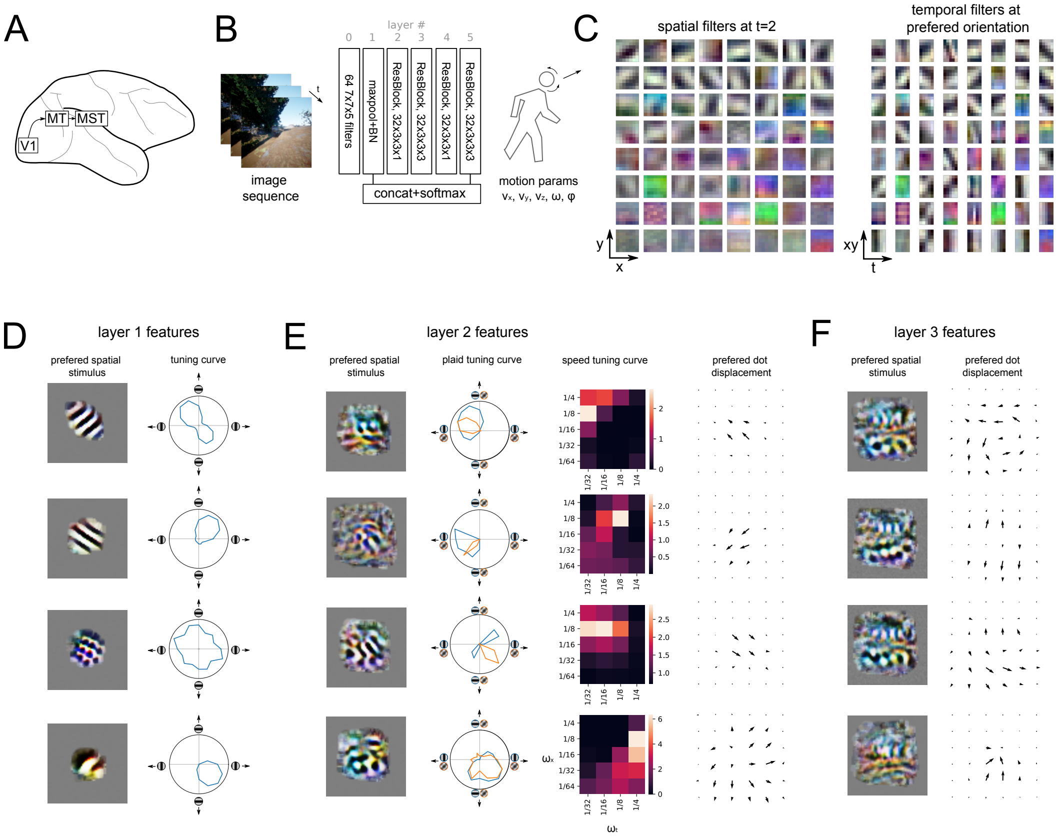

Figure 1: A 3D ResNet trained for self-motion estimation learns dorsal-like representations. A. The goal of this study is to model the dorsal visual stream, including V1, MT and MST. B. The 3D Resnet model is trained to estimate self-motion parameters from image sequences. C. Weights of the first layer. First layer filters are selective in space-time. D. Sample tuning of layer 1 features. Layer 1 contains many direction-selective cells reminiscent of V1. E. Sample tuning of layer 2 features. Many layer 2 features exhibit tuning for rigid motion, similar to MT. F. Sample tuning curve of layer 3 features. Many cells in layer 3 are tuned for complex optic flow, like MST

Layer 2 units tended to prefer more spatially broadband moving stimuli, not unlike the plaids conventionally used in probing MT cells [21] (Figure 1E, left column). Indeed, probing the representations with sums of gratings revealed similar selectivity to a single grating in a subset of cells (Figure 1E, middle column; pattern selectivity plots in Figure S1 in the appendix). These cells likely encode stimulus velocity in a manner that is invariant of the composition of the stimulus [21]. Like MT cells, subunits in this layer tended to be highly direction selective, with the average circular variance of the direction tuning curves being .41.

MT cells are also known to be selective for stimulus speed, which is the ratio of temporal to spatial frequencies [59]. A similar kind of selectivity emerged in layer 2 of the model, where many units preferred higher temporal frequencies when the spatial frequencies were higher (example tuning curve in Figure 1E). To quantify this selectivity, we probed the model units with a range of spatiotemporal frequencies and fit the data with slanted Gaussian functions [60], which revealed a mean speed selectivity index of -.14 in layer 1, compared to 0.58 in layer 2, the latter being similar to the value of 0.52 reported in MT [60]. Probing layer 2 units with moving dots, we found a majority of neurons with simple receptive fields that prefer linear motion, with a smaller number of complex receptive fields (Figure 1E, bottom right).

Finally, we found many cells in layer 3 that combined the outputs of lower-level units to generate selectivity for more complex motion patterns (example cells in Figure 1F). Dot pattern probes revealed selectivity for rotations, spirals or single axis expansion. As in primate area MST, these units tended

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.451701; this version posted October 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

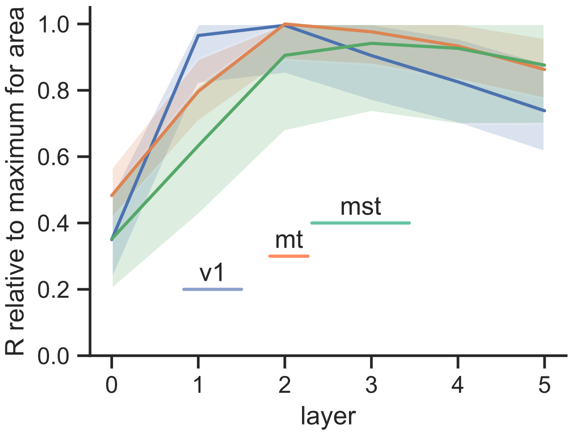

Figure 2: Layers 1, 2, and 3 of DorsalNet best match areas V1, MT and MST. Lines show correlation (R) relative to maximum for area. Horizontal lines: 95% CI of layer with maximal alignment to area.

to emphasize expansion motion rather than contraction, similar to the bias experienced during forward navigation (Figure S1) [61].

Area	Dataset	Data	Sampling	Stimulus
V1	crcns-pvc1 [62, 63]	23 multi-units	30Hz	Color movies
	crcns-pvc4 [64, 65, 66]	25 single units	75Hz	B&W movies
MT	crcns-mt1 [44, 67]	88 single units	30Hz	optic flow kinematograms
	crcns-mt2 [46, 68]	44 single units	83Hz	B&W motion-enhanced movies
MST	packlab-mst [47]	36 single units	30Hz	optic flow kinematograms

Table 1: Datasets

Regression analysis of representations Given that the trained network recapitulated many qualitative properties of the dorsal stream, we next investigated whether they quantitatively matched dorsal stream areas, for which single-neuron data was available. We used ridge regression to learn a mapping from latent representations at each layer of the network to single neural responses to complex stimuli, including black and white and color movies, along with random dot kinematograms (See Methods and Table 1 for details). We learned a separate mapping for each layer of the network, allowing us to match the depth of the network to each brain area. As seen in Figure 2, this showed a hierarchical progression, with higher-level cortex matching higher-level layers in the ResNet. The average best matching layer across cells with report correlation greater than .01, illustrated by the horizontal lines, was 1.1 for V1 cells [(0.9, 1.5) 95% CI, bootstrap across cells], 2.0 for MT cells [(1.8, 2.3)] and 2.9 for MST cells [(2.3, 3.4)].

We noticed that the mapping was less distinct for layer 3. We investigated this further by measuring the response of the network to a gauntlet of stimuli from the Airsim dataset. Centered kernel alignment (CKA) [69] revealed that while layers 0, 1, and 2 had highly distinct representations, subsequent layers were less distinct S3. We also investigated the robustness of the mapping to a change of scale of the input sequences [70]. We saw some minor shifts: the median mean layer assignment for V1 was 1.0 at 0.66X scale, 1.1 at the standard 1X scale, and 1.3 at 1.5X scale (Figure S2). Overall, however, mappings were robust to a change of scale. Thus, broadly speaking, layers 1, 2 and 3 of the network recapitulated V1, MT and MST, respectively.

4.2 Networks with alternative objectives do not account for responses in the dorsal stream

Action recognition networks To examine the specificity of these results, we tested other networks trained with different objective functions. Action recognition is a popular computer vision task,

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.451701; this version posted October 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Category	Name	Dataset	License	Notes
SlowFast [49]	slowfast i3d	Kinetics400	Apache	Fast branch only
R3D [50]	r3d_18 r2plus1_18 mc3_18	Kinetics400	BSD	
CPC [30]	cpc_ucf cpc_airsim	UCF101 Airsim	own work	R3D with 10 res blocks R3D with 10 res blocks
Gabors [46]	gabor gabor_nomotion	- -	own work	 Opposite dirs averaged
MotionNet [26]	motionnet	shifted images	CCBY4.0	
DorsalNet	dorsalnet	Airsim	own work	R3D with 4 res blocks

Table 2: Models tested

	V1		MT		MST
	pvc1	pvc4	mt1	mt2	mst
slowfast	**.471** (.034)	.361 (.042)	.211 (.018)	.281 (.015)	.189 (.044)
i3d	**.457** (.036)	**.389** (.046)	.213 (.018)	.284 (.015)	.219 (.044)
r3d_18	.403 (.032)	**.383** (.042)	.217 (.018)	.289 (.015)	.224 (.046)
r2plus1d_18	.428 (.035)	**.382** (.042)	.215 (.018)	.282 (.015)	.226 (.043)
mc3_18	.405 (.034)	**.393** (.045)	.218 (.018)	.276 (.014)	.228 (.045)
cpc_ucf	.271 (.044)	**.394** (.046)	.214 (.018)	.241 (.016)	.190 (.045)
cpc_airsim	.422 (.036)	**.384** (.045)	**.250** (.020)	.360 (.017)	.292 (.045)
gabor_nomotion	.273 (.035)	.353 (.038)	.212 (.018)	.188 (.014)	.248 (.045)
gabor	.325 (.036)	**.366** (.037)	**.249** (.019)	.301 (.015)	.394 (.054)
motionnet	.276 (.042)	**.364** (.039)	.238 (.018)	.333 (.016)	**.441** (.053)
dorsalnet	.364 (.043)	**.370** (.039)	**.251** (.019)	**.381** (.017)	**.454** (.054)

Table 3: DorsalNet quantitatively performs best across the dorsal stream. Table shows normalized pearson correlation (R; see Methods for definition) of different models on different datasets. In parenthesis: standard error of the mean over cells.

and so we tested 3D ResNets trained on Kinetics400 [71]. These networks performed admirably in explaining V1 responses, reaching an average R > .4 on the pvc1 dataset. However, across our MT and MST datasets, performance was poor, failing to exceed that of a null model [72] consisting of a 3D Gabor pyramid (Table 3). We note that only a small fraction of V1 neurons project to the dorsal stream, with the majority projecting to ventral stream areas; we interpret the relative performance in V1 vs. MT and MST as a sign that these networks learned representations more aligned with the *ventral* stream, supporting object recognition and by extension action recognition. Consistent with this interpretation, we found that the first layer of 3D ResNets trained for action recognition did not learn motion in the traditional sense (Figure S5). Instead, their filters were mostly separable in space and time, meaning they were not selective for motion energy *per se*.

CPC Our results indicate that learning to estimate self-motion in a simulated environment creates representations similar to those in the primate dorsal stream. The neural network architecture (3D ResNets) was similar for the self-motion estimation objective and action recognition tasks. However, both the task - prediction of self-motion parameters - and the stimulus ensemble - self-motion sequences in the Airsim environment - differed. To tease apart the relative importance of these two factors, we tested the ability of contrastive predictive coding (CPC) networks [30] to account for responses in the dorsal stream when trained over different stimulus ensembles. CPC is a self-supervised training method that finds predictive latent representations that can distinguish

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.451701; this version posted October 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

between image sequences. Importantly, it is possible to apply the CPC objective to different stimulus ensembles, thereby differentiating between task and stimulus ensemble effects. We trained an 11-layer network with a CPC objective on the UCF101 dataset and our Airsim dataset. The Airsim-trained network performed significantly better than the UCF101-trained network, approaching the performance of DorsalNet in MT but not in MST. Examining first layer filters revealed direction-selective receptive fields after training on the Airsim dataset but not with UCF101 (Figure S5). This is consistent with the training set being necessary, though not sufficient, to match primate dorsal stream neurons.

MotionNet We next tested a much simpler 2-layer network from the neuroscience literature, which was trained to estimate the linear motion of black and white image patches [26]. The original model was a fully connected architecture working on small image patches, and we made it convolutional by tiling. We used the checkpoints shared by the authors as the model weights. This model had not previously been directly benchmarked against neural data, and given the small size of its stimulus ensemble, we did not expect it to perform well. Surprisingly, it scored far better in predicting MT and MST responses than action recognition networks (Table 3). We found in a control analysis (Table S3 in the appendix) that the relative performance of MotionNet could be improved still by spatially scaling up the stimulus, matching the performance of DorsalNet on 2 out of 3 MT and MST datasets. These results are consistent with solving 2D motion being an important sub-goal of the dorsal stream.

We next asked whether there existed a one-to-one or few-to-one relationship between model subunits and single neurons. Using sparse regression, we found that DorsalNet better matched individual neurons across all MT and MST datasets than MotionNet, regardless of scaling (Tables S4 and S5 in the appendix). Thus, DorsalNet subunits were more directly aligned to single neurons across the dorsal stream.

5 Self-motion estimation performance correlates with dorsal stream match

Across our baselines, there was a large range in the ability of different models to reproduce dorsal stream data. We asked whether this heterogeneity could be linked to performance on a self-motion estimation task. We froze the weights of our baseline networks and trained linear decoders to estimate self-motion parameters on the AirSim dataset from hidden layer representations. We excluded DorsalNet and Airsim-trained CPC from these comparisons. Across our baselines, there was a highly significant correlation between self-motion estimation performance and match to MT and MST neurons (Figure S4; Table S2 in the appendix). Interestingly, when looking at individual self-motion parameters, head rotation estimation accuracy was most correlated with performance on MT and MST datasets. Thus, those networks which happen to be best at self-motion estimation, especially head rotation, can best explain responses in the dorsal stream, consistent with a formative role of self-motion estimation in dorsal stream representations.

6 Limitations

Multiple interpretations We show that learning to estimate one's self-motion from visual cues leads to representations which are similar to those of the dorsal stream. We benchmark against several other candidate models, including localized frequency detectors, which form a sparse basis for images [73], predictive coding models, models trained for action recognition, and models trained trained to estimate the motion of small image patches. While DorsalNet performed best overall across the dorsal stream, we found that MotionNet and a CPC-based network trained on our AirSim dataset were close contenders. With the available data, we cannot conclusively rule out that these alternative objectives, with the right tweaks, could not account for the data. One interesting possibility is that, as MotionNet hints, solving rigid 2D motion is a sub-goal of the dorsal stream; and, as DorsalNet shows, the supervisory signal needed to learn to solve that sub-goal could come from head movements, especially head rotations, via efference copy. An open benchmark in the style of [14] could reveal other objectives compatible with the data and refine these results.

Data limitations To the best of our knowledge, we used all of the relevant publicly available non-human primate data for this study. Most of this data was collected more than a decade ago in time-consuming single-electrode experiments, with electrode drift, loss of fixation, short recording

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.451701; this version posted October 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

times and small numbers of recordings per experiment being significant limitations. The MST dataset in particular is not very discriminative across models. Differences in stimuli and number of repetitions make absolute comparisons across areas difficult. Improvements in recording technology as well as better-designed hypothesis-driven studies will allow the collection of more discriminative data in the future. Our study paves the way for closed-loop experiments to verify that the estimated stimuli indeed maximally drive dorsal stream neurons [74, 75].

7 Discussion

Systems neuroscience aims to explain how the brain solves behavioral tasks at the algorithmic level [17]. While a rich literature has linked the ventral visual stream to the task of object recognition, little work has focused on understanding how and why dorsal streams representations emerge. Noting the critical role of self-motion estimation across the animal kingdom [76], we hypothesized that training an artificial neural net on self-motion estimation from image sequences would lead to representations similar to the dorsal stream. We verified this qualitatively by probing networks with artificial stimuli and by finding maximizing stimuli. We confirmed these findings quantitatively by benchmarking existing computer vision networks on a gauntlet of neural data [14].

In the framework of [17], the objective, learning rule and architecture specify how a task is to be solved by an artificial or biological neural network. Implicit in the framework is a fourth critical ingredient: the dataset, or distribution of training examples. Our work focuses on how an objective, learning rule and dataset interact to form representations similar to the dorsal stream. In contradistinction with previous work, we focus on a single architecture of 3D ResNets, a coarse approximation to early and intermediate visual processing stages, highlighting the formative role of objective, learning rule and dataset in the creation of useful representations for action.

Maximizing stimuli reveal selectivity Systems identification has long been used in systems neuroscience to estimate preferred stimuli in different brain areas [55, 56, 43, 46, 51, 20, 47, 77, 44]. More recently, systems identification has been used to better understand mechanisms of selectivity in deep neural nets [12]. Given the breadth of available systems identification results in brains, we suggest that systems identification is a particularly powerful tool to relate brains and artificial neural nets, especially when combined with benchmarking: it can offer clues as to why certain networks perform better than others. In this article, we identified direction selectivity in the first layer as a strong clue that networks develop good motion representations.

Action recognition is poorly aligned to the dorsal stream Our work shows that ANNs trained on the standard computer vision task of action recognition fail to learn motion representations that correlate with single neurons in MT and MST. [78] reported that on Kinetics400 and UCF101, a single image is sufficient to get within 6% of the action recognition accuracy of a full image sequence, indicating that motion has a limited role in action recognition in these datasets. Motion selectivity can be reintroduced via a parallel optic flow pathway [79] or by enforcing that the network reproduce dense optic flow following early layers [80, 81], with modest improvements in classification accuracy. Our benchmarks strongly suggest that current action recognition datasets can be solved without motion and that good motion representations don't emerge from supervised learning on them alone.

Self-supervision through cross-modal prediction We train DorsalNet in a supervised way. From the agent's perspective, however, corollary discharges of the motor plan are available, as well as vestibular inputs. Thus, the objective can be viewed as a self-supervised objective which aims to predict one modality or channel of the input from the other, in line with other proxy tasks including colorization and audiovisual alignment [82, 83, 84]. Because multisensory integration and corollary discharges are ubiquitous across mobile animals [85], self-supervision through cross-modal prediction could be potentially widely used across species to learn useful representations.

Evolution and learning in sensory systems Thompson [86] identifies four set of constraints against which in silico models of sensory systems can be evaluated:

- Whether it can perform a relevant task
- Whether it accounts for neural activity

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.451701; this version posted October 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- Whether it is biologically plausible
- Whether it could have evolved

We presented a model of the dorsal stream that is trained to estimate self-motion. It accounts for neural responses in 3 different areas, taken from 5 different datasets. The model weights can be learned by the agent through biologically plausible self-supervision, since the approximate parameters of self-motion are known to the agent, via corollary discharges and vestibular and proprioceptive inputs [28]. Self-motion estimation is particularly important for gaze stabilization, which evolved in tandem with the earliest visual functions [87, 88], and continues to be necessary for visual processing, including that performed in the ventral pathway [89]. Given this evolutionary pressure, some aspects of the dorsal pathway are likely hard-coded in the genome, while others are learned through development [90]; further work will focus on better understanding the relative role of evolution vs. learning in dorsal stream processing. This work and its follow-ups thus have the potential to elucidate long-standing questions about how sensory systems evolved.

Acknowledgments and Disclosure of Funding

This work was supported by a NSERC (Discovery Grant: RGPIN-2020-05105; Discovery Accelerator Supplement: RGPAS-2020-00031), Healthy Brains, Healthy Lives (New Investigator Award: 2b-NISU-8; Innovative Ideas Grant: 1c-II-15), and CIFAR (Canada AI Chair; Learning in Machine and Brains Fellowship). CCP was funded by a CIHR grant (MOP-115178). This work was also funded by the Canada First Research Excellence Fund (CFREF Competition 2, 2015-2016) awarded to the Healthy Brains, Healthy Lives initiative at McGill University, through the Helmholtz International BigBrain Analytics and Learning Laboratory (HIBALL).

References

[1] Leslie G. Ungerleider and Mortimer Mishkin. Two cortical visual systems. *Analysis of visual behavior*, pages 549–586, 1982.

[2] A. Pasupathy and C. E Connor. Population coding of shape in area V4. *Nature Neuroscience*, 5(12):1332–1338, 2002.

[3] Nicole C. Rust and James J. DiCarlo. Selectivity and Tolerance ("Invariance") Both Increase as Visual Information Propagates from Cortical Area V4 to IT. *The Journal of Neuroscience*, 30(39):12978 –12995, 2010.

[4] S. L Brincat and C. E Connor. Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature Neuroscience*, 7(8):880–886, 2004.

[5] Jeremy Freeman, Corey M. Ziemba, David J. Heeger, Eero P. Simoncelli, and J. Anthony Movshon. A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, 16(7):974–981, July 2013.

[6] Andrew J Parker. Intermediate level cortical areas and the multiple roles of area V4. *Current Opinion in Physiology*, 16:61–67, August 2020.

[7] Daniel Yamins, Ha Hong, Charles Cadieu, and James J. DiCarlo. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. 2013. Publisher: Neural Information Processing Systems Foundation.

[8] Daniel LK Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014. Publisher: National Acad Sciences.

[9] Daniel LK Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016. Publisher: Nature Publishing Group.

[10] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014. Publisher: Public Library of Science San Francisco, USA.

[11] Grace W. Lindsay. Convolutional neural networks as a model of the visual system: past, present, and future. *Journal of cognitive neuroscience*, pages 1–15, 2020. Publisher: MIT Press.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.451701; this version posted October 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[12] Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, and Ludwig Schubert. Thread: Circuits. *Distill*, 5(3):e24, March 2020.

[13] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015. Publisher: Soc Neuroscience.

[14] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, and Kailyn Schmidt. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018. Publisher: Cold Spring Harbor Laboratory.

[15] Talia Konkle and George Alvarez. Deepnets do not need category supervision to predict visual system responses to objects. *Journal of Vision*, 20(11):498–498, October 2020. Publisher: The Association for Research in Vision and Ophthalmology.

[16] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, and Daniel L. K. Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3), January 2021. Publisher: National Academy of Sciences Section: Biological Sciences.

[17] Blake A. Richards, Timothy P. Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, and Surya Ganguli. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019. Publisher: Nature Publishing Group.

[18] Alexander J. E. Kell, Daniel L. K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, and Josh H. McDermott. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3):630–644.e16, May 2018.

[19] Alexander JE Kell and Josh H. McDermott. Deep neural network models of sensory systems: windows onto the role of task constraints. *Current opinion in neurobiology*, 55:121–132, 2019. Publisher: Elsevier.

[20] Christopher C. Pack, Bevil R. Conway, Richard T. Born, and Margaret S. Livingstone. Spatiotemporal structure of nonlinear subunits in macaque visual cortex. *Journal of Neuroscience*, 26(3):893–907, 2006. Publisher: Soc Neuroscience.

[21] J. Movshon, E. H. Adelson, M. S. Gizzi, and William T. Newsome. The analysis of moving visual patterns. *Pattern recognition mechanisms*, pages 117–151, 1985. Publisher: Vatican Press.

[22] Charles J. Duffy and Robert H. Wurtz. Sensitivity of MST neurons to optic flow stimuli. I. A continuum of response selectivity to large-field stimuli. *Journal of neurophysiology*, 65(6):1329–1345, 1991. Publisher: American Physiological Society Bethesda, MD.

[23] Martin A. Giese and Tomaso Poggio. Neural mechanisms for the recognition of biological movements. *Nature Reviews. Neuroscience*, 4(3):179–192, March 2003.

[24] Hueihan Jhuang. *Dorsal stream : from algorithm to neuroscience*. Thesis, Massachusetts Institute of Technology, 2011. Accepted: 2011-09-27T18:31:39Z.

[25] Janneke F. M. Jehee, Constantin Rothkopf, Jeffrey M. Beck, and Dana H. Ballard. Learning receptive fields using predictive feedback. *Journal of Physiology-Paris*, 100(1):125–132, July 2006.

[26] Reuben Rideaux and Andrew E. Welchman. But still it moves: static image statistics underlie how we see motion. *Journal of Neuroscience*, 40(12):2538–2552, 2020. Publisher: Soc Neuroscience.

[27] William T. Newsome, Robert H. Wurtz, and Hidehiko Komatsu. Relation of cortical areas MT and MST to pursuit eye movements. II. Differentiation of retinal from extraretinal inputs. *Journal of neurophysiology*, 60(2):604–620, 1988. Publisher: American Physiological Society Bethesda, MD.

[28] Yong Gu, Paul V. Watkins, Dora E. Angelaki, and Gregory C. DeAngelis. Visual and nonvisual contributions to three-dimensional heading selectivity in the medial superior temporal area. *Journal of Neuroscience*, 26(1):73–85, 2006. Publisher: Soc Neuroscience.

[29] R. A. Andersen, L. H. Snyder, D. C. Bradley, and J. Xing. Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience*, 20:303–330, 1997.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.451701; this version posted October 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748 [cs, stat]*, January 2019. arXiv: 1807.03748.

[31] Semir M. Zeki. Functional organization of a visual area in the posterior bank of the superior temporal sulcus of the rhesus monkey. *The Journal of physiology*, 236(3):549–573, 1974. Publisher: Wiley Online Library.

[32] Farhan A. Khawaja, James MG Tsui, and Christopher C. Pack. Pattern motion selectivity of spiking outputs and local field potentials in macaque visual cortex. *Journal of Neuroscience*, 29(43):13702–13709, 2009. Publisher: Soc Neuroscience.

[33] Alexander C. Huk and Michael N. Shadlen. Neural activity in macaque parietal cortex reflects temporal integration of visual motion signals during perceptual decision making. *Journal of Neuroscience*, 25(45):10420–10436, 2005. Publisher: Soc Neuroscience.

[34] Bijan Pesaran, John S. Pezaris, Maneesh Sahani, Partha P. Mitra, and Richard A. Andersen. Temporal structure in neuronal activity during working memory in macaque parietal cortex. *Nature neuroscience*, 5(8):805–811, 2002. Publisher: Nature Publishing Group.

[35] P. H. Schiller. The effects of V4 and middle temporal (MT) area lesions on visual performance in the rhesus monkey. *Visual Neuroscience*, 10(4):717–746, August 1993.

[36] Steven J. Nowlan and Terrence J. Sejnowski. A selection model for motion processing in area MT of primates. *Journal of Neuroscience*, 15(2):1195–1214, 1995. Publisher: Soc Neuroscience.

[37] Richard S. Zemel and Terrence J. Sejnowski. A model for encoding multiple object motions and self-motion in area MST of primate visual cortex. *Journal of Neuroscience*, 18(1):531–547, 1998. Publisher: Soc Neuroscience.

[38] Claudio Galletti and Patrizia Fattori. Neuronal mechanisms for detection of motion in the field of view. *Neuropsychologia*, 41(13):1717–1727, January 2003.

[39] J. A. Perrone. Model for the computation of self-motion in biological systems. *Journal of the Optical Society of America A, Optics and Image Science*, 9(2):177–194, 1992.

[40] M. Lappe. Computational mechanisms for optic flow analysis in primate cortex. *International Review of Neurobiology*, pages 235–268, 2000.

[41] Yong Gu, Christopher R. Fetsch, Babatunde Adeyemo, Gregory C. DeAngelis, and Dora E. Angelaki. Decoding of MSTd population activity accounts for variations in the precision of heading perception. *Neuron*, 66(4):596–609, 2010.

[42] Oliver W. Layton and N. Andrew Browning. A unified model of heading and path perception in primate MSTd. *PLoS Comput Biol*, 10(2):e1003476, 2014. Publisher: Public Library of Science.

[43] Nicole C. Rust, Odelia Schwartz, J. Anthony Movshon, and Eero P. Simoncelli. Spatiotemporal elements of macaque v1 receptive fields. *Neuron*, 46(6):945–956, 2005.

[44] Yuwei Cui, Liu D. Liu, Farhan A. Khawaja, Christopher C. Pack, and Daniel A. Butts. Diverse suppressive influences in area MT and selectivity to complex motion features. *Journal of Neuroscience*, 33(42):16715–16728, 2013. Publisher: Soc Neuroscience.

[45] Nicole C. Rust, Valerio Mante, Eero P. Simoncelli, and J. Anthony Movshon. How MT cells analyze the motion of visual patterns. *Nature neuroscience*, 9(11):1421–1431, 2006. Publisher: Nature Publishing Group.

[46] Shinji Nishimoto and Jack L. Gallant. A three-dimensional spatiotemporal receptive field model explains responses of area MT neurons to naturalistic movies. *Journal of Neuroscience*, 31(41):14551–14564, 2011. Publisher: Soc Neuroscience.

[47] Patrick J. Mineault, Farhan A. Khawaja, Daniel A. Butts, and Christopher C. Pack. Hierarchical processing of complex motion along the primate dorsal visual pathway. *Proceedings of the National Academy of Sciences*, 109(16):E972–E980, 2012. Publisher: National Acad Sciences.

[48] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles. In *Field and Service Robotics*, 2017. _eprint: arXiv:1705.05065.

[49] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.451701; this version posted October 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[50] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[51] Michael C.-K. Wu, Stephen V. David, and Jack L. Gallant. Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.*, 29:477–505, 2006. Publisher: Annual Reviews.

[52] Wyeth Bair and Christof Koch. Temporal precision of spike trains in extrastriate cortex of the behaving macaque monkey. *Neural computation*, 8(6):1185–1202, 1996. Publisher: MIT Press.

[53] Maneesh Sahani and Jennifer F. Linden. How linear are auditory cortical responses? *Advances in neural information processing systems*, pages 125–132, 2003. Publisher: MIT; 1998.

[54] Tengda Han, Weidi Xie, and Andrew Zisserman. Video Representation Learning by Dense Predictive Coding. *arXiv:1909.04656 [cs]*, September 2019. arXiv: 1909.04656.

[55] M. S. Livingstone. Mechanisms of direction selectivity in macaque V1. *Neuron*, 20(3):509–526, March 1998.

[56] Nicole C. Rust, Odelia Schwartz, J. Anthony Movshon, and Eero Simoncelli. Spike-triggered characterization of excitatory and suppressive stimulus dimensions in monkey V1. *Neurocomputing*, 58:793–799, 2004. Publisher: Elsevier.

[57] Russell L. De Valois, E. William Yund, and Norva Hepler. The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, 22(5):531–544, January 1982.

[58] J. Anthony Movshon and William T. Newsome. Visual response properties of striate cortical neurons projecting to area MT in macaque monkeys. *Journal of Neuroscience*, 16(23):7733–7741, 1996. Publisher: Soc Neuroscience.

[59] J. A. Perrone and A. Thiele. Speed skills: measuring the visual speed analyzing properties of primate MT neurons. *Nat Neurosci*, 4(5):526–32, May 2001.

[60] Mark M. Churchland, Nicholas J. Priebe, and Stephen G. Lisberger. Comparison of the Spatial Limits on Direction Selectivity in Visual Areas MT and V1. *Journal of Neurophysiology*, 93(3):1235–1245, March 2005. Publisher: American Physiological Society.

[61] Michael SA Graziano and Charles G. Gross. The representation of extrapersonal space: A possible role for bimodal, visual-tactile neurons. *The cognitive neurosciences*, pages 1021–1034, 1995.

[62] Ian Nauhaus and Dario L. Ringach. Precise alignment of micromachined electrode arrays with V1 functional maps. *Journal of neurophysiology*, 97(5):3781–3789, 2007. Publisher: American Physiological Society.

[63] Dario Ringach and Ian Nauhaus. Single- and multi-unit recordings from monkey primary visual cortex.

[64] William E. Vinje and Jack L. Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000. Publisher: American Association for the Advancement of Science.

[65] Stephen V. David, William E. Vinje, and Jack L. Gallant. Natural stimulus statistics alter the receptive field structure of v1 neurons. *Journal of Neuroscience*, 24(31):6991–7006, 2004. Publisher: Soc Neuroscience.

[66] Stephen V David, William E Vinje, and Jack L Gallant. Single electrode recordings from primary visual cortex.

[67] Y Cui, DL Liu, FA Khawaja, CC Pack, and DA Butts. Spiking activity in area MT of awake adult macaques in response to complex motion features.

[68] S Nishimoto and JL Gallant. Extracellular recordings from area MT of awake macaques in response to naturalistic movies.

[69] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

[70] Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, and Alexander S. Ecker. Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS computational biology*, 15(4):e1006897, 2019. Publisher: Public Library of Science San Francisco, CA USA.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.451701; this version posted October 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[71] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset. *arXiv:1705.06950 [cs]*, May 2017. arXiv: 1705.06950.

[72] Nicolas Pinto, David D. Cox, and James J. DiCarlo. Why is Real-World Visual Object Recognition Hard? *PLOS Computational Biology*, 4(1):e27, January 2008. Publisher: Public Library of Science.

[73] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[74] Carlos R. Ponce, Will Xiao, Peter F. Schade, Till S. Hartmann, Gabriel Kreiman, and Margaret S. Livingstone. Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. *Cell*, 177(4):999–1009.e10, May 2019.

[75] Pouya Bashivan, Kohitij Kar, and James J. DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439), May 2019. Publisher: American Association for the Advancement of Science Section: Research Article.

[76] Ben J. Hardcastle and Holger G. Krapp. Evolution of Biological Image Stabilization. *Current biology: CB*, 26(20):R1010–R1021, October 2016.

[77] Dario L. Ringach, Guillermo Sapiro, and Robert Shapley. A subspace reverse-correlation technique for the study of visual neurons. *Vision research*, 37(17):2455–2464, 1997.

[78] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What Makes a Video a Video: Analyzing Temporal Information in Video Understanding Models and Datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7366–7375, June 2018.

[79] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the Neural Information Processing Systems (NIPS)*, 2014.

[80] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2019.

[81] Jonathan Stroud, David Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. D3d: Distilled 3d networks for video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 625–634, 2020.

[82] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful Image Colorization. In *arXiv:1603.08511 [cs]*, October 2016. arXiv: 1603.08511.

[83] Andrew Owens and Alexei A. Efros. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features. *arXiv:1804.03641 [cs, eess]*, October 2018. arXiv: 1804.03641.

[84] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-Supervised Learning by Cross-Modal Audio-Video Clustering. *arXiv:1911.12667 [cs]*, October 2020. arXiv: 1911.12667.

[85] Trinity B. Crapse and Marc A. Sommer. Corollary discharge across the animal kingdom. *Nature Reviews Neuroscience*, 9(8):587–600, August 2008. Number: 8 Publisher: Nature Publishing Group.

[86] Jessica AF Thompson. *Characterizing and comparing acoustic representations in convolutional neural networks and the human auditory system*. PhD thesis, Universite de Montreal, 2020.

[87] Georg F. Striedter and R. Glenn Northcutt. *Brains Through Time: A Natural History of Vertebrates*. Oxford University Press, 2019.

[88] Frederick Albert Miles and Joshua Wallman. *Visual Motion and its Role in the Stabilization of Gaze*, volume 5. Elsevier Science Limited, 1993.

[89] Richard Le Grand, Catherine J. Mondloch, Daphne Maurer, and Henry P. Brent. Early visual experience and face processing. *Nature*, 410(6831):890–890, 2001. Publisher: Nature Publishing Group.

[90] Uri Hasson, Samuel A. Nastase, and Ariel Goldstein. Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. *Neuron*, 105(3):416–434, February 2020.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.451701; this version posted October 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section 1.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We have been careful not to overclaim (for example, not claiming that this is the only possible explanation of dorsal stream processing), and we describe limitations of our work in the Limitations section.
   (b) Did you describe the limitations of your work? [Yes] Yes, we describe limitations of our work in the Limitations section.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We don't anticipate negative social impacts from this work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] Yes, we have read the guidelines and ensured our paper conforms to them.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] We do not have theoretical results.
   (b) Did you include complete proofs of all theoretical results? [N/A] We do not have theoretical results.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Yes, data, code and instructions are online
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Yes, discussed in the Methods section
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We reported standard error of the mean calculated over cells in Table 3.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We reported our compute in section 3.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Yes, cited in section 3 and table 2.
   (b) Did you mention the license of the assets? [Yes] Yes, listed in section 3 and table 2.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Yes, the data generated in airsim is available online
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We mention that we received permission from the authors where applicable in section 3. In other cases, reuse was allowed by the license under which data was released.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We did not use human data.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not collect human data.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We did not collect human data.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We did not collect human data.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.451701; this version posted October 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A Appendix

Parameter	Value
Environments	AirSimNH, TrapCamera
Dataset size	39645 movies, 112x112x10 frames
Nominal FPS	30
Heading (yaw)	$\mathrm{VonMises}(0, 2.5)$ (rad)
Heading (pitch)	$\mathrm{VonMises}(0, 16)$ (rad)
Head rotation (yaw)	$\mathrm{Normal}(\sigma = \pi/6)$ (rad/s)
Head rotation (pitch)	$\mathrm{Normal}(\sigma = \pi/18)$ (rad/s)
Walking speed	$\mathrm{Uniform}(0, 3)$ (m/s)
Height from ground	$\mathrm{Uniform}(1.4, 2)$ (m)
Step size	0.003
Training epochs	100
Layers	0: 64 7x7x5 conv filters, stride 1x1x1 1: leaky ReLU, 3x3x1 maxpooling, 2x downsampling, batch norm 2: residual block branch 1: 64 filters projected to 32 via 1x1x1 convs branch 2: 32 1x1x1 filters, 8 3x3x1, 32 1x1x1, batch norm 3: residual block, 32 1x1x3 filters, 8 3x3x1, 32 1x1x1, batch norm 4: residual block, 32 1x1x1 filters, 8 3x3x1, 32 1x1x1, batch norm 5: residual block, 32 1x1x3 filters, 8 3x3x1, 32 1x1x1, batch norm
Boosting step size	0.1
Boosting max iterations	100

Table S1: Airsim dataset and training parameters

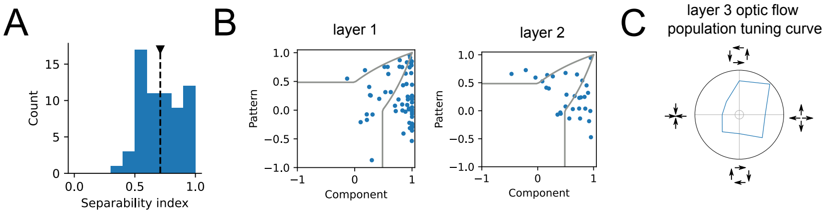

Figure S1: A: Separability index of layer 1. B: pattern index for layers 1 and 2. C: population curves for optic flow in layer 3

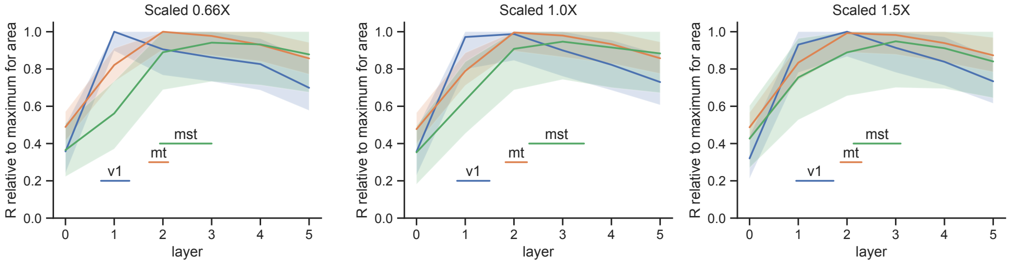

Figure S2: Alignment between layers of DorsalNet and datasets when resizing stimuli. V1 alignment shifts slightly higher as scale is increased, as expected. Alignment is nevertheless broadly similar across different scales.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.451701; this version posted October 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

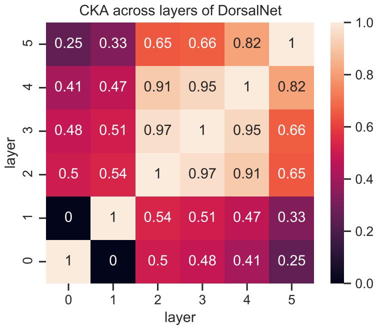

Figure S3: CKA across layers. We used a battery of all stimuli from the airsim dataset to compare representations across layers. We extracted the response of the central pixel of a representation in a given layer and computed alignment between internal representations using centered kernel alignment (CKA) [69]. 0 indicates no alignment, while 1 indicates perfect alignment between layers.

metric area	overall	pitch	yaw	rotation pitch	rotation yaw	speed
v1	-0.39	0.13	0.11	-0.36	-0.54	-0.12
mt	-0.66	-0.05	-0.02	-0.51	-0.64	-0.40
mst	-0.53	0.05	0.05	-0.51	-0.69	-0.13

Table S2: Correlation between loss on heading task and performance on data from different areas across models and layers. Performance on predicting head rotation parameters (rotation pitch and rotation yaw) is most correlated with match to different brain areas.

		V1		MT		MST
	scaling	pvc1	pvc4	mt1	mt2	mst
motionnet	0.66X	.303 (.044)	**.373** (.041)	.221 (.018)	.306 (.016)	.403 (.052)
	1X	.276 (.042)	**.364** (.039)	.238 (.018)	.333 (.016)	**.441** (.053)
	1.5X	**.343** (.040)	**.371** (.039)	**.252** (.019)	.346 (.016)	**.452** (.050)
dorsalnet	0.66X	**.358** (.041)	**.380** (.040)	.245 (.018)	**.388** (.016)	**.460** (.056)
	1X	**.364** (.043)	**.370** (.039)	**.251** (.019)	**.381** (.017)	**.454** (.054)
	1.5X	**.389** (.034)	**.359** (.038)	**.252** (.020)	.370 (.017)	.411 (.052)

Table S3: Relative performance of DorsalNet and MotionNet across different scalings of the input, measured with ridge regression. MotionNet generally benefits from scaling up the videos (1.5X), presumably because of its large second layer receptive fields (27x27). DorsalNet performance is relatively constant across scalings. Table shows normalized pearson correlation (R; see Methods for definition) of different models with different input scaling on different datasets.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.451701; this version posted October 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

		MT		MST
	scaling	mt1	mt2	mst
motionnet	0.66X	-	-	.336 (.050)
	1X	.159 (.012)	.284 (.012)	.361 (.048)
	1.5X	.160 (.011)	.298 (.012)	.385 (.047)
dorsalnet	0.66X	-	-	**.464** (.054)
	1X	**.228** (.017)	**.370** (.016)	**.474** (.051)
	1.5X	**.230** (.017)	.362 (.016)	.434 (.051)

Table S4: DorsalNet quantitatively performs best across the dorsal stream across different scalings, as measured with boosting after downsampling. Table shows normalized pearson correlation (R; see Methods for definition) of different models with different input scaling on different datasets.

		V1		MT		MST
	scaling	pvc1	pvc4	mt1	mt2	mst
motionnet	0.66X	.371 (.048)	**.319** (.036)	.157 (.012)	.258 (.013)	.345 (.049)
	1X	.426 (.050)	**.311** (.034)	.158 (.012)	.271 (.012)	.359 (.048)
	1.5X	**.460** (.051)	.313 (.036)	.158 (.012)	.282 (.012)	.365 (.047)
dorsalnet	0.66X	.471 (.051)	**.355** (.038)	.212 (.016)	**.353** (.016)	**.435** (.052)
	1X	**.491** (.049)	.313 (.039)	**.217** (.016)	**.348** (.016)	**.415** (.055)
	1.5X	**.503** (.051)	**.313** (.034)	.209 (.016)	.328 (.016)	.356 (.052)

Table S5: DorsalNet quantitatively performs best across the dorsal stream across different scalings, as measured with boosting after subsampling. Table shows normalized pearson correlation (R; see Methods for definition) of different models with different input scaling on different datasets.

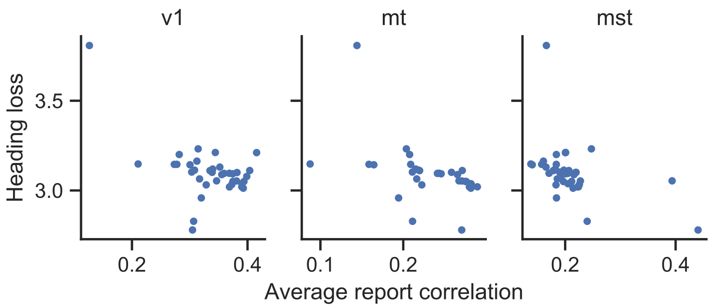

Figure S4: Correlation between heading loss and performance on dorsal stream datasets across networks and layers. Networks and layers which perform better at heading discrimination tend to better match the dorsal stream.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.451701; this version posted October 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

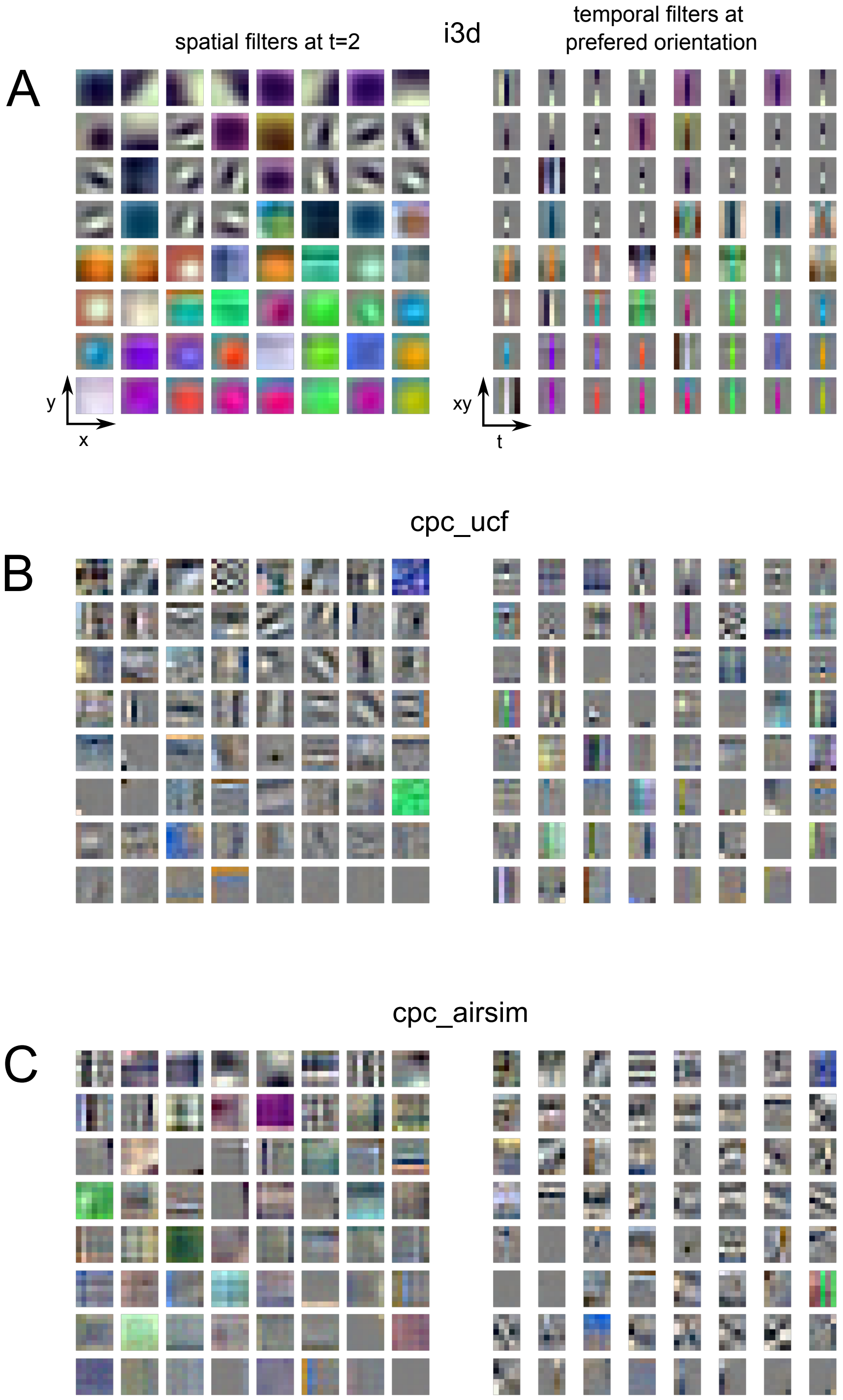

Figure S5: First layer filters for alternative networks. i3d and CPC on UCF learn orientation selectivity but not direction selectivity. CPC on Airsim learns both.