Neural Networks 154 (2022) 383–396

Contents lists available at ScienceDirect

Neural Networks

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

Estimating heading from optic flow: Comparing deep learning network and human performance[☆]

Natalie Maus[a], Oliver W. Layton[b,*]

[a] Department of Computer Science, University of Pennsylvania, Philadelphia, 19104, PA, USA
[b] Department of Computer Science, Colby College, Waterville, 04901, ME, USA

ARTICLE INFO

Article history:
Received 22 November 2021
Received in revised form 17 June 2022
Accepted 7 July 2022
Available online 25 July 2022

Keywords:
Optic flow
Deep learning
Heading
Vision
Self-motion

ABSTRACT

Convolutional neural networks (CNNs) have made significant advances over the past decade with visual recognition, matching or exceeding human performance on certain tasks. Visual recognition is subserved by the ventral stream of the visual system, which, remarkably, CNNs also effectively model. Inspired by this connection, we investigated the extent to which CNNs account for human heading perception, an important function of the complementary dorsal stream. Heading refers to the direction of movement during self-motion, which humans judge with high degrees of accuracy from the streaming pattern of motion on the eye known as optic flow. We examined the accuracy with which CNNs estimate heading from optic flow in a range of situations in which human heading perception has been well studied. These scenarios include heading estimation from sparse optic flow, in the presence of moving objects, and in the presence of rotation. We assessed performance under controlled conditions wherein self-motion was simulated through minimal or realistic scenes. We found that the CNN did not capture the accuracy of heading perception. The addition of recurrent processing to the network, however, closed the gap in performance with humans substantially in many situations. Our work highlights important self-motion scenarios in which recurrent processing supports heading estimation that approaches human-like accuracy.

1. Introduction

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) catalyzed historic developments in deep convolutional neural networks (CNNs) (Lecun et al., 2015; LeCun et al., 1995; Russakovsky et al., 2015). The challenge benchmarked the accuracy with which competing algorithms classified 1000 categories of natural images in the large ImageNet dataset (Deng et al., 2009). CNNs won the ILSVRC in 2012 starting with AlexNet (Krizhevsky et al., 2012) and the error in image classification produced by the top performing network reached a milestone level in 2015: it matched the accuracy of human image recognition (Russakovsky et al., 2015). CNNs have since been shown to demonstrate human-like or better levels of performance on other image recognition and segmentation tasks (Cireşan et al., 2012; He et al., 2015; Lee et al., 2017; Phillips et al., 2018).

The similarity in performance between humans and CNNs at image recognition may not be coincidental: CNNs share many components with hierarchical neural networks that model the primate ventral stream, part of the visual system that is involved with object recognition. For example, Fukushima (1980)'s pioneering hierarchical model Neocognitron contains the rectified-linear (ReLU) activation function and interleaved convolutional and max-pooling layers — integral components of CNNs like AlexNet (Krizhevsky et al., 2012). From a neuroscience perspective, ReLU prevents negative firing rates and introduces nonlinearities that help capture important properties of neurons (Wu et al., 2006). Convolution enables weight sharing and models biologically-plausible local connectivity among neurons. The max pooling operation gives rise to invariant neural responses to objects that occupy different positions within the visual scene (Fukushima, 1980; Riesenhuber & Poggio, 1999). CNNs with these building blocks have been shown to develop similar representations and effectively predict the responses of ventral stream neurons involved with object recognition (Guclu & Van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Lindsay, 2021; Yamins & Dicarlo, 2016; Yamins et al., 2013, 2014).

Despite their successes at image recognition and as a model of the primate ventral stream, CNNs have not yet been used to examine the function of the dorsal stream, the complementary visual pathway involved with motion. One primary function of the dorsal stream is believed to be the perception of self-motion as we move through our environment (Britten, 2008). Indeed, causal

[☆] Funding: This work was supported by ONR N00014-18-1-2283 (OWL).

[*] Corresponding author.
E-mail address: oliver.layton@colby.edu (O.W. Layton).

https://doi.org/10.1016/j.neunet.2022.07.007

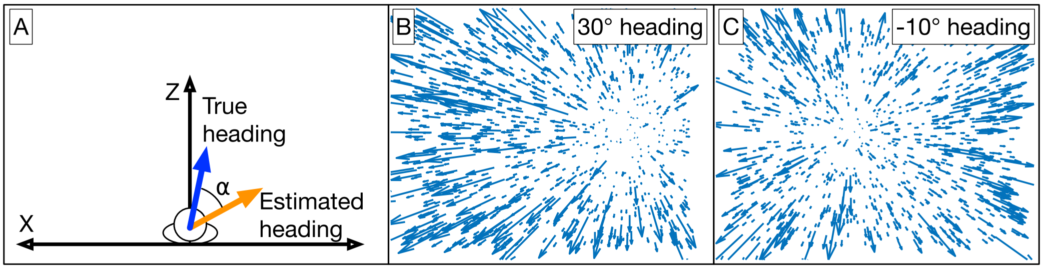

Fig. 1. World coordinate system (a) and sample optic flow fields depicting simulated self-motion through a 3D dot cloud along 30° (b) and −10° headings (c). The eye of the observer begins centered at the origin of the world coordinate system (X, Y, Z) and gaze is parallel to the positive Z axis. Positive (negative) heading angles define translation to the right (left) to the straight-ahead (Z axis). Heading error (α) is defined as the difference between the estimated heading and the true heading (*predicted* − *true*).

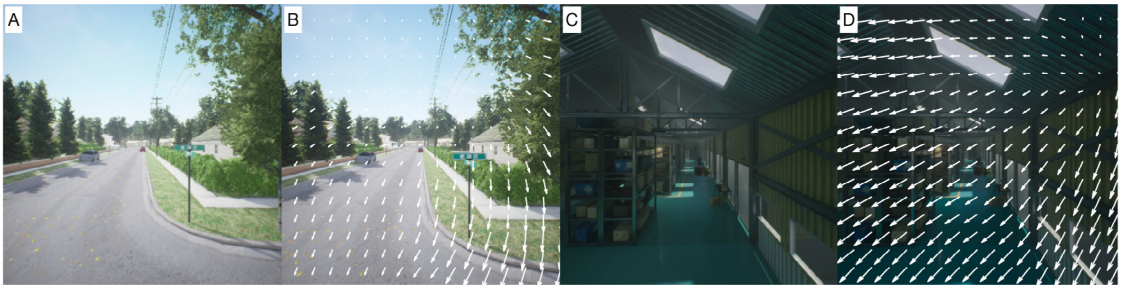

Fig. 2. Sample frames of Neighborhood (a) and Warehouse (c) datasets and corresponding optic flow (b, d).

connections have been established between neurons in the dorsal medial superior area (MSTd) and the observer's direction of travel (heading) (Gu et al., 2012). MSTd neurons demonstrate sensitivity to the expansive patterns of motion (optic flow) that arise while moving along different heading directions (Fig. 1b–c) (Britten, 2008; Duffy & Wurtz, 1995; Gu et al., 2006). The singularity in the optic flow that arises during forward self-motion is known as the focus of expansion (FoE) and coincides with the heading direction (Gibson, 1950). Humans are capable of judging heading with a high degree of accuracy, to within ≈1° for central headings (Foulkes et al., 2013; Warren et al., 1988). The standard stimulus that has been used to characterize human heading perception is simulated self-motion through virtual scenes consisting of sparse arrays of randomly positioned dots. The minimal composition of the stimulus means that the visual system must rely on optic flow to estimate heading, thereby mitigating the influence of other cues that may correlate with heading in richer scenes. Fig. 1b–c show examples of the instantaneous optic flow fields when dots are positioned in a 3D cloud in front of the observer. Remarkably, thresholds with which humans discriminate heading are only ≈3° in scenes with two dots (Warren et al., 1988).

Given how well CNNs match human performance at certain image recognition tasks, we explored the extent to which CNNs effectively model human heading perception from optic flow. We compared the heading estimates produced by CNNs with human heading judgments on a number of scenarios from human studies wherein the task was to judge heading from optic flow. These scenarios include heading estimation from sparse optic flow, in the presence of independently moving objects, and with rotation. Given the sufficiency of optic flow for driving MSTd signals and human heading perception, optic flow vectors served as the input to the CNNs. This constrained the CNNs to learn based on optic flow and is consistent with many biologically inspired models of heading estimation in MSTd (Beyeler et al., 2016; Cameron et al., 1998; Elder et al., 2009; Lappe & Rauschecker, 1993; Layton et al., 2012; Royden, 2002; Warren & Saunders, 1995). We considered optic flow from scenes consisting of randomly positioned dots (Fig. 1), similar to those used in studies of human heading perception, and from more realistic outdoor and warehouse environments generated with the Unreal video game engine (Fig. 2).

Because human heading perception depends on the temporal evolution of optic flow (Layton & Fajen, 2016c) and CNNs do not account for this, we additionally consider the heading estimation performance of a recurrent neural network (RNN). The CNN and RNN differ only by a long short-term memory (LSTM) layer to facilitate their comparison. We hypothesized that the RNN would produce superior accuracy in scenarios wherein events unfold over time, such as in the presence of moving objects. In these cases, recurrent connections that integrate information over time may improve the robustness of heading estimates, as has been shown to occur in biologically inspired models of MSTd (Layton & Fajen, 2016a).

2. Methods

2.1. Optic flow datasets

Table 1 specifies the video datasets of simulated self-motion through a 3D dot cloud. Each video consists of 45 frames of optic flow (1.5 s duration at 30 frames-per-second). We designed the conditions to facilitate comparison to data from studies of human heading perception. For consistency with these studies, we distributed heading angles along the horizontal azimuthal (X) axis and fixed the elevation to the vertical midline ($Y = 0$). We considered scenarios wherein heading azimuth and elevation both vary in separate simulation experiments involving the Neighborhood and Warehouse datasets (described below).

On each frame of video, we computed the optic flow using a pinhole camera model (Raudies & Neumann, 2012) and standard analytic equations (Longuet-Higgins & Prazdny, 1980). Table 2 summarizes the parameters that specify the simulated observer and 3D dot cloud environment. We clipped and replaced dots that exited the field of view or valid depth range to ensure that the same number of dots always remained visible.

Table 1
Datasets of simulated self-motion through a 3D cloud of randomly positioned dots. *Uniform* indicates sampling from a uniform random distribution with the specified endpoints. Negative (positive) object speed indicates leftward (rightward) movement.

Dataset	Description	Independent variables
Constant heading	525 videos of simulated self-motion along a constant heading direction	Heading azimuth: *Uniform* $(-60°, 60°)$
Approaching moving object	525 videos of simulated self-motion in the presence of a large 20 × 20 m square object moving at 12 m/s. The object approached the observer in depth over time.	Heading azimuth: *Uniform* $(-15°, 15°)$ Object path angle: *Uniform* $(-25°, 25°)$ Object horizontal offset (m): *Uniform* $(-2, 2)$
Fixed-depth moving object	18 videos of simulated self-motion in the presence of a square independently moving object that maintained a fixed depth with respect to the observer, moving only either leftward or rightward in the observer's reference frame.	Horizontal object position (m): $[-25, -18.75, \ldots, 25]$ Object speed (m/s): ±3
Simulated rotation	540 videos of simulated self-motion with horizontal rotation (yaw; about Y axis)	Heading azimuth: *Uniform* $(-4°, 4°)$ Yaw rotation rate (°/s): $[-5.7, -4.3, \ldots, 5.7]$

Table 2
Parameters specifying self-motion through 3D dot cloud environment.

Parameter	Value
Spatial resolution	64 × 64 pixels
Translational speed	3 m/s
Camera focal length	1.74 cm
Field of view	90°
Eye height	1.61 m
Dots in 3D dot cloud	2000
Dots in moving object[a]	2000
Observer-relative depth range of dots	1–50 m

[a] Object only present in moving object datasets.

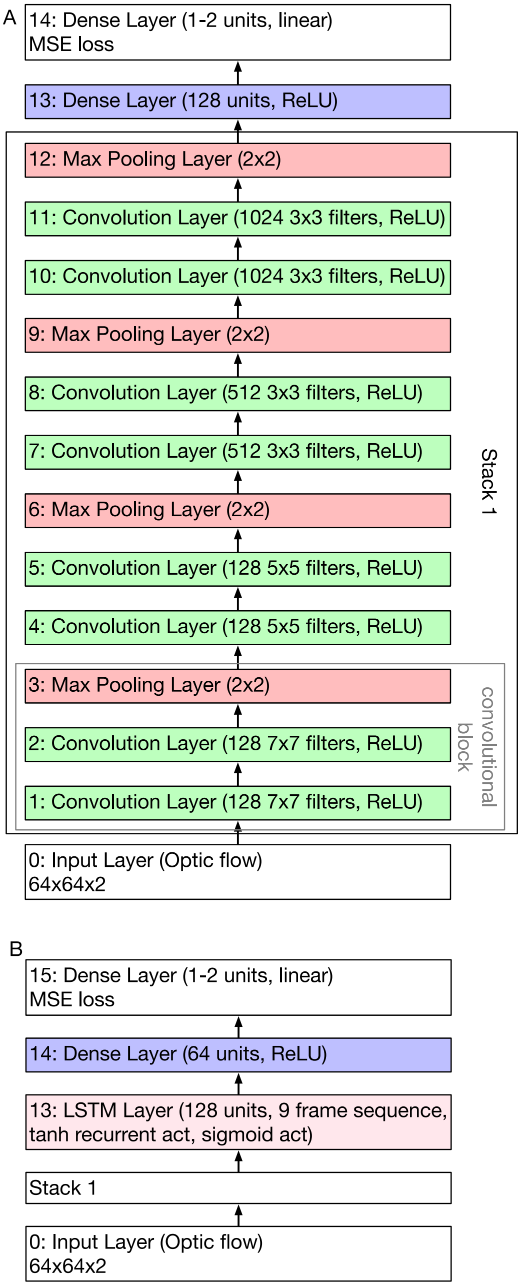
A
14: Dense Layer (1-2 units, linear)
MSE loss
13: Dense Layer (128 units, ReLU)
12: Max Pooling Layer (2x2)
11: Convolution Layer (1024 3x3 filters, ReLU)
10: Convolution Layer (1024 3x3 filters, ReLU)
9: Max Pooling Layer (2x2)
8: Convolution Layer (512 3x3 filters, ReLU)
7: Convolution Layer (512 3x3 filters, ReLU)
6: Max Pooling Layer (2x2)
5: Convolution Layer (128 5x5 filters, ReLU)
4: Convolution Layer (128 5x5 filters, ReLU)
3: Max Pooling Layer (2x2)
2: Convolution Layer (128 7x7 filters, ReLU)
1: Convolution Layer (128 7x7 filters, ReLU)
Stack 1
convolutional block
0: Input Layer (Optic flow)
64x64x2
B
15: Dense Layer (1-2 units, linear)
MSE loss
14: Dense Layer (64 units, ReLU)
13: LSTM Layer (128 units, 9 frame sequence, tanh recurrent act, sigmoid act)
Stack 1
0: Input Layer (Optic flow)
64x64x2

Fig. 3. Architecture of the simulated CNN (a) and RNN (b).

The Neighborhood and Warehouse datasets were generated with Microsoft AirSim, a simulation environment for drones that renders realistic scenes using the Unreal game engine (Shah et al., 2017). The neighborhood scene consisted of a bright outdoor setting with grass, trees, houses, fences, and streets, and other objects (Fig. 2a). The warehouse environment consisted of a darker indoor scene with tall shelving, skylight reflections, and boxes (Fig. 2c). We made 30 5-s 64 × 64 videos of each scene wherein the camera flew along a piecewise-linear random trajectories that changed 2D heading direction (azimuth and elevation) every 30 frames. The same video dataset was used in Steinmetz et al. (2022). We computed the optic flow from the videos using DeepFlow2 (Weinzaepfel et al., 2013).

2.2. Neural networks

Both CNN and RNN were trained to perform regression on the horizontal heading angle and minimized mean-squared error (MSE) loss. We added a second output neuron in the case of the Neighborhood and Warehouse datasets to estimate both the azimuth and elevation heading components (multi-output network).

2.2.1. Convolutional neural network

Fig. 3a depicts the 14-layer CNN used in our simulation experiments. The architecture resembles AlexNet with interleaved convolution and max-pooling layers, followed by fully connected layers. All convolutions use stride of 1 and 'same' padding. Max pooling operations use strides of 2.

We determined the 14-layer depth of the network through grid search optimization (Table 3). We assumed that the network would consist of one or more "convolutional blocks", defined as two convolutional layers followed by a max pooling layer, connected to a fully connected layer and an output regression layer that minimizes MSE loss. Beginning with a stack of two

Table 3
CNN hyperparameter values used in grid search.

Hyperparameter	Values
Number of convolutional blocks	2, 4, 8[a], 16
Number of filters in convolutional block	32, 64, 128[a], 256, 512[a], 1024[a]
Convolutional filter size	3[a], 4, 5[a], 6, 7[a], 8
Max pooling window size	2[a], 4, 8
Number of fully connected layers before output layer	1[a], 2, 3, 4, 5, 6
Number of units in fully connected layers before output layer	32, 64, 128[a], 256, 512, 1024
Adam learning rate	1e−2, 1e−3, 1e-4[a], 1e−5
Mini-batch size	32, 64, 128, 256[a], 512

[a] Final values selected for one or more network layers.

Table 4
RNN hyperparameter values used in grid search.

Hyperparameter	Values
Number of fully connected layers before output layer	1[a], 2, 3, 4, 5, 6
Number of units in fully connected layers before output layer	32, 64[a], 128, 256, 512, 1024
Number of units in LSTM layer	32, 64, 128[a], 256, 512, 1024

[a] Final values selected for one or more network layers.

convolutional blocks (Fig. 3), we performed a grid search on the remaining network hyperparameters to optimize performance on a validation set from the Constant Heading dataset (see Training protocol section below). We repeated this process with the other numbers of convolutional blocks enumerated in Table 3. The final network depicted in Fig. 3 achieved the minimum MSE on the validation set.

2.2.2. Recurrent neural network

To facilitate comparison with the CNN, the RNN builds on the final architecture of the CNN and its hyperparameters (Fig. 3b). The RNN only differs structurally in the inclusion of a long short-term memory (LSTM) layer in between the final max pooling layer and the first fully connected output layer. We performed another grid search to optimize hyperparameters related to the LSTM and subsequent layers (Table 4).

2.2.3. Training protocol

We trained the neural networks on optic flow features using a supervised learning paradigm. One frame of video constituted two 64 × 64 "images" with the horizontal (dx) and vertical (dy) vector components treated as separate channels. Individual frames served as the samples used to train the CNN (input shape: 64 × 64 × 2). Sequences of 9 contiguous frames represented the samples used to train the RNN (input shape: 9 × 64 × 64 × 2). In the case of the Neighborhood and Warehouse datasets, we ensured that heading remained constant within each RNN sample.

We trained both networks using the Adam optimizer with a learning rate of 1e-4 and mini-batch size of 256 samples, as determined by the grid search described above. We trained the CNN and RNN for 25 and 140 epochs, respectively, at which point the validation loss stopped decreasing. The CNN and RNN took ≈30 mins and ≈80 mins to train, respectively, on an NVIDIA RTX 2080 Ti GPU when implemented with TensorFlow 2.

Unless noted otherwise, we trained the networks with a 80/10/10 split on the Constant Heading dataset: reserving 80% of samples for training, 10% for validation, and 10% for the test set. The networks were trained on the same number of optic flow fields (18 900). It is noteworthy that one CNN sample consists of a single optic flow field, whereas one RNN sample consists of a 9 frame contiguous sequence of optic flow fields within each 45 frame video. This meant that the CNN and RNN training sets consisted of 18 900 and 2100 samples, respectively. The size of the training set was determined by a grid search procedure: we increased the training set in the Constant Heading dataset until the validation loss stopped decreasing. We treated the remaining 3D dot cloud datasets (Table 1) as test sets. For the Neighborhood and Warehouse datasets, we trained the networks on 80% of the Neighborhood videos (80/10/10 split) and evaluated performance on both the remaining Neighborhood samples and Warehouse dataset.

2.2.4. Heading error

When summarizing the accuracy of heading estimates, we may report both MSE and mean absolute error (MAE). MSE was minimized during training; MAE has units of degrees and facilitates comparison with human error.

3. Results

3.1. Accuracy of heading estimation from optic flow

We begin by investigating the accuracy with which three specific instances of the convolutional and recurrent networks estimate heading from optic flow. The first set of networks was trained on optic flow fields composed of random numbers of optic flow vectors in the range [2, 2000] (henceforth "RandCNN", "RandRNN"). This training strategy was intended to emulate the diverse range of optic flow densities experienced by the visual system. The second set of networks was trained on samples containing 200 optic flow vectors, a number that has been used in studies of human heading perception (Foulkes et al., 2013; Warren et al., 1988) (henceforth "200CNN", "200RNN"). The third set of models was trained on samples containing 2000 optic flow vectors and served as a comparison to evaluate the effect of optic flow density from training (henceforth "2000CNN", "2000RNN"). As described in Methods, we trained each of the networks on the same number of optic flow fields. Optic flow samples came from the Constant Heading dataset, which contains videos of simulated forward self-motion along straight paths through a 3D cloud of dots.

Our first test evaluates the accuracy of heading estimates over a wide range of headings ([−60, −60]°) when test samples contain either 200 or 2000 dots, matching the density used to train 200CNN, 200RNN, 2000CNN, and 2000RNN. Fig. 4 shows the error in heading estimates produced by the networks trained on diverse optic flow densities (Fig. 4a–d, top row), networks trained on samples with 200 optic flow vectors (Fig. 4e–f, left box of middle row), and networks trained on samples with 2000 optic flow vectors (Fig. 4g–h, right box of middle row). To facilitate comparison between the neural networks and human

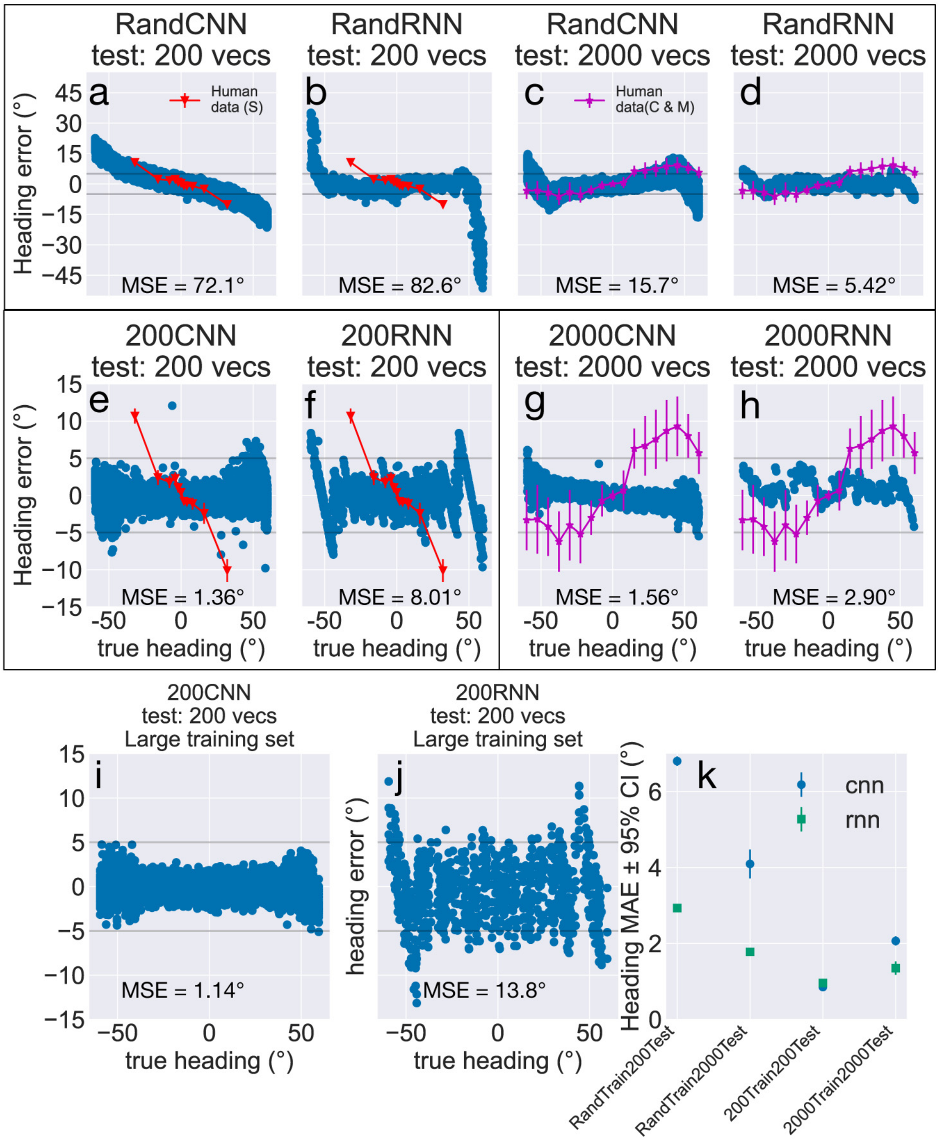

Fig. 4. Heading estimates obtained from three sets of CNNs and RNNs: (a–d) those trained on optic flow fields with varied numbers of vectors (RandCNN, RandRNN), (e–f) 200 vectors (200CNN, 200RNN), (g–h) 2000 vectors (2000CNN, 2000RNN). Optic flow test samples taken from the Constant Heading dataset wherein videos simulate self-motion through a 3D dot cloud contain either 200 or 2000 optic flow vectors. Error computed as *predicted — true*. MSE signifies mean squared error and MAE represents mean absolute error. Error bars show 95% confidence intervals (CIs). Red curve with triangular markers plots human data (mean ± CI) from Sun et al. (2020). Magenta curve with star markers plots human data (mean ± CI) from Cuturi and Macneilage (2013). (i,j) Heading error obtained when 200CNN and 200RNN trained on four times as much data. (k) Summary of MAE parameter estimates for each network and test scenario. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

heading judgments, we include relevant human data in Fig. 4a–h. When test samples contain 200 optic flow vectors, we plot human data from a study by Sun and colleagues wherein humans judged their simulated heading through an virtual 3D cloud of 200 dots (Sun et al., 2020). We show data from a study by Cuturi and MacNeilage in the case of 2000 vector test samples (Cuturi & Macneilage, 2013). Their stimulus contained 16 900 dots instead of 2000, but the comparison is valuable nonetheless because the accuracy of human heading perception appears to stabilize when the number of dots in the visual display exceeds several hundred dots (Foulkes et al., 2013). Additionally, the Cuturi and MacNeilage study assesses 360° heading estimation whereas most other studies focus a limited range of central headings. Fig. 4a, c show that the accuracy of RandCNN estimates are generally consistent with the pattern of human heading errors. RandRNN estimates peripheral headings more accurately (Fig. 4b, d), though 200 vector test samples with large, peripheral headings are clear exceptions (Fig. 4b). In this case errors increase considerably compared to more central headings, where error does not generally exceed ≈ ±5°. For example, when heading is ≈ −60° errors reach ≈ 40° and when heading is ≈ 60° errors reach ≈ −50°. This pattern indicates substantial bias toward the center (e.g. estimate a −60° heading as −20°). The center bias likely occurs because the FoE, the most informative region of the optic flow field about heading (Crowell & Banks, 1996), is not visible within the observer's field of view (Yumurtaci & Layton, 2021). Under these conditions heading estimation is challenging because all the motion vectors appear nearly parallel. Interestingly, humans also demonstrate bias in their judgments toward the center (e.g. Fig. 4a, red curve) (Sun et al., 2020; Yumurtaci & Layton, 2021).

In most cases, network estimates of peripheral headings exhibit increased variability compared to central headings (Fig. 4a–h), a phenomenon that also arises in human heading judgments

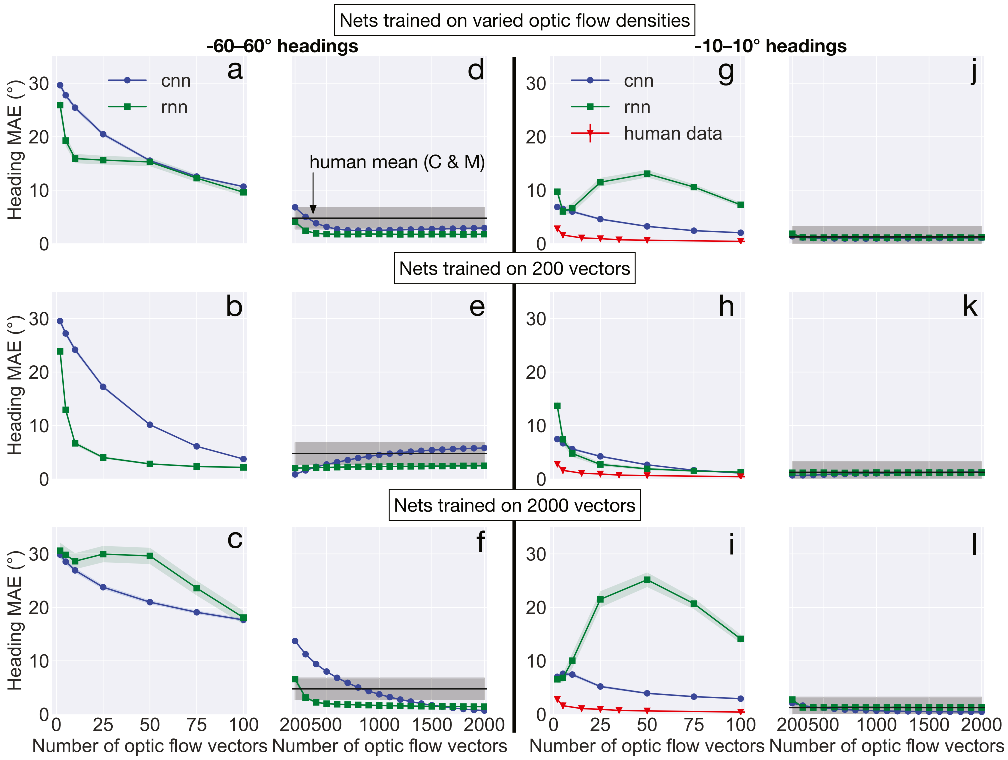

Fig. 5. Accuracy of heading estimates on optic flow density generalization test. (a–f) Model accuracy computed over (−60, 60)° headings when test samples contain the indicated number of optic flow vectors (2, 2000). (g–l) Model accuracy computed over central (−10, 10)° headings. (top row) Accuracy of networks trained on varied optic flow densities. (middle row) Accuracy of networks trained on samples with 200 vectors. (bottom row) Accuracy of networks trained on samples with 2000 vectors. MAE signifies mean absolute error. Black horizontal line (d–f, j–l) depicts MAE of human judgments over the applicable heading range computed based on data from Cuturi and Macneilage (2013). Gray bands show the mean standard deviation of human data computed over the applicable range of headings. Error bands on model MAE estimates represent 95% CIs.

(Cuturi & Macneilage, 2013). The networks that were trained on constant density optic flow also exhibit increased variability and error in their estimates of peripheral headings, albeit to a lesser extent than RandCNN and RandRNN (Fig. 4e–h; note difference in *y* axis scale). The constant density networks produce more accurate estimates overall than RandCNN and RandRNN, particularly for peripheral headings (compare MSE values between Fig. 4a–d and e–h).

It is noteworthy that the RNNs exhibit negatively sloped, diagonal structure in their errors, which is most pronounced in the case of 200RNN (Fig. 4f). This pattern indicates that the RNN assigns a single heading to samples that have headings in a similar range. For example, the large leftmost diagonal in Fig. 4f indicates that the network predicts a heading of $\approx -50°$ for optic flow samples with $[-60, -45]°$ headings. One possible explanation is that the RNN encounters too few unique headings during training and the diagonal error patterns reflect coarse learning. We investigated this possibility by fitting CNN200 and RNN200 with an independently generated training set composed of four times as many optic flow samples as the original set. Fig. 4j reveals that the diagonal structure persists in the RNN errors, which suggests that the pattern does not result from too few unique headings in the training set. The overall accuracy with the larger training set was comparable (200CNN: 1.14° vs. 1.36° MSE, 200RNN: 13.8° vs. 8.01° MSE).

Fig. 4k summarizes the accuracy of each model on each test case with 95% confidence intervals (CIs) constructed on the mean absolute error obtained by corresponding CNN and RNN pairs. With the exception of 200CNN and 200RNN, whose MAEs are virtually identical, the RNNs produce more accurate mean heading estimates than the corresponding CNN, regardless of the density of optic flow used in training.

3.2. Sparse optic flow

We tested how well the three sets of networks estimate heading from optic flow fields that have different numbers of vectors. For the networks trained on samples with a constant number of vectors, this is a test of generalization. Our analysis focuses on heading estimation from sparse optic flow, conditions in which humans excel. Fig. 5a–c show that sparse optic flow consisting of 100 or fewer optic flow vectors pose a challenge for the networks, as the precipitous rise in mean error demonstrates. 2000CNN and 2000RNN consistently produce the least accurate mean estimates among networks of their respective types (Fig. 5c). On the other hand, 200RNN garners the most accurate estimates over the widest range of densities (compare Fig. 5b with Fig. 5a and c). The error produced by 200RNN is small compared to the other networks and stable until it rises sharply for samples with 10 or fewer vectors. When the test samples contain only two optic flow vectors, the predictions from all the networks approach ≈ 30° MAE, which represents the heading halfway between the center (0°) and periphery (±60°). That is, the high uncertainty that stems from the sparseness of the optic flow field yields estimates that are no better than an average between the smallest and largest possible headings. RandCNN and RandRNN generate predictions with mean accuracies that fall in between those produced by the networks trained on constant densities. For example, the MAE values that characterize the performance of RandRNN (Fig. 5a) are generally smaller than those of 2000RNN (Fig. 5c), but larger than those of 200RNN (Fig. 5b). Except for the sparsest conditions, however, none of the other networks generates nearly as accurate heading estimates as 200RNN.

Fig. 5d–f reveal that the mean network estimates for dense optic flow are generally at least as accurate as the 4.8° mean

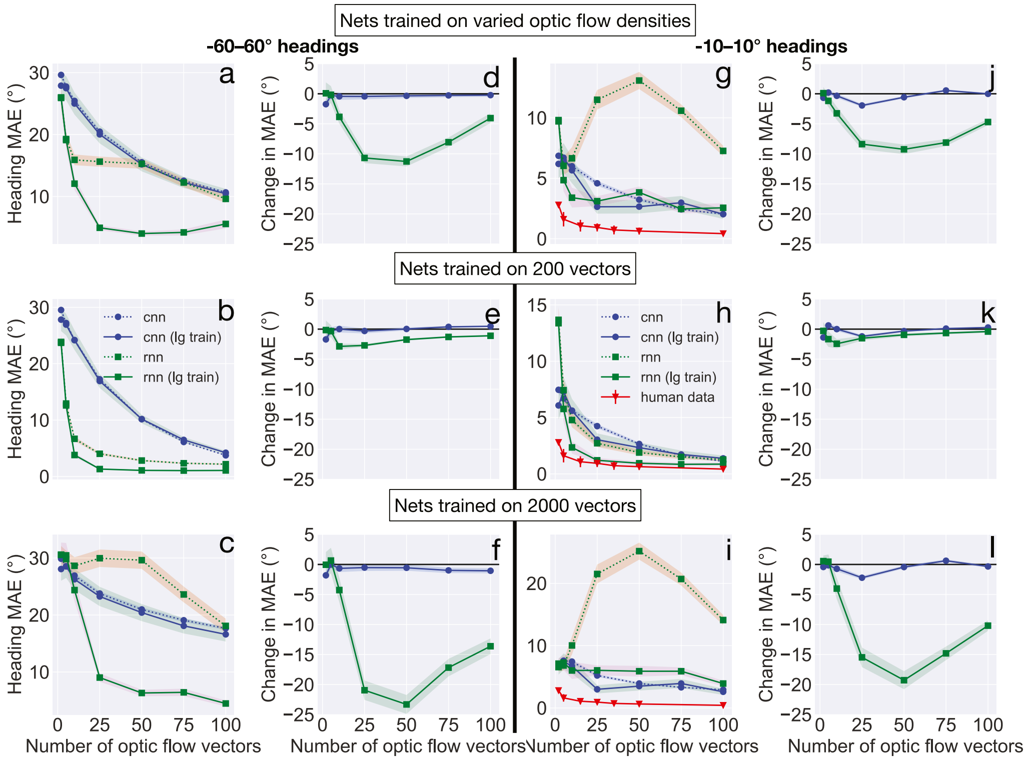

Fig. 6. Effect of additional training on heading generalization across optic flow densities. Format is similar to Fig. 5. (a–c, g–i) Solid curves show MAE values achieved by networks fit with larger training set and dashed curves show values from Fig. 5. Error bands on model MAE estimates represent 95% CIs. (g–i) Red curve with triangular markers shows human data (mean ± CI) from Foulkes et al. (2013). (d–f, j–l) Difference in network MAE values based on training set size (dashed curve–solid curve). Negative values indicate that additional training lowers MAE.

human error obtained over the same range of headings (Cuturi & Macneilage, 2013). The RNNs generally achieve the lowest error over the largest range of optic flow densities. Taken together, these results show that in most cases the RNNs perform at least as well as, and sometimes substantially better than (Fig. 4b, e), the CNNs. The accuracy of 2000RNN on sparse optic flow is an exception (Fig. 5c), however, in this case the mean accuracy for both 2000CNN and 2000RNN is poor.

To compare network performance with human heading judgments from sparse optic flow, we repeated our test while constraining headings to the central 20°, the range used in relevant psychophysical experiments (Foulkes et al., 2013; Warren et al., 1988). None of the networks achieves human levels of accuracy from sparse optic flow fields with 50 or fewer vectors (Fig. 5g–i). On the other hand, the MAE achieved by the networks for dense optic flow did not exceed 2° (Fig. 5j–l), which is comparable to the precision in human judgments of central headings (Crowell & Banks, 1993).

Recall that the goal of the present experiment is to test how well the pretrained networks generalize across different optic flow densities; we did not optimize training for sparse optic flow. Given the large MAE fluctuations among the RNNs (Fig. 5g–i), we explored whether network estimates from sparse optic flow would improve with additional training data. We repeated the density generalization experiment with the networks trained on four times as many optic flow samples. Fig. 6 shows that additional training data improves generalization in the RNNs, but not the CNNs. The RNN error decreases over the full range of headings (Fig. 6d–f) and over central headings (Fig. 6j–l). In the case of RandRNN (Fig. 6d, j) and 2000RNN (Fig. 6f, l) the improvements in RNN mean accuracy are substantial. With additional training data, RandRNN and 2000RNN no longer yield large fluctuations in MAE values (Fig. 6a–c, g–i) and the fit of all the RNNs to the human data improves, though 200RNN remains the best fit (Fig. 6h). The accuracy of the CNNs remains largely unchanged, indicating that the additional training data does not improve generalization to sparse optic flow in these networks. Additional training does not improve the accuracy of any network in the sparsest conditions. Taken together, 200RNN achieves the best accuracy and fit to the human data with sparse optic flow.

3.3. Heading estimation in dynamic scenarios

Next, we evaluated network performance in dynamic scenarios in which properties of the environment changed over time. The aim was to evaluate the consequences of recurrent processing, the key difference between the networks. Recurrent signals allow the RNN to integrate optic flow over time, unlike the CNN that treats the optic flow on successive frames of video as independent. We considered heading estimation in the presence of an independently moving object (IMO), a dynamic scenario that arises frequently in everyday locomotion. IMOs make heading estimation more challenging because they occupy large portions of the visual field and introduce discrepant optic flow (Fig. 7a–b) that is not informative about heading (Layton & Fajen, 2016b; Warren & Saunders, 1995). By integrating sequences of optic flow fields over time, the RNN may be more resilient to the discrepant optic flow and produce more accurate heading estimates than the CNN. However, another possibility is that the RNN may demonstrate increased sensitivity to the moving object since the motion inside the object border resembles self-motion along a discrepant direction (Fig. 7b). We considered two types of objects in the presence of which human heading perception has been studied: moving objects that approach the observer in depth ("approaching objects") and those that only move laterally in the observer's frame of reference ("fixed-depth objects"). These two

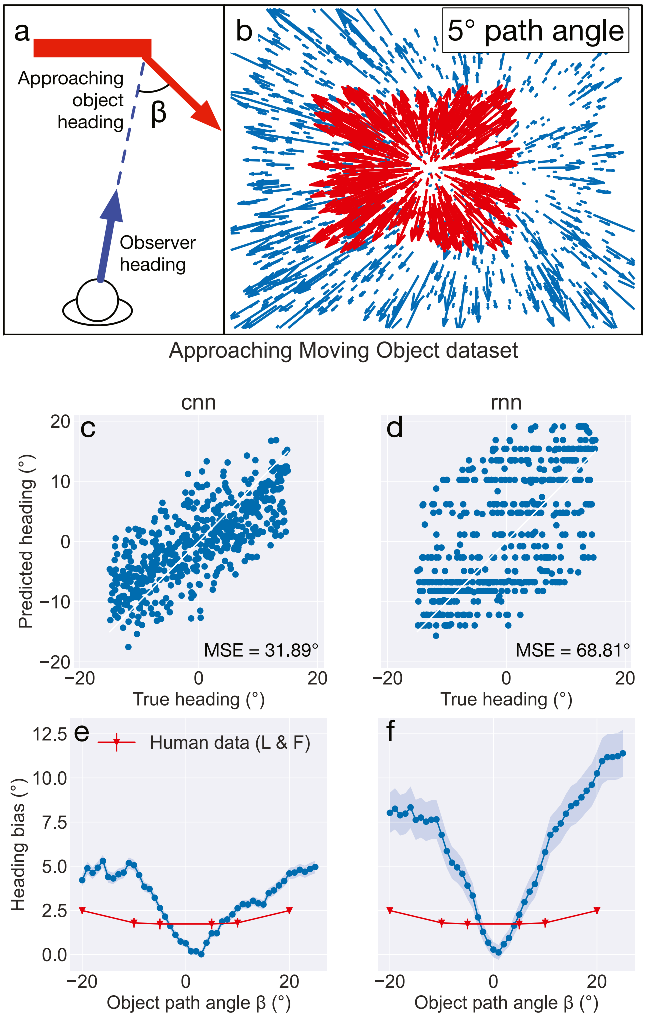

Fig. 7. Heading estimation in the presence of an approaching independently moving object. (a) Approaching objects move toward the observer in depth over time. Object moves along a trajectory defined by its path angle, the angle β between the observer and object headings. Positive (negative) path angles indicate object movement to the right (left) relative to the observer. (b) Optic flow produced with 15° observer heading and 5° path angle. Optic flow associated with the object appears in red. Fewer optic flow vectors shown compared to those used in simulations for visual clarity. (c, d) Heading predictions of the CNN and RNN over the Approaching Moving Object dataset. White curve shows unity line. (e, f) Heading error as a function of object path angle, expressed as object-relative heading bias. Positive bias indicates error toward the object's direction of approach. Red curve with triangular markers depicts human data (mean ± CI) from "After Cross Near" condition of Layton and Fajen (2016a). Error bands on model MAE estimates represent 95% CIs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

types of IMOs are of interest because human heading judgments are biased in opposite directions: the fixed-depth IMO yields bias in the direction of object motion, whereas the bias is toward the direction of approach for the approaching IMO. We examined whether the heading estimates produced by the CNN and RNN would produce similar directions of bias. In these experiments, we continued to use the 3D dot cloud environment and focused on the performance of the CNN and RNN trained on 2000 optic flow vectors.

3.3.1. Approaching moving objects

Starting with the approaching object, we define the trajectory of the object with respect to the heading of the observer with the path angle β (Fig. 7a) (Layton & Fajen, 2016b; Warren & Saunders, 1995). Positive path angles indicate rightward object trajectories relative to the observer's heading, 0° indicates a parallel trajectory to the observer's, and negative path angles indicate leftward object trajectories relative to the observer's heading. The path angle influences the optic flow inside the boundary of the object (Fig. 7b; red arrows). Within this aperture, the motion resembles the optic flow that an observer would experience while moving through a rigid environment without the moving object. Unless the path angle is 0°, the motion vectors inside the object (Fig. 7b; red arrows) do not correspond to the observer's true heading relative to the stationary scene (Fig. 7b; blue arrows), which makes heading estimation more challenging. Although the object and background optic flow vectors are colored different in Fig. 7b, the neural networks do not receive object-background segmentation labels.

Simulations of self-motion in the presence of the approaching object show that both the CNN (Fig. 7c; MSE = 31.89; MAE = 4.61°) and RNN (Fig. 7d; MSE = 68.81; MAE = 6.52°) produce larger heading errors than from the static scene (Fig. 4). The mean absolute error exceeds the ≈1–3° error found in human heading judgments under similar circumstances (Layton & Fajen, 2016b, 2017; Royden & Hildreth, 1996; Warren & Saunders, 1995). To examine this further, we plotted the heading error obtained by the networks for the different object path angles (Fig. 7e–f). For comparison, we plotted human judgments from a study by Layton and Fajen that focused on heading perception in the presence of approaching moving objects (Layton & Fajen, 2016b). We plotted data from the "After Cross Near" object movement condition, given that it had the largest impact on heading judgments. While the networks garner better accuracy for small path angles when the object moves along a trajectory that closely aligns with the observer's heading, the error exceeds that of human judgments at large path angles.

Despite this discrepancy, human and network errors both increase with path angle and exhibit bias toward the object's direction of approach (positive bias; Fig. 7e–f). For objects that approach from the left ($\beta < 0°$) this corresponds to error to the left (negative heading error). For objects that approach from the right ($\beta > 0°$) this corresponds to error to the right (positive heading error).

3.3.2. Including moving objects in training

The heading estimates garnered on the Approaching Object dataset reflect training on the Constant Heading dataset, which did not contain moving objects. We explored whether performance could be improved by varying the proportion of training samples that included an approaching object. We examined the influence of the large 20 × 20 m approaching object used previously (Fig. 7) as well as a smaller 7 × 7 m object.

Fig. 8 shows that including objects of either size in a portion of training samples substantially improves the accuracy of the RNN estimates. To reach the accuracy garnered without the moving object, the RNN required moving objects in 40%–50% of training samples. By contrast, the accuracy of the CNN does not consistently improve, regardless of the proportion of training samples that contain moving objects.

To accurately estimate heading, the networks must learn to discount, ignore, or compensate for the optic flow vectors that belong to the moving object, since they may be inconsistent with the observer's heading. Recall that this is challenging because neither network receives labels that specify which vectors belong to the object. Given the lack of consistent improvement in the accuracy of the CNN when training samples contain the moving object, it is unlikely that the network learns to segment the object from the background based on single independent frames of optic

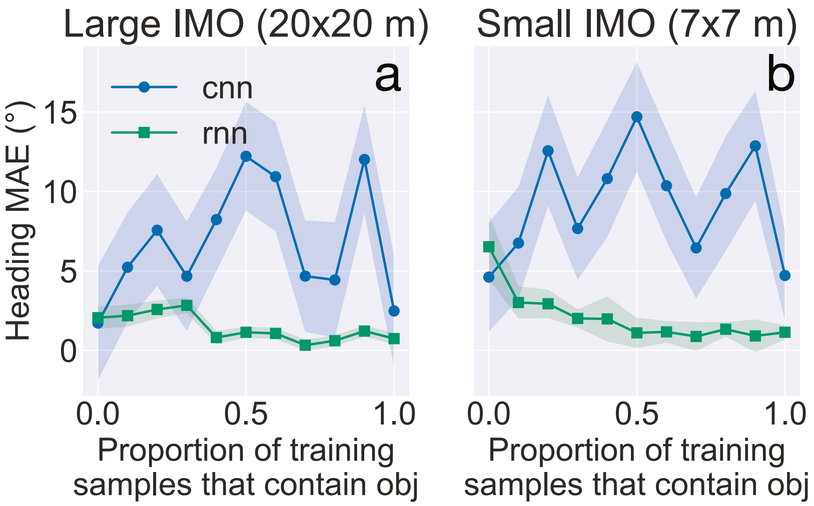

Fig. 8. The accuracy of heading estimates produced by the CNN and RNN on the Approaching Moving Object dataset when different proportions of training set samples contain either a large (a) or small (b) moving object. Error bands on model MAE estimates represent 95% CIs.

flow. On the other hand, the reliable improvement achieved by the RNN suggests that the RNN exploits spatio-temporal information to disambiguate the optic flow field. This information appears to specifically pertain to the moving object, since the RNN produced less accurate estimates than the CNN when the training samples did not contain moving objects (Fig. 7c–d).

3.3.3. Fixed-depth moving objects

To evaluate performance in the presence of the other object type, the fixed-depth object, we once again considered the models trained on the Constant Heading dataset. The fixed-depth object moves leftward or rightward while retreating to occupy a constant size within the optic flow field over time (Fig. 9a–b). Emulating studies that address human heading perception in this scenario (Layton & Fajen, 2017; Royden & Hildreth, 1996), we tested the networks on fixed-depth objects that occupied different horizontal positions (Table 1). We fixed the observer heading to the straight-ahead (0°).

Fig. 9c shows that the CNN demonstrates heading error toward the position of the fixed-depth object: negative error when the object appears to the left, positive error when the object appears to the right. This pattern deviates from the depicted human judgments in several important ways. First, human judgments only reliably demonstrate nonzero error when the object comes close to the object's heading direction (0° in Fig. 9c) (Royden & Hildreth, 1996). The CNN exhibits the opposite pattern: small error when the object appears close to the observer's 0° heading direction, large error when it appears far away. Second, humans exhibit bias in the direction of object motion: error toward the left when the object moves leftward and toward the right when the object moves rightward. Instead, the CNN estimates are biased by object position — leftward (rightward) heading bias when the object is positioned to the left (right), regardless of the object's direction of motion. Third, the maximal error found in human judgments is ≈1°, whereas it reaches 5–6° with the CNN.

The RNN yields a qualitatively different pattern of heading estimates than the CNN (Fig. 9d). The RNN produces a fixed bias of ≈ −2° when the object is far from the heading direction. This represents a ≈ −2° offset from human judgments, which are unbiased in this case (≈ 0°). Similar to humans, the RNN estimates are biased when the moving object is close to the heading direction ($x = 0°$) (Royden & Hildreth, 1996). The differential bias produced by the RNN when the object is close to the heading direction compared to when it is far (≈ 4°), however, is larger than that of humans (≈ 1°). The direction of bias is also not consistent with human judgments: the RNN estimates demonstrate rightward (positive) bias regardless of whether the object moves leftward or rightward. Humans judgments, on the other hand, exhibit bias in the direction of object motion — leftward (negative) for leftward object motion and rightward (positive) for rightward object motion. In sum, neither model captures human heading estimates in the presence of the fixed-depth object, but the RNN does capture the tendency in the human data for the bias to peak when the FoE is obscured by the moving object.

3.4. Simulated rotation

We explored how well the neural networks tolerate the presence of rotation in the optic flow field, which arises during eye movements. This scenario is challenging because rotation causes global distortions to optic flow and may result in a singularity that no longer corresponds to the heading direction. While the networks simulated here do not receive input from the non-visual signals that would normally accompany eye movements, human studies have characterized heading perception when eye movements are simulated within the visual display ("simulated rotation") (Banks et al., 1996; Royden et al., 1994). In this case, rotation arises in the optic flow field due to rotation of the virtual camera while the human observer's eyes remain stationary. Fig. 10a schematizes the virtual camera dynamics during simulated rotation used by Royden et al. (1994). In light of the debate surrounding the extent to which the visual system explicitly compensates for rotation through visual mechanisms when estimating heading (Danz et al., 2020; Lappe et al., 1999; Li & Warren Jr, 2000), we characterized how the networks estimate heading in the presence of simulated rotation when trained on the Constant Heading dataset, which does not contain rotation.

Fig. 10b shows that the heading estimates produced by the CNN and RNN with different rates of yaw rotation exhibit a sigmoidal shape, similar to the pattern of human judgments (Royden et al., 1994). Model heading estimates demonstrate ≈ 10° greater error than humans for the larger rotation rates (Fig. 10b). There is little difference between the CNN and RNN estimates, indicating that recurrent signals and temporal evolution do not influence performance.

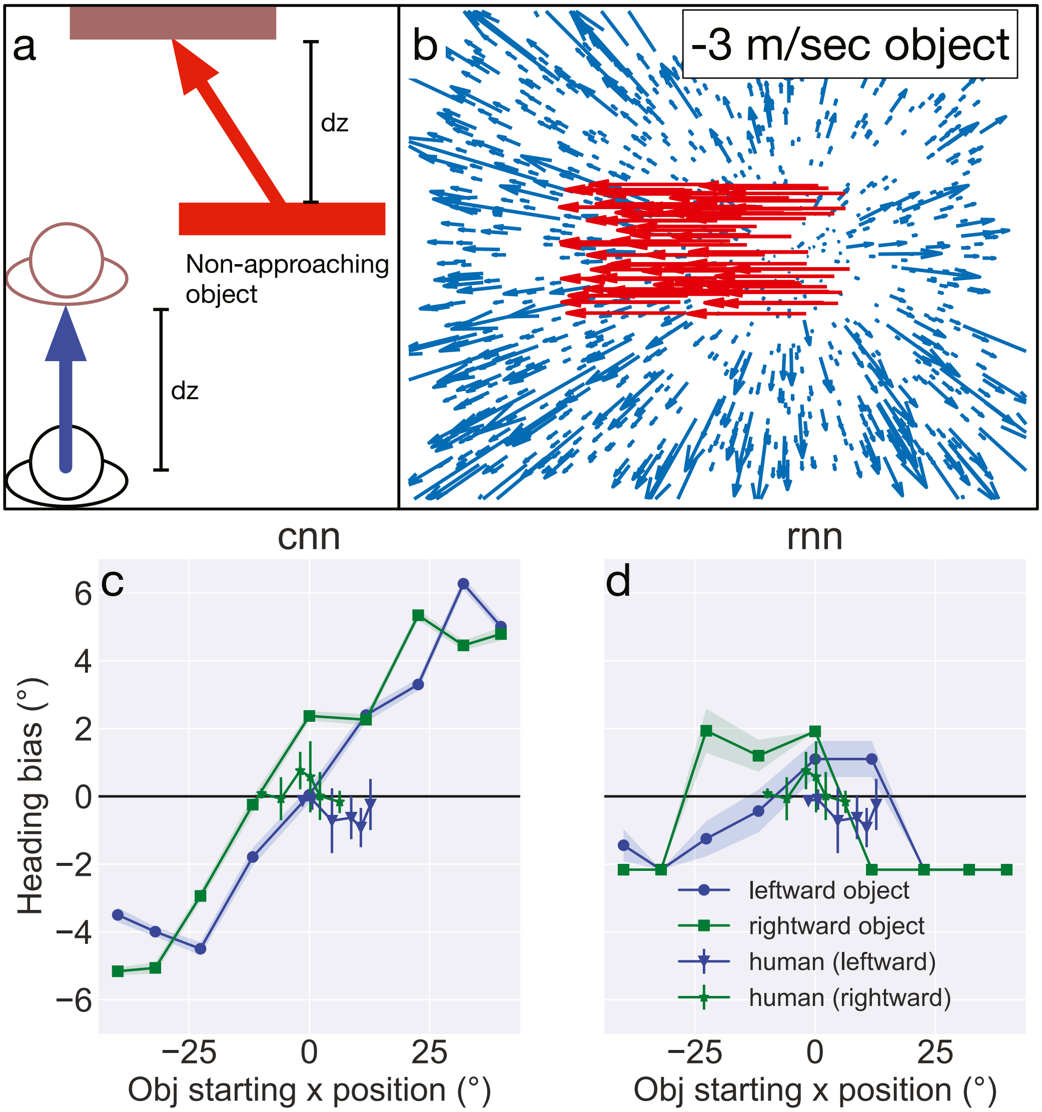

Fig. 9. Heading estimates in the presence of the fixed-depth object. (a) Fixed objects move laterally and retreat to maintain a fixed depth with respect to the observer over time. Positive (negative) speed indicates rightward (leftward) movement. (b) Optic flow produced with 15° observer heading and fixed-depth object moving 3 m/s to the left. Fewer optic flow vectors shown compared to those used in simulations for visual clarity. Bottom row: Error in heading estimates produced by the CNN (c) and RNN (d) on the Fixed Depth Moving Object Dataset. Curves with triangular and star markers show human data (mean ± CI) from Royden and Hildreth (1996) for leftward and rightward moving objects, respectively. Error bands on model bias estimates represent 95% CIs.

3.5. Realistic scenes

In addition to scenarios motivated by studies of human heading perception, we assessed heading estimation from optic flow detected from simulated motion through visually realistic neighborhood and warehouse scenes (Fig. 3). We trained the networks on optic flow from the Neighborhood dataset and evaluated the accuracy of estimates on both the Neighborhood and Warehouse datasets. We used optic flow detected from video rather than the ground truth. Unlike with the 3D dot cloud environment, the virtual camera moved in both azimuth (X) and elevation (Y) directions.

Fig. 11 shows that both networks estimate the azimuth (X) and elevation (Y) heading components with comparable accuracy. Averaging across azimuth and elevation, MAE is 3–5 times higher in the CNN (Neighborhood: 3.55°; Warehouse: 4.81°) and 2 times higher in the RNN (Neighborhood: 2.27°; Warehouse: 2.70°) compared to the Constant Heading dataset (Figs. 4–5). The RNN yields considerably more accurate estimates on the Neighborhood and Warehouse datasets than the CNN (compare Fig. 11a–d with Fig. 11e–h). Both networks perform better on the Neighborhood dataset, reflecting the fact that the training set only contains videos of the neighborhood scene. While the Warehouse Dataset garners less accurate estimates, the accuracy of the RNN estimates decreases modestly (compare Fig. 11e–f with Fig. 11g–h). On the other hand, the accuracy of the CNN estimates decreases by a large extent (compare Fig. 11a–b with Fig. 11c–d).

Fig. 12 summarizes the mean accuracy achieved by each network with 95% confidence intervals. Together, these simulations show that while both networks estimate heading less accurately in the more realistic scenes than the random dot scenes (compare Figs. 12 and 4k), the RNN outperforms the CNN and better generalizes to the novel warehouse environment.

4. Discussion

CNNs demonstrate compelling accuracy on visual classification tasks that under some circumstances matches or exceeds human

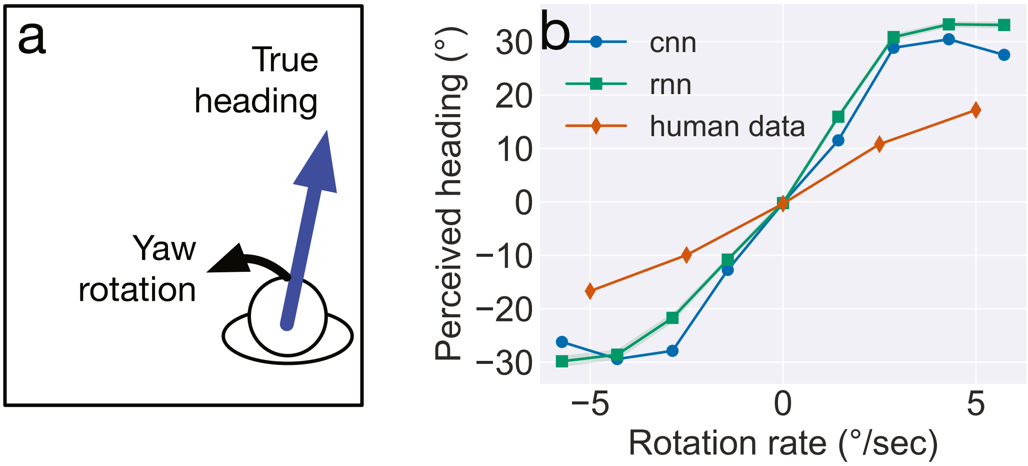

Fig. 10. Heading estimation in the presence of simulated yaw rotation. (a) Simulated self-motion scenario used by Royden et al. (1994). The virtual camera moves along a straight path through a 3D dot cloud while undergoing yaw rotation. (b) CNN and RNN heading estimates on the Simulated Rotation Dataset. Diamond markers indicate judgments of a 0° heading in the simulated rotation scenario made by a representative human subject (Royden et al., 1994). Error bands on model heading estimates represent 95% CIs.

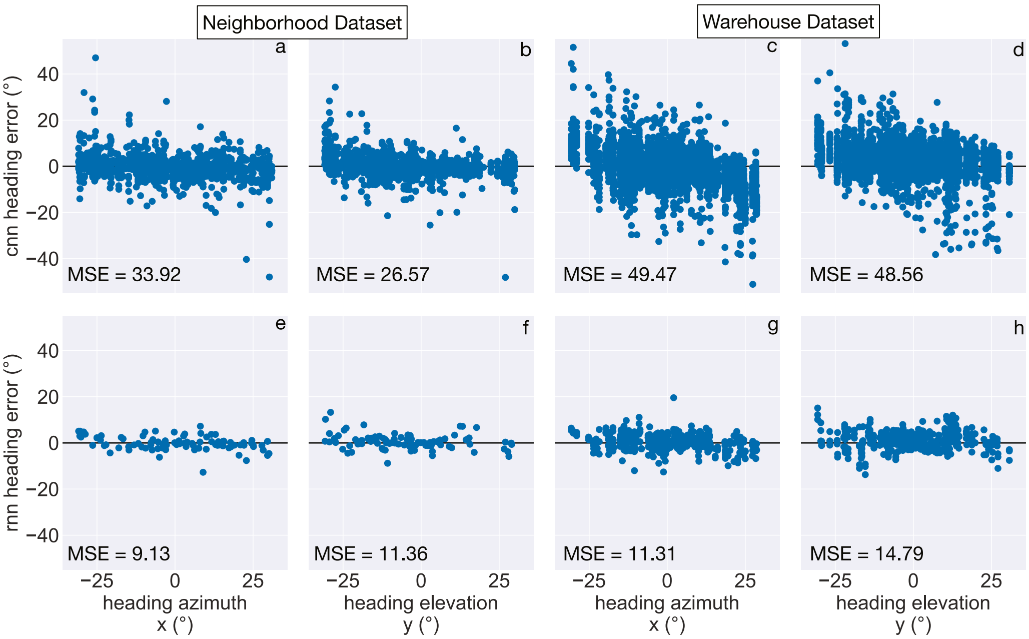

Fig. 11. Heading estimates produced by the CNN and RNN on the Neighborhood and Warehouse datasets. Each network estimates the horizontal (azimuth) and vertical (elevation) component of the heading direction.

performance. This is intriguing because CNNs and biologically-inspired models of the primate ventral stream share convolution, max pooling, and other fundamental computations. Our findings suggest that these canonical mechanisms may not be sufficient for capturing the accuracy of human heading perception from optic flow, a task linked to area MSTd in the complementary primate dorsal stream. Key findings that support this notion include:

- The CNNs did not capture the accuracy of human heading judgments from sparse optic flow (Fig. 5), even when we expanded the training set (Fig. 6).
- The CNN did not capture the accuracy of human heading judgments in the presence of approaching moving objects (Fig. 7), even when explicitly trained on samples that contain the moving object (Fig. 8).
- The pattern of heading bias produced by the CNN deviates in direction and magnitude from human judgments in the presence of the fixed-depth moving object (Fig. 9).

This apparent insufficiency is somewhat surprising given that MSTd resides at a similar depth along the visual hierarchy as IT, an area in the ventral stream associated with object recognition that CNNs effectively model (Yamins & Dicarlo, 2016; Yamins et al., 2013, 2014). The insufficiency of the CNN is unlikely to stem from its specific configuration, given that the CNN achieved accurate performance on the Constant Heading test set (Figs. 4e, g, i and 5d–f, j–l) and performed at least as well as the RNN (Figs. 7 and 10) on other tasks. One possible explanation for the discrepancy with human performance is that heading estimation in the primate dorsal stream relies on additional mechanisms not captured by the CNN. Recurrent signals that integrate optic flow signals over time may represent one such mechanism.

4.1. The role of recurrent signals in heading estimation

While the addition of recurrent signals did not allow the network to fully account for human heading perception, the

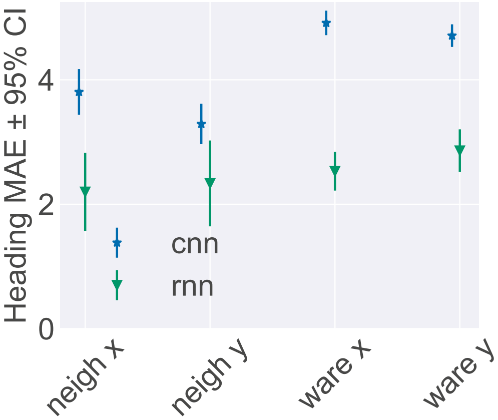

Fig. 12. Comparison of mean accuracy achieved by the networks on the Neighborhood and Warehouse datasets. Error bars on MAE estimates represent 95% CIs.

RNN came closer in important ways. These improvements may illuminate circumstances in which recurrent integration of optic flow signals may enhance the accuracy of heading estimation.

One such scenario may be heading estimation when the density of optic flow differs from that used in training (Fig. 5). The RNN substantially outperformed the CNN, achieving accurate, stable, human-like estimates over most densities tested. The gap in accuracy is particularly large on very sparse optic flow fields (50 vectors or fewer; Fig. 5b). This suggests that integrating optic flow sequences supports accurate heading estimates when limited visual information is present at any one time. It is noteworthy that the performance advantage of the RNN and its consistency with human data only holds for the network trained on the sparser optic flow patterns — the CNN and RNN both performed poorly on sparse optic flow when trained on dense optic flow. This finding is compatible with the success of biological models that rely on sparse representations in accounting for neurophysiological properties of primate area MSTd and human heading data (Beyeler et al., 2016; Layton et al., 2019).

We found that recurrent signals support human or better levels of heading accuracy in the presence of the approaching moving object, but only when the object appears in at least a small proportion of training samples. This is likely because the optic flow inside the contours of the approaching object resembles self-motion along a discrepant heading direction (red vectors; Fig. 7b). Without a teaching signal to indicate otherwise, the RNN possesses no mechanism to discount motion that conflicts with the observer's heading direction. Integrating optic flow signals over time appears to play a crucial role in learning to disregard the conflicting object motion, given that the inclusion of the object in training did not improve the accuracy of the CNN estimates. This agrees with the demonstration that recurrent signals stabilize and support accurate heading estimation in the Competitive Dynamics model of MSTd (Layton & Fajen, 2016a).

The presence of recurrent signals coincides with substantial improvements in heading accuracy on the Neighborhood dataset and generalization to the Warehouse dataset (Fig. 11). This suggests that recurrent signals improve tolerance to noise and deviations in the optic flow from the ground truth introduced through the motion estimation algorithm. These ideas warrant further investigation.

4.2. Heading perception in the presence of rotation

The mechanisms underlying heading perception when the optic flow field contains rotation have been subject to longstanding debate. Royden and colleagues demonstrated that human heading judgments are accurate when the rotation is caused by eye movements, but not when the rotation is simulated within the virtual environment (Royden et al., 1994). Neurophysiological studies showed that MSTd neurons demonstrate compensation in their heading signals to rotation introduced by eye movements, suggesting a mechanism that discounts rotation through nonvisual signals (Bradley et al., 1996; Perrone & Krauzlis, 2008; Shenoy et al., 1999). However, recent evidence calls into question the extent to which nonvisual signals compensate for rotation (Danz et al., 2020) and demonstrates that MSTd neurons compensate for rotation visually to varying extents. Indeed, an entirely visual solution may exist given that humans are capable of accurately judging heading from optic flow with simulated rotation when the optic flow is sufficiently dense (Li & Warren Jr, 2000).

In our simulations, the CNN and RNN both yield qualitatively similar sigmoidal heading estimates to humans in the presence of simulated rotation (Fig. 10). It is notable that the networks were both trained on translational optic flow. This may suggest that the heading judgments largely depend on neurons tuned to radial expansion patterns if the networks share mechanisms with the visual system. Clearly, not all MSTd neurons exhibit tuning to expansion (Duffy & Wurtz, 1995; Graziano et al., 1994) and those tuned to other patterns could contribute to a mechanism that visually compensates for rotation. This could explain the gap in accuracy between humans and the networks. The lack of distinction between the accuracy achieved by the CNN and RNN suggests that recurrent signals do not enhance rotation tolerance based on the representations developed by translational optic flow. With our focus on heading estimation, the present study does not address rotation estimation and compensation. Future work should further investigate whether deep networks could offer insight about mechanisms in the visual system that stabilize self-motion perception in the presence of rotation.

4.3. Comparison between network and MSTd receptive field sizes

The grid search that we used to optimize the network structure selected an intermediate depth among the values tested (Table 3). Because the receptive field (RF) size of units in the network increases with depth, this suggests that some RF sizes may better support heading estimation than others. Each signal sent to the first fully connected layer of the simulated networks corresponds to 22.5 × 22.5° portions of the 90 × 90° optic flow field. A circle that circumscribes this square region has a diameter of 32°, which is smaller than the 46° mean receptive field (RF) diameter in MSTd (Tanaka et al., 1986). One possible explanation for the discrepancy is the weight sharing property of convolutional neural networks, whereby each unit's filter weights are updated based on the signals that arise when the filter is centered at different positions across the input. This is unlike biological neurons, which only integrate signals within a fixed RF aperture. Because convolutional units integrate signals over many apertures within the optic flow field, perhaps comparably smaller RF sizes may be sufficient for heading estimation in CNNs.

The 32° diameter from the network is also smaller than the RF sizes that garner the most accurate estimates in simulations of a biologically inspired neural model of MSTd (Yumurtaci & Layton, 2021), which show that larger RF sizes improve the accuracy of heading estimations, but only to a point — RFs larger than 44° only modestly improve the accuracy of heading estimates. In the simulations of Yumurtaci and Layton, RF size had the greatest influence on estimates of peripheral headings (see Fig. 7 of Yumurtaci and Layton (2021)), which raises the possibility that networks with larger RF units may perform better than the presently configured networks when peripheral heading estimation is the priority.

4.4. Input representation

In the present study, the instantaneous optic flow field serves as the input to the deep networks. Each motion vector in this representation is specified by its speed and direction, properties to which neurons in primate brain area MT exhibit tuning (Born & Bradley, 2005). This is noteworthy because MT neurons are believed to provide a primary feedforward input signal to MSTd neurons. It should be noted, however, that the encoding of speed and direction in MT differs from an instantaneous vector field. For example, MT neurons encode speed and direction in a population code and tuning to these properties is not uniform (Yumurtaci & Layton, 2021). Investigating how different front-ends and input representations influence heading estimates would be a natural extension of the present work.

4.5. Comparison to other models

Recently, Mineault and colleagues have trained a 3D ResNet network on naturalistic videos created with AirSim and compared the network properties to areas along the dorsal stream (Mineault et al., 2021). There are several key differences between the present study and theirs. First, their focus is on comparisons with neural data whereas we are concerned with comparisons with human heading judgments. Second, images serve as the input to their network, whereas ours use optic flow. While image-based inputs more closely approximate those from the retina to the visual system, network performance is more difficult to interpret since additional factors likely contribute to the network's predictions. For example, their network appears to extract optic flow signals, but also contrast, color, and other image properties. We used optic flow features to constrain what the networks learn so that we could examine whether optic flow signals coupled with specific network mechanisms are sufficient for accounting for human performance. Third, we simulated networks that are simpler than 3D ResNet. As outlined in Introduction, the basic forward computations in our CNNs have been used in biologically inspired models of the visual system for decades. 3D ResNet contains additional mechanisms, such as residual connections and batch normalization.

Deep convolutional networks have only recently been developed to estimate self-motion from optic flow (Costante & Ciarfuglia, 2017; Costante et al., 2015; Kashyap et al., 2019). Constante and colleagues demonstrated that CNNs outperform state-of-the-art algorithms on the KITTI dataset. Consistent with our findings, the addition of recurrent processing yields substantial accuracy improvements on KITTI compared to this CNN architecture (Pandey et al., 2021; Zhao et al., 2021). We contribute to this collection of work a characterization of how a similar recurrent network architecture performs under controlled, parametrically varying self-motion scenarios, which is not possible with KITTI. Our simulations brings into focus several important conditions that are not well represented in KITTI. As a ground vehicle driving dataset, heading in KITTI does not deviate significantly from the straight-ahead for long stretches. This is noteworthy because we found large disparities in CNN and RNN performance when estimating central and peripheral headings (Fig. 4). Additionally, heading varied both in azimuth and elevation in Neighborhood and Warehouse datasets, more closely resembling drone flight than ground vehicle movement. Clearly, heading estimation is more challenging with the extra degree of freedom. Future work should explore strategies for improving network performance under these conditions.

Declaration of competing interest

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

References

Banks, M., Ehrlich, S., Backus, B., & Crowell, J. (1996). Estimating heading during real and simulated eye movements. *Vision Research*, *36*(3), 431–443.

Beyeler, M., Dutt, N., & Krichmar, J. (2016). 3D visual response properties of MSTd emerge from an efficient, sparse population code. *The Journal of Neuroscience*, *36*(32), 8399–8415.

Born, R., & Bradley, D. (2005). Structure and function of visual area MT. *Annual Review of Neuroscience*, *28*, 157–189.

Bradley, D., Maxwell, M., Andersen, R., Banks, M., & Shenoy, K. (1996). Mechanisms of heading perception in primate visual cortex. *Science*, *273*(5281), 1544–1547.

Britten, K. (2008). Mechanisms of self-motion perception.. *Annual Review of Neuroscience*, *31*, 389–410.

Cameron, S., Grossberg, S., & Guenther, F. (1998). A self-organizing neural network architecture for navigation using optic flow. *Neural Computation*, *10*(2), 313–352.

Cireşan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. (pp. 3642–3649). arXivCVPR, 1202.2745v1.

Costante, G., & Ciarfuglia, T. (2017). LS-VO: Learning dense optical subspace for robust visual odometry estimation. ArXiv, 1709.06019v2.

Costante, G., Mancini, M., Valigi, P., & Ciarfuglia, T. (2015). Exploring representation learning with cnns for frame-to-frame ego-motion estimation. *IEEE Robotics and Automation Letters*, *1*, 18–25.

Crowell, J., & Banks, M. (1993). Perceiving heading with different retinal regions and types of optic flow. *Perception & Psychophysics*, *53*(3), 325–337.

Crowell, J., & Banks, M. (1996). Ideal observer for heading judgments. *Vision Research*, *36*(3), 471–490.

Cuturi, L., & Macneilage, P. (2013). Systematic biases in human heading estimation. *PLoS One*, *8*(2), Article e56862.

Danz, A., Angelaki, D., & Deangelis, G. (2020). The effects of depth cues and vestibular translation signals on the rotation tolerance of heading tuning in macaque area MSTd. *ENeuro*, *7*(6).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Imagenet: A large-scale hierarchical image database* (pp. 248–255). IEEE.

Duffy, C., & Wurtz, R. (1995). Response of monkey MST neurons to optic flow stimuli with shifted centers of motion. *Journal of Neuroscience*, *15*(7), 5192–5208.

Elder, D., Grossberg, S., & Mingolla, E. (2009). A neural model of visually guided steering, obstacle avoidance, and route selection.. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(5), 1501.

Foulkes, A., Rushton, S., & Warren, P. (2013). Flow parsing and heading perception show similar dependence on quality and quantity of optic flow. *Frontiers in Behavioral Neuroscience*, *7*, 49.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*, 193–202.

Gibson, J. J. (1950). *The perception of the visual world*. Houghton Mifflin.

Graziano, M., Andersen, R., & Snowden, R. (1994). Tuning of MST neurons to spiral motions. *The Journal of Neuroscience*, *14*(1), 54–67.

Gu, Y., Deangelis, G., & Angelaki, D. (2012). Causal links between dorsal medial superior temporal area neurons and multisensory heading perception. *Journal of Neuroscience*, *32*(7), 2299–2313.

Gu, Y., Watkins, P., Angelaki, D., & Deangelis, G. (2006). Visual and nonvisual contributions to three-dimensional heading selectivity in the medial superior temporal area.. *The Journal of Neuroscience*, *26*(1), 73–85.

Guclu, U., & Van Gerven, M. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Delving deep into rectifiers: surpassing human-level performance on imagenet classification* (pp. 1026–1034).

Kashyap, H., Fowlkes, C., & Krichmar, J. (2019). Sparse representations for object and ego-motion estimation in dynamic scenes. ArXiv, 1903.03731v1.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), Article e1003915.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

Lappe, M., Bremmer, F., & Van Den Berg, A. (1999). Perception of self-motion from visual flow. *Trends in Cognitive Sciences*, *3*(9), 329–336.

Lappe, M., & Rauschecker, J. (1993). A neural network for the processing of optic flow from ego-motion in man and higher mammals. *Neural Computation*, *5*(3), 374–391.

Layton, O., & Fajen, B. (2016a). Competitive dynamics in MSTd: A mechanism for robust heading perception based on optic flow. *PLoS Computational Biology*, *12*(6), Article e1004942.

Layton, O., & Fajen, B. (2016b). Sources of bias in the perception of heading in the presence of moving objects: Object-based and border-based discrepancies.. *The Journal of Visual*, *16*(1), 9.

Layton, O., & Fajen, B. (2016c). The temporal dynamics of heading perception in the presence of moving objects.. *Journal of Neurophysiology*, *115*(1), 286–300.

Layton, O., & Fajen, B. (2017). Possible role for recurrent interactions between expansion and contraction cells in MSTd during self-motion perception in dynamic environments.. *The Journal of Visual*, *17*(5), 5.

Layton, O., Mingolla, E., & Browning, N. (2012). A motion pooling model of visually guided navigation explains human behavior in the presence of independently moving objects. *Journal of Vision*, *12*(1), 20.

Layton, O., Steinmetz, S., Powell, N., & Fajen, B. (2019). Computational investigation of sparse MT-MSTd connectivity and heading perception. *Journal of Vision*, *19*(10), 237a.

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

LeCun, Y., Jackel, L., Bottou, L., Brunot, A., Cortes, C., Denker, J., Drucker, H., Guyon, I., Muller, U., & Sackinger, E. (1995). Comparison of learning algorithms for handwritten digit recognition. In *Comparison of learning algorithms for handwritten digit recognition, vol. 60* (pp. 53–60). Perth, Australia.

Lee, K., Zung, J., Li, P., Jain, V., & Seung, H. (2017). Superhuman accuracy on the SNEMI3D connectomics challenge. ArXiv, 1706.00120v1.

Li, L., & Warren Jr, W. (2000). Perception of heading during rotation: Sufficiency of dense motion parallax and reference objects. *Vision Research*, *40*(28), 3873–3894.

Lindsay, G. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, *33*(10), 2017–2031.

Longuet-Higgins, H., & Prazdny, K. (1980). The interpretation of a moving retinal image. *Proceedings of the Royal Society of London. Series B*, *208*(1173), 385–397.

Mineault, P., Bakhtiari, S., Richards, B., & Pack, C. (2021). Your head is there to move you around: Goal-driven models of the primate dorsal pathway.

Pandey, T., Pena, D., Byrne, J., & Moloney, D. (2021). Leveraging deep learning for visual odometry using optical flow. *Sensors (Basel)*, *21*(4).

Perrone, J., & Krauzlis, R. (2008). Vector subtraction using visual and extraretinal motion signals: a new look at efference copy and corollary discharge theories.. *The Journal of Visual*, *8*(14), 24.1–2414.

Phillips, P., Yates, A., Hu, Y., Hahn, C., Noyes, E., Jackson, K., Cavazos, J., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., Chen, J., Castillo, C., Chellappa, R., White, D., & O'toole, A. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms.. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(24), 6171–6176.

Raudies, F., & Neumann, H. (2012). A review and evaluation of methods estimating ego-motion. *Computer Vision and Image Understanding*, *116*(5), 606–633.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Royden, C. (2002). Computing heading in the presence of moving objects: a model that uses motion-opponent operators. *Vision Research*, *42*(28), 3043–3058.

Royden, C., Crowell, J., & Banks, M. (1994). Estimating heading during eye movements. *Vision Research*, *34*(23), 3197–3214.

Royden, C., & Hildreth, E. (1996). Human heading judgments in the presence of moving objects.. *Percept Psychophys*, *58*(6), 836–856.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Shah, S., Dey, D., Lovett, C., & Kapoor, A. (2017). AirSim: High-fidelity visual and physical simulation for autonomous vehicles. ArXiv, 1705.05065v2.

Shenoy, K., Bradley, D., & Andersen, R. (1999). Influence of gaze rotation on the visual response of primate mstd neurons. *Journal of Neurophysiology*, *81*(6), 2764–2786.

Steinmetz, S., Layton, O., Powell, N., & Fajen, B. (2022). A dynamic efficient sensory encoding approach to adaptive tuning in neural models of optic flow processing. *Frontiers in Computational Neuroscience*, *16*(844289).

Sun, Q., Zhang, H., Alais, D., & Li, L. (2020). Serial dependence and center bias in heading perception from optic flow.. *The Journal of Visual*, *20*(10), 1.

Tanaka, K., Hikosaka, K., Saito, H.-A., Yukie, M., Fukada, Y., & Iwai, E. (1986). Analysis of local and wide-field movements in the superior temporal visual areas of the macaque monkey. *Journal of Neuroscience*, *6*(1), 134–144.

Warren, W., Morris, M., & Kalish, M. (1988). Perception of translational heading from optical flow.. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(4), 646.

Warren, W., & Saunders, J. (1995). Perceiving heading in the presence of moving objects. *Perception*, *24*(3), 315–331.

Weinzaepfel, P., Revaud, J., Harchaoui, Z., & Schmid, C. (2013). DeepFlow: Large displacement optical flow with deep matching. In *DeepFlow: Large displacement optical flow with deep matching*. IEEE.

Wu, M.-K., David, S., & Gallant, J. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, *29*, 477–505.

Yamins, D., & Dicarlo, J. (2016). Using goal-driven deep learning models to understand sensory cortex.. *Nature Neuroscience*, *19*(3), 356–365.

Yamins, D., Hong, H., Cadieu, C., & DiCarlo, J. J. (2013). Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. In *Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream, vol. 26*. Neural Information Processing Systems Foundation.

Yamins, D., Hong, H., Cadieu, C., Solomon, E., Seibert, D., & Dicarlo, J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Yumurtaci, S., & Layton, O. (2021). Modeling physiological sources of heading bias from optic flow. *ENeuro*, *8*(6).

Zhao, B., Huang, Y., Wei, H., & Hu, X. (2021). Ego-motion estimation using recurrent convolutional neural networks through optical flow learning. *Electronics*, *10*(3), 222.