# How Do Neural Networks Estimate Optical Flow A Neuropsychology-Inspired Study

De Jong, David Benjamin; Paredes-Valles, Federico; De Croon, Guido Cornelis Henricus Eugene

**DOI**
10.1109/TPAMI.2021.3083538

**Publication date**
2022

**Document Version**
Final published version

**Published in**
IEEE Transactions on Pattern Analysis and Machine Intelligence

# How Do Neural Networks Estimate Optical Flow? A Neuropsychology-Inspired Study

David B. de Jong, Federico Paredes-Vallés, and Guido C. H. E. de Croon, *Member, IEEE*

**Abstract**—End-to-end trained convolutional neural networks have led to a breakthrough in optical flow estimation. The most recent advances focus on improving the optical flow estimation by improving the architecture and setting a new benchmark on the publicly available MPI-Sintel dataset. Instead, in this article, we investigate how deep neural networks estimate optical flow. A better understanding of how these networks function is important for (i) assessing their generalization capabilities to unseen inputs, and (ii) suggesting changes to improve their performance. For our investigation, we focus on FlowNetS, as it is the prototype of an encoder-decoder neural network for optical flow estimation. Furthermore, we use a filter identification method that has played a major role in uncovering the motion filters present in animal brains in neuropsychological research. The method shows that the filters in the deepest layer of FlowNetS are sensitive to a variety of motion patterns. Not only do we find translation filters, as demonstrated in animal brains, but thanks to the easier measurements in artificial neural networks, we even unveil dilation, rotation, and occlusion filters. Furthermore, we find similarities in the refinement part of the network and the perceptual filling-in process which occurs in the mammal primary visual cortex.

**Index Terms**—Optical flow, convolutional neural networks, Gabor filters, neuropsychology

## 1 Introduction

Optical flow is a visual cue defined as the projection of the apparent motion of objects in a scene onto the image plane of a biological vision system or a visual sensor [1]. This cue is important for the behavior of animals of varying size [2], ranging from small flying insects [3] to humans [4], as it allows these animals to estimate their ego-motion and to have a better understanding of the visual scene. Optical flow is also important in computer vision and robotics applications for tasks such as object tracking [5] and autonomous navigation [6].

Many algorithms have been introduced to determine optical flow [7], including correlation-based matching methods [8], [9], frequency-based methods [10], [11], and differential methods [12], [13]. Correlation-based matching methods try to maximize the similarity between different intensity regions across multiple frames. Finding the best match then corresponds to finding the shift which maximizes the similarity score. Frequency-based methods exploit either the amplitude or phase component of the complex valued response of a Gabor quadrature filter pair [14] convolved with an image sequence. Lastly, differential methods compute optical flow based on a Taylor expansion of the image signal, subject to the brightness constancy assumption.

All these methods assume that the brightness of a moving pixel remains constant over time and, when applied locally, are subject to the *aperture problem* [15]. Only motion components normal to the orientation of an edge in the image can be resolved.

A global smoothness constraint has been added for differential methods, which assumes that neighboring pixels undergo a similar motion [12]. This has led to *variational* methods that minimize a global energy function consisting of a data and a smoothness term. These methods have played a dominant role for many years due to their high performance. However, a main drawback is that the iterative minimization of the energy function leads to long computation times. Moreover, the brightness constancy assumption is a coarse approximation to reality and thus limits performance [16]. Research has focused on extra energy terms to deal with deviations from the brightness constancy assumption and improve the robustness of global smoothness constraints, leading to slow but steady progress.

As in many other computer vision areas, currently, the best-performing algorithms are trained deep neural networks. Initially, training such networks was challenging due to the lack of ground-truth optical flow data and the excessive human effort required for manual optical flow labeling. *Dosovitskiy et al.* [17] were the first to successfully train deep neural networks to estimate optical flow by using a synthetically generated dataset with optical flow ground truth. Their networks, FlowNetS and FlowNetC, initially performed slightly worse than the state-of-the-art variational methods [18]. However, trained deep neural networks became the new state-of-the-art method for optical flow estimation by subsequent researchers who focused on improving the architecture and training data [19], [20], [21].

Until now, the functioning of these networks is poorly understood. In this article we investigate *how* deep neural

- *The authors are with the Micro Air Vehicle Laboratory (MAVLab), Department of Control and Simulation, Faculty of Aerospace Engineering, Delft University of Technology, 2629 HS Delft, The Netherlands. E-mail: daviddejong@me.com, fedeparedesv@gmail.com, g.c.h.e.decroon@tudelft.nl.*

*Manuscript received 20 Apr. 2020; revised 28 Mar. 2021; accepted 14 May 2021. Date of publication 25 May 2021; date of current version 3 Oct. 2022.*
*(Corresponding author: Federico Paredes-Vallés.)*
*Recommended for acceptance by E. B. Sudderth.*
*Digital Object Identifier no. 10.1109/TPAMI.2021.3083538*

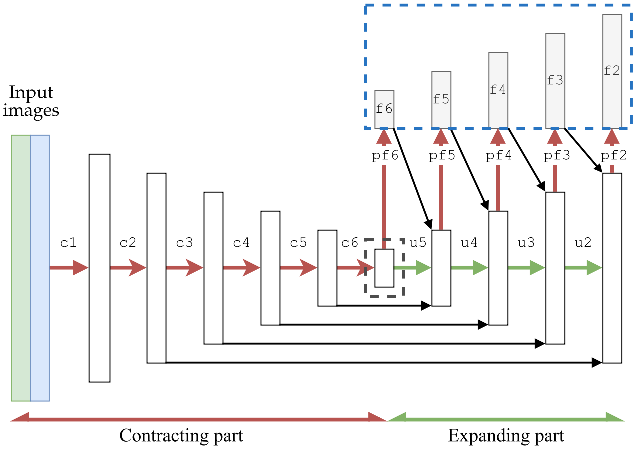


Fig. 1. Schematic of the FlowNetS architecture [17]. The contracting part compresses spatial information through the use of strided convolutions (c), while the expanding part uses upconvolutions (u) for refinement. The predict-flow (pf) layers transform feature map activations into dense flow estimates (f). The feature map corresponding to the output of the c6 layer (gray dashed box) is studied in Sections 4 and 5, while the flow refinement process (blue dashed box) is discussed in Section 6.

networks perform optical flow estimation. Besides satisfying curiosity, there are two main reasons why this is important. First, understanding the method's functioning brings insights into its limits and robustness, for example concerning generalization to test distributions. Second, it may lead to valuable recommendations for improving the performance, for instance, by changing properties of the architecture or training data.

In our analysis of deep optical flow networks, we make use of a method that has helped unveiling the workings of motion-sensitive brain areas in neuropsychology [22]. Specifically, we measure the response of neurons in FlowNetS [17] to stimuli with varying spatiotemporal frequencies and construct a spectral response profile. The input stimuli used are translating plane waves, as this input type proved to be more selective in the frequency domain than moving bars [23]. Based on the earlier findings of Gabor filters [14] in biological vision systems [24], [25] and other learning-based methods [26], [27], we expect to find these filters in FlowNetS as well. Therefore, we fit a Gabor function to the spectral response profile of neurons in the network and study the residual error patterns. We find that the Gabor translational motion filter model is suitable for the majority of the neurons. Additionally, we find neurons sensitive to motion patterns such as dilation, rotation, and occlusion. Interestingly, neurons sensitive to these motion patterns have not been mentioned in neuropsychology. Furthermore, our analysis strongly suggests that the resolution in the temporal frequency domain can be significantly improved if more than two frames would be used as input to the neural network. Lastly, we find that the optical flow refinement process in the decoder part of the network behaves similarly in function to flow refinement in biological vision systems.

The remainder of the article is structured as follows. In Section 2, related work in neuropsychology and deep-learning is discussed. In Section 3, an explanation is given of the architecture of FlowNetS (see Fig. 1). In Section 4, the network's neural responses to translating wave patterns are studied, and compared to translational Gabor filters. Subsequently, Section 5 discusses the response of neurons to dilating and rotating waves. In Section 6, it is studied how FlowNetS resolves the aperture problem. Finally, the results of this work are discussed in Section 7, and conclusions are drawn in Section 8.

## 2 Related Work

### 2.1 Dense Optical Flow Estimation With CNNs

Ever since the pioneering work of *Horn et al.* [12], variational optical flow methods [28] have played a dominant role in optical flow estimation due to their high performance. Most modern variational optical flow estimation pipelines consist of four stages: matching, filtering, interpolation, and variational refinement. Various improvements have been proposed over time to deal with issues such as long-range matching [29] and occlusion [30]. Furthermore, improvements such as dense correspondence matching based on convolution response maps of the reference image with the target image [31], and supervised data-driven interpolation of a sparse optical flow map [32] were also proposed. These last two improvements introduced elements of deep learning into the variational optical flow estimation pipeline.

*Dosovitskiy et al.* [17], however, were the first to introduce a supervised end-to-end trained Convolutional Neural Network (CNN). CNNs have three major advantages when it comes to estimating optical flow. First, CNNs outperform variational optical flow estimation methods in terms of accuracy [19], [20], [21]. Second, the runtime of CNN-based optical flow algorithms, when executed on the appropriate hardware, is significantly lower than variational methods [19]. Third, CNN-based methods can learn from data and can exploit statistical patterns not realized by a human designer. This is an advantage over variational methods which require explicit, and sometimes inaccurate, assumptions on the input. However, CNNs also have three disadvantages. First, the results depend on the quality and size of the training data. Second, CNN-based methods face the risk of overfitting, which is relevant for optical flow estimation because it is difficult to obtain ground truth [18]. Third, there is no guarantee that the trained models will generalize to scenarios not contained in the training dataset. Due to the *"black-box"* nature of the solution, it is difficult to get insight into its workings and limitations.

In [17], *Dosovitskiy et al.* introduced two networks based on the U-net architecture [33]: FlowNetS and FlowNetC. While FlowNetS is an encoder-decoder network consisting of simple convolutions, FlowNetC creates two separate processing streams and combines them in a *correlation-layer*. This layer performs a multiplicative patch comparison between feature maps. Due to the explicit use of a correlation-layer, it is more straightforward to understand the workings of FlowNetC. However, not much is known about the workings of FlowNetS. Inspired by this architecture, *Ranjan et al.* [34] introduced SpyNet, a spatial image pyramid with simple convolutional layers at each pyramid level and a warping operation between pyramid levels. SpyNet's coarse-to-fine approach brings a higher computational and memory efficiency at the cost of a more limited set of perceivable motion types. *Ranjan et al.* also visualized the weights of the first layer of their network and observe that these filters resemble Gabor filters [14], which provided a glimpse into the working principle of this architecture. Finally, *Teney et al.* [35] built a shallow CNN-architecture by integrating domain knowledge, such as invariance to brightness and in-plane rotations. On small motion their architecture performs well, but performance declines on large motion near occlusions. They conclude good occlusion performance requires reasoning over a larger spatiotemporal extent, which their shallow architecture is not able to do.

The generalization performance of CNN-based methods can be evaluated for specific instances by determining the epistemic uncertainty [36]. Indeed, *Ilg et al.* [37] used a modified FlowNetC that produces multiple hypotheses per forward pass, which are then merged to a single distributional flow output. They showed that their network produces highly uncertain flow estimates when optical flow estimation is difficult (shadows, translucency, etc.). Lastly, *Ranjan et al.* [38] highlighted another downside of deep neural networks, which is the ability of adversarial examples to fool neural networks and produce erroneous results. They showed that especially networks using an encoder-decoder architecture are affected, while networks using a spatial pyramid framework are less vulnerable. None of the works above, however, explain how their architecture estimates optical flow.

### 2.2 Receptive Field Mapping

There are two main threads of research to understand what neural networks have learned: attribution and feature visualization. Attribution methods [39], [40] are used to *attribute* filter outputs, like optical flow, to parts of the input by visualizing the gradient. However, it is hard to see where an optical flow estimate comes from. Feature visualization is concerned with understanding what neurons, filters, or layers in a neural network are sensitive to by optimizing the input [41]. The result is usually an image with noisy and visually difficult to interpret high-frequency patterns [42]. Three methods of regularization can be applied to cope with this phenomenon. First, frequency penalization discourages the forming of these patterns. The downside is that this approach also discourages the forming of legitimate high-frequency patterns which are of interest for optical flow estimation. Second, small transformations like scaling, rotation, or translation can be applied in between optimization steps [43]. This approach is also not viable because transformation affects the ground truth of optical flow. Third, priors can be used which can keep the optimized input interpretable. Such approaches typically involve learning a generative model [44] or enforcing priors based on statistics from the training data [45]. This approach is often very complex and it may be unclear what can be attributed to the prior and what can be attributed to what the network has learned.

Due to these reasons, we look at the field of neuropsychology and specifically study what methods researchers have used to determine what stimuli activate neurons in mammalian vision systems and what functions best describe the neural responses. It was shown that Gabor functions [14] best modeled the spatial response of simple cells in the mammal visual cortex [24]. It can be shown that Gabor filters are optimal for simultaneously localizing a signal in the spatial and frequency domain [46], making them ideal for motion estimation. Later, *DeAngelis et al.* [47] examined the spatiotemporal response of cells and their space-time separability. In functional form, space-time separable Gabor filters are frequency-tuned with a stationary Gaussian envelope and space-time inseparable Gabor filters are velocity-tuned with a moving Gaussian envelope [48]. In this work we only consider fitting frequency-tuned Gabor filters, due to their simplicity and the low number of input frames used by the FlowNet architectures.

Two approaches to receptive field mapping in neuropsychology can be discerned: the reverse-correlation approach and the spectral response profile approach. The former presents a rapid random sequence of flashing bars at various imaging locations to the mammal. The spike train emitted by the neuron in the subject is correlated to the sequence in which the stimuli were presented. This approach allows for a rapid measurement of the receptive field profile in the spatiotemporal domain [25]. Instead, the spectral response profile approach presents translating plane waves to the mammal at varying orientations and spatiotemporal frequencies [49], [50]. *Jones et al.* used both the reverse-correlation approach to construct a spatial receptive field profile [51] and measured the response to plane waves to construct a spectral response profile [22]. Subsequently, the spatial and spectral responses were compared to the Gabor filter model in the spatial and frequency domain, and the filter parameters obtained from both methods proved to be highly correlated [24]. A similar correspondence in outcome between the methods was found by *Deangelis et al.* [47], [50] in the visual cortex of cats.

In this work we extend the approach of *Jones et al.* [24] to the spatiotemporal domain and measure spectral responses of the network to translating plane waves, to which frequency-tuned spatiotemporal Gabor filters are fitted. A benefit of measuring the spatiotemporal spectral responses for optical flow is that translation is more easily described in the frequency domain [48].

### 2.3 Aperture Problem

Optical flow estimations methods are only able to resolve motion components normal to the orientation of an edge in the intensity pattern. This is known as the aperture problem [15]. In CNNs the size of the aperture of a neuron is referred to as the receptive field, which is defined as the region in the input which affects the activation of the neuron. In this work we show that the receptive field size is related to the aperture problem by training different versions of FlowNetS with varying receptive field sizes.

In neuropsychology, *Komatsu* [52] has shown the existence of a perceptual filling-in mechanism in the mammalian visual cortex for cues such as color, brightness, texture, or motion. While the precise neural workings are still under discussion, edge structure [53] and the interaction between neighboring neurons play an important role in this process [54]. In neural networks attempts have been made to implement such a mechanism as well. To allow for the interaction between neurons, a recurrent model can be used [55]. *Zweig et al.* [32], however, used an unfolded feed-forward version of a recurrent network and a multi-layer loss to allow for interaction between neurons. Their CNN-based motion interpolation architecture takes a sparse flow map and edge structure as input. They showed their motion interpolation method refines motion estimates similarly to the human visual cortex by demonstrating the filling-in effect of the network on a Kanizsa illusion. FlowNetS also features a multi-layer loss, and, in Section 6, the ability of the expanding part of FlowNetS to interpolate and refine flow maps is highlighted.

## 3 Model Details

Fig. 1 shows a schematic representation of the FlowNetS architecture, which takes two consecutive images as input. Multiple versions of FlowNetS exist. *Dosovitskiy et al.* [17] mention the use of the ReLU activation function in their work. The release of their pre-trained models, however, uses a leakyReLU activation function.[1] In order to facilitate interpretability of the motion filter analysis, we choose to use the ReLU version. With the same aim, we introduce two small adjustments. First, the bias terms are removed in the predict-flow `pf` layers because the flow is assumed to be zero-centered. Second, the kernel size in the `pf` layers is reduced from $3 \times 3$ to $1 \times 1$ to allow clearer location identification. The full details of our version of FlowNetS can be found in Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2021.3083538.

Regarding training, as in [17], we use the same data augmentation *on both* frames, but we do not use incremental flow and color augmentation *between* frames, since the authors do not specify the parameters of these mechanisms. Furthermore, the network is trained for fewer iterations (300K iterations versus 600K iterations) due to limited availability of computational resources. Evaluation on the MPI-Sintel [56] and FlyingChairs [17] datasets shows comparable performance between the slightly modified FlowNetS and the original version, as can be seen in Appendix A, available in the online supplemental material.

The synthetic dataset FlyingChairs [17], which was used to train the original and our slightly modified FlowNetS, consists of approximately 22k image pairs. The image pairs are composed of a varying numbers of chairs and background images from natural scenes. Between image pairs, a composition of translation, rotation, and scaling motion is applied. As stated in the supplementary material of [17], the size of the chairs[2] is sampled from a Gaussian with a mean and standard deviation of 200 pixels, clamped between 50 and 640 pixels. Note that the synthetic scenes also contain occlusion. Further details about the composition of affine motion can be found in [17].

## 4 Gabor Spectral Response Profile Fitting for Translation

We investigate to what motion patterns the neurons in FlowNetS are sensitive. In neuropsychology, the responses of simple cells turned out to be captured very well by Gabor filters [22], [24], [50], [57]. That simple cells act like Gabor filters makes sense, since Gabor filters are known to be optimal in the sense that they achieve maximal resolution in both the spatiotemporal and the associated frequency domains. As a consequence, they require a minimal number of filters to represent spatiotemporal information [24], [50].

Although artificial neural networks are very different in many aspects from biological ones, they were inspired by them and inherit similar traits. In particular, they seem suitable to represent spatiotemporal filters and may be subject to a similar pressure as biological networks to succinctly represent spatiotemporal patterns when having to estimate optical flow. This was our motivation to first investigate whether FlowNetS' neural responses resemble those of Gabor filters. In our investigation, we mainly focus on the deepest encoding layer in the network, the `c6` layer. As shown in Fig. 1, the activations of the feature maps of these layers are directly, linearly transformed (via `pf6`) into an initial coarse-scale horizontal and vertical flow estimate (i.e., `f6`), which is later used as the basis for refinement. Hence, the coarsest, most direct representation of optical flow is encoded in this layer. Although we focus our analysis on `c6`, the earlier layers play an important role as well. They do this not only by the determination of the activations in layer `c6` but also (in the case of `c2 - c5`) by contributing to the refinement of optical flow via skip connections.

In this section, first the theory behind Gabor filters and the spectral response fitting method is discussed, followed by the results obtained. Thereafter, we discuss the resolution in the temporal frequency domain of the fitted Gabor filters.

### 4.1 Methodology

As in [10], [14], [48], the spatiotemporal frequency-tuned Gabor filter $g$ in Cartesian coordinates centered at the origin can be written as the product of a Gaussian $w$ and a translating plane wave $s$

$$g(x, y, t) = s(x, y, t)w(x, y, t). \tag{1}$$

The (non-normalized) Gaussian $w$ is defined by

$$w(x, y, t) = \exp\left(-\frac{1}{2}\left(\frac{x_R^2}{\sigma_x^2} + \frac{y_R^2}{\sigma_y^2} + \frac{t^2}{\sigma_t^2}\right)\right), \tag{2}$$

where $\sigma_x$, $\sigma_y$, and $\sigma_t$ control the spread of the spatiotemporal Gaussian window. To decrease the number of parameters in the fitting process, it is assumed that the center of the Gaussian coincides with the center pixel of the receptive field. Furthermore, the subscript $R$ denotes a rotation operation which allows the Gaussian to be aligned along orientation $\theta_0$, and is defined as

$$\begin{aligned} x_R &= x\cos(\theta_0) + y\sin(\theta_0) \\ y_R &= -x\sin(\theta_0) + y\cos(\theta_0), \end{aligned} \tag{3}$$

where a positive value of $\theta_0$ corresponds to a clockwise rotation with respect to the positive $x$-axis. The subscript 0 indicates the parameter value corresponding to the peak response of the Gabor filter. This orientation, which

[1] https://lmb.informatik.uni-freiburg.de/resources/binaries/flownet/flownet-release-1.0.tar.gz

[2] Note that, in [17], the authors do not specify how the size of a chair is determined, so there is a certain ambiguity around this parameter.

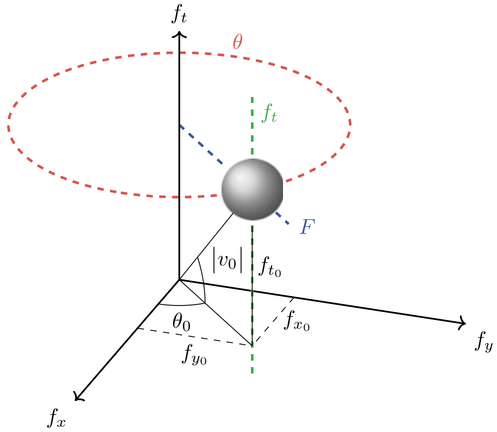


Fig. 2. Illustration of the half-magnitude profile in the 3D frequency domain of a spatiotemporal Gabor filter. The three ranges along which the responses of the Gabor half-magnitude profile are evaluated for the spectral response profile fitting process are shown in color.

corresponds to the preferred direction of motion of the filter, is related to the spatial frequencies via $\theta_0 = \tan^{-1}(f_{y_0}/f_{x_0})$.

A translating plane wave $s$ in the Cartesian coordinate system can be written as

$$s(x,y,t) = \cos\left(2\pi\left(F_0 x_R - f_{t_0} t\right) + \varphi_0\right), \tag{4}$$

where the spatial frequency magnitude $F_0$ is related to the spatial frequencies via $F_0 = (f_{x_0}^2 + f_{y_0}^2)^{1/2}$, $f_{t0}$ indicates the temporal frequency, and $\varphi_0$ denotes the phase of the filter. The dependence of $s$ on $y$ is due to $x_R$, which is a function of $x$ and $y$ (see Eq. (3)). A Gabor filter is said to be even when $\varphi_0 = 0$ and odd when $\varphi_0 = \pm\pi$. Further, note that the preferred velocity of the filter $v_0$ is related to $F_0$ and the temporal frequency $f_{t_0}$ via $v_0 = f_{t_0}/F_0$, as in [10]. A higher spatial frequency $F_0$ allows tracking of motion of thinner image structures. When a signal is sampled in time or space, frequency components which are larger than or equal to 0.5 cycles per frame (i.e., the Nyquist frequency) become undersampled and aliasing occurs. Thus, if we limit ourselves to signals which do not suffer from aliasing, the maximum velocity a signal can have is limited by its $F_0$. Fig. 2 shows the 3D frequency space with the half-magnitude profile of a Gabor filter.

Because we will fit the response of phase-sensitive Gabor filters, we highlight three phase-dependent convolution phenomena. Note that a valid convolution[3] of two tensors with equal size corresponds to their dot product. First, because a sine is an odd signal, the dot product of two sines at opposite frequencies is negative. Second, the dot product of a cosine at opposite frequencies will be positive due to the even nature of the function. Third, sine and cosine are decorrelated and thus the dot product will be zero between these two signals.

*Gabor Spectral Response Profile Fitting.* In the Gabor spectral response fitting process, translating grayscale plane waves $s$ are used as input to the network, and we try to minimize the difference in response between filters in the `c6` layer of our FlowNetS and spatiotemporal Gabor filters $g$. To better approximate the response of `c6` filters, we enhance the Gabor filter output with a gain term $K$, a bias term $b$, and pass the response through a ReLU non-linearity. Then, the response $r$ to a convolution with a translating plane wave $s$ and a Gabor filter $g$ is given by:

$$r = \mathrm{ReLU}(K(s(x,y,t) * g(x,y,t)) + b), \tag{5}$$

where $r$ is a function of nine parameters (i.e., $F_0$, $\theta_0$, $f_{t_0}$, $\varphi_0$, $\sigma_x, \sigma_y, \sigma_t$, $K$, $b$), which are estimated in a two-step process.

First, a gridsearch is performed to determine the location in the spatiotemporal frequency domain with the highest response per filter in the `c6` layer. We denote the response of the filters in the network by $\hat{r}$, and their peak response value by $\hat{r}_0$. Because the fitted Gabor filters are phase sensitive, this amounts to estimating four parameters (i.e., $F_0$, $\theta_0$, $f_{t_0}$, $\varphi_0$). Therefore, a four-dimensional grid of translating plane waves (i.e., the input to the network) is constructed using all combinations of these parameters within a given range and step size (see Appendix B, available in the online supplemental material). The range for the value of half spatial wavelength $\lambda/2 = 1/2F$ is chosen so that it captures the sizes of the chairs present in the training dataset (as explained in Section 3).

Second, once the peak response of the `c6` filters is found, we estimate the spatiotemporal spread of the Gaussian (determined by $\sigma_x, \sigma_y, \sigma_t$), the gain $K$, and the bias $b$. This is done by minimizing the difference in response between the fitted Gabor filters $r$ (see Eq. (5)) and the corresponding `c6` filters $\hat{r}$ along three separate ranges in the spatiotemporal frequency space ($F$, $\theta$, and $f_t$). These ranges are illustrated in Fig. 2, and further described in Appendix B, available in the online supplemental material. We define the cost function $\mathcal{L}$ in response to a convolution with a translating plane wave $s$ as

$$\begin{aligned}\mathcal{L} &= \sum_i (r_i - \hat{r}_i)_F^2 + \sum_j (r_j - \hat{r}_j)_\theta^2 + \sum_k (r_k - \hat{r}_k)_{f_t}^2 \\ &= \mathcal{L}_F + \mathcal{L}_\theta + \mathcal{L}_{f_t},\end{aligned} \tag{6}$$

where $\mathcal{L}_F$, $\mathcal{L}_\theta$, and $\mathcal{L}_{f_t}$ denote the sum of squared errors over the respective ranges. We constrain the bounds of the Gabor filter parameters to obtain reasonable values, which leads to a non-linear bounded convex optimization problem which is solved using the robust trust-region-reflective algorithm [58]. In order to compare the obtained cost values between `c6` filters, we construct a normalized cost value $\mathcal{L}_{norm}$ by dividing the cost by the squared peak response of the filter: $\mathcal{L}_{norm} = \mathcal{L}/\hat{r}_0^2$.

### 4.2 Results

We found 592 of the 1024 filters in the `c6` layer of FlowNetS to have an activation larger than zero when using the aforementioned input waves. The location of the peak response of the active `c6` filters in terms of half spatial wavelength $\lambda_0/2$, orientation $\theta_0$, and temporal frequency $f_{t_0}$ can be seen in Fig. 3 (left). As shown, the locations of the peak responses of the filters are well distributed over all angles. Radially, there is a concentration around a half spatial wavelength of 200 pixels. Two possible explanations for this are the fact that (i) the average size of the chairs in the training dataset

[3] We use "convolution" to refer to the correlation of a filter over an image to remain consistent with the CNN terminology.

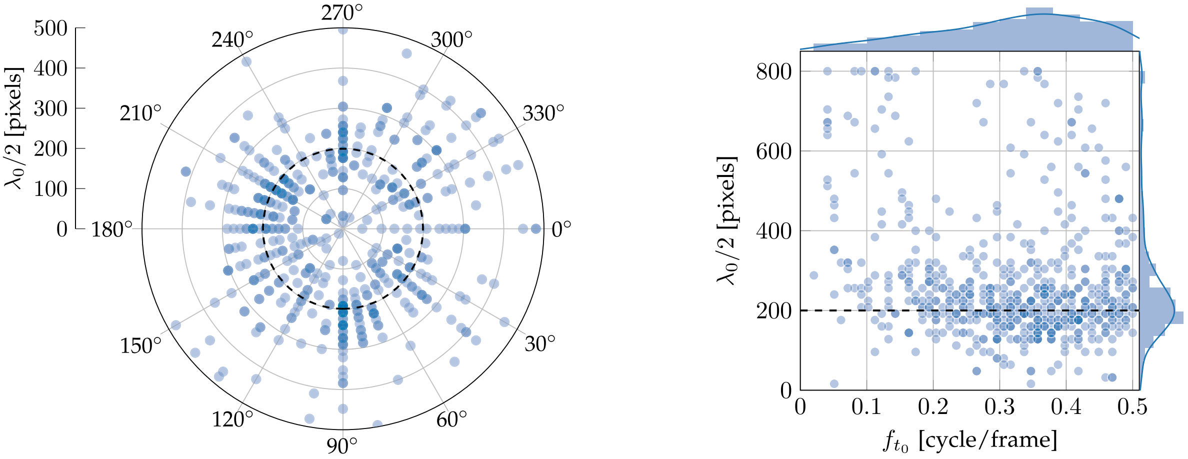


Fig. 3. Location of peak response $\hat{r}_0$ per c6 filter in the spatiotemporal frequency domain in response to translating plane waves. *Left:* Half spatial wavelength $\lambda_0/2$ and orientation $\theta_0$ corresponding to peak response $\hat{r}_0$ per filter. *Right:* Half spatial wavelength $\lambda_0/2$ and temporal frequency $f_{t_0}$ corresponding to peak response $\hat{r}_0$ per filter. In both plots, the black dashed lines indicate the peak of the distribution in the half spatial wavelength dimension, which is around 200 pixels.

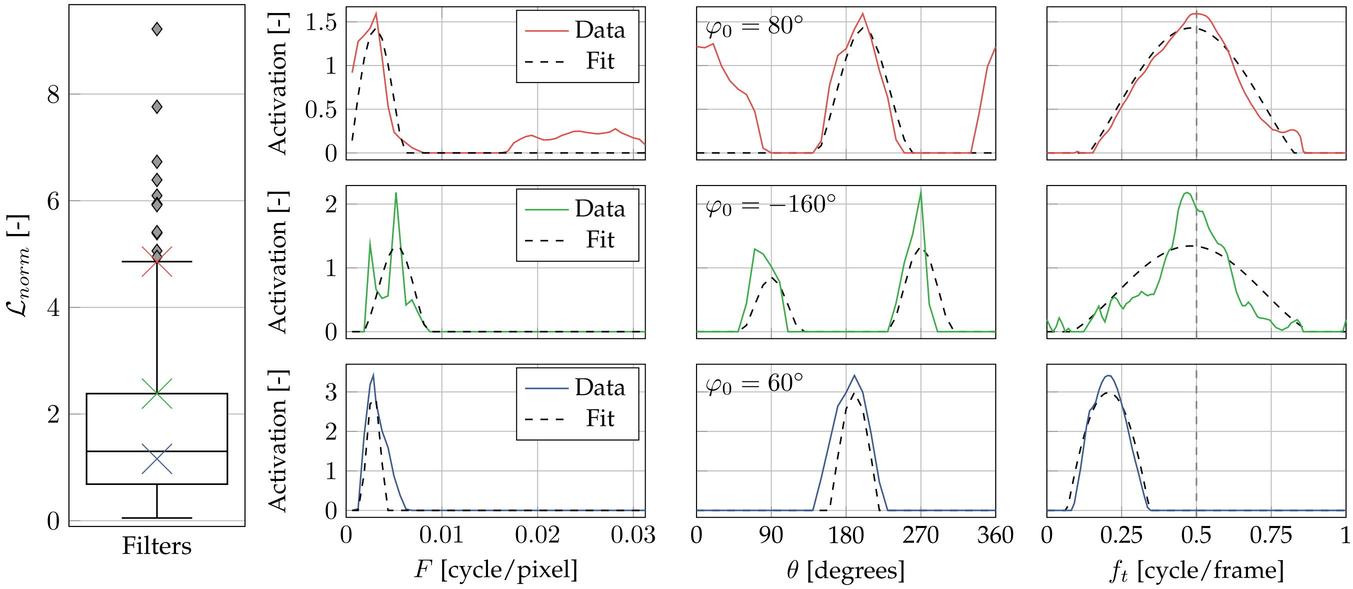


Fig. 4. Quantitative results of the spectral Gabor filter fitting process. *Left:* Boxplot containing the total normalized cost $\mathcal{L}_{\text{norm}}$ per filter (592 filters). *Right 3x3 plots:* Row-wise, the measured responses of three different c6 filters and their corresponding Gabor fits. The blue, green, and red c6 filters correspond to the crosses at the median, near the 75th percentile and near the upper whisker limit of the boxplot, respectively.

is 200 pixels, or that (ii) the half of the receptive field size of c6 filters is 192 pixels. The concentration of the peak responses becomes even more apparent in Fig. 3 (right), which shows the distribution along the temporal and half spatial wavelength axes. Furthermore, we note that the distribution of the temporal frequencies is skewed toward the Nyquist limit of 0.5 cycles per frame. A possible reason for this is the low resolution in the temporal frequency due to the low number frames used as input to the network. This is further discussed in Section 4.3.

The main observation of our spectral analysis is that the fitted modified Gabor functions (i.e., Eq. (5)) capture the spatiotemporal frequency selectivity of the active c6 filters of FlowNetS accurately. In order to give insight into the goodness of fits for all neural responses in the c6 layer, we show three example responses corresponding to different normalized cost values $\mathcal{L}_{norm}$ in Fig. 4. Note that the fitted Gabor filters correspond well to the response of the blue and green c6 filters (with $\mathcal{L}_{norm}$ at 50 percent, 75 percent of the distribution); but, in the red case (an outlier), the fitted Gabor shows a substantial deviation from the measured c6 response near $\theta = 0$.

This experiment was also performed for the other convolutional layers of the network's encoder segment. As shown in Table 1, the lower the layer, the smaller the receptive field size and hence the upper limit for the half spatial wavelength is decreased. According to the average (normalized) fitting error per layer $\mathcal{L}_{norm}$, the response of neurons in the c3–c6 layers fits well the translational Gabor filter model, while our methodology suggests that neurons in c1 and c2 are not yet as motion-selective as Gabor filters. Table 1 also shows that c6 is characterized by a higher $\mathcal{L}_{norm}$ than its preceding layer. A possible explanation for this is that, in the earlier layers, the network is only able to perceive less complex motions which better fit the Gabor filter model.

TABLE 1
Result of the Gabor Spectral Response Fitting Procedure for Different Convolutional Layers of the Encoder Part of FlowNetS

| Layer | $\mathcal{L}_{norm}$ | Max. $\lambda/2$ | Num. active filters/filters |
|---|---|---|---|
| conv6_1 | 1.65 | 800 | 592/1024 |
| conv5_1 | 1.42 | 270 | 372/512 |
| conv4_1 | 1.44 | 270 | 408/512 |
| conv3_1 | 1.67 | 95 | 234/256 |
| conv2 | 3.37 | 47 | 62/128 |
| conv1 | 4.71 | 10 | 64/64 |

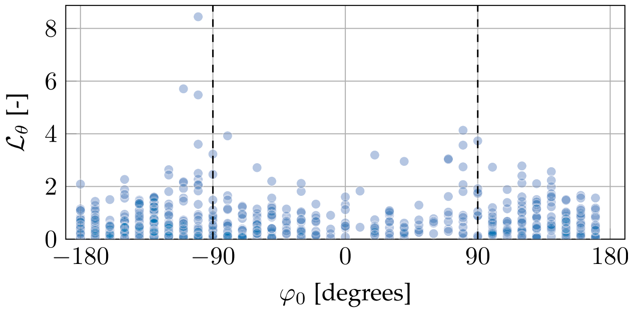


Fig. 5. Orientation cost $\mathcal{L}_\theta$ per filter as a function of $\varphi_0$.

Coming back to c6, the good fit for the majority of neurons supports the choice for the Gabor filter as opposed to other types of models. Of course, one can argue that the Gabor filter does not perfectly capture the response and a more complex model may lead to a better fit. Below, we will extensively delve into the cases in which the Gabor model seems to fall short of explaining c6's neural responses. Here, it is important to note that in principle, we already have such a complex model: the neural network itself. The advantage of the Gabor model is that it has a low number of parameters that can be readily interpreted. Indeed, in neuropsychology, the step to more complex filters was only made when it became necessary for characterizing "complex" cells that did not respond to simple stimuli [57]. The fits and error patterns above the 75 percent percent threshold (corresponding to the green c6 filter) are very interesting, and we visually inspected them for systematic deviations. Visual inspection is performed instead of an auto-correlation procedure since the latter is not possible due to a non-uniformly spaced polar 3D frequency grid[24]. Fig. 6 contains the qualitative results used for this analysis, while Appendix C, available in the online supplemental material, evaluates the generalizability of the fitted Gabor filters to more complex natural stimuli.

Similarly to the blue filter in Fig. 4, Fig. 6A shows a c6 filter whose response fits nicely in the Gabor filter framework. On the other hand, we find three types of systematic deviations (i.e., Figs. 6B, 6C, and 6E) from the Gabor model, and also conclude that some patterns are too complex for interpretation, such as the c6 filter shown in Fig. 6D.

The filter in Fig. 6B shows a deviation from the fitted Gabor 180 degrees away from $\theta_0$. This filter is responsive to edge structure (i.e., $|\varphi_0| \approx 90^\circ$) and is thus approximately odd, since the dot product of two odd signals at opposite frequencies results in a negative value. However, this filter still produces a positive activation at the opposite spatial frequency, corresponding to 180 degrees away from $\theta_0$. In Fig. 5 the distribution of the phase values $\varphi_0$ versus orientation cost $\mathcal{L}_\theta$ for all filters is depicted. As shown, there are multiple filters responsive to edge structure that have a high $\mathcal{L}_\theta$ (e.g., the red filter in Fig. 4). One possible reason for this systematic deviation from the Gabor response is that the network is able to learn flow filters that are invariant to polarity (meaning white-black or black-white transitions).

We find two c6 filters that exhibit weak directional bias, an example of which can be found in Fig. 6C. Moreover, we also find filters that exhibit two or more Gaussian peaks with similar peak response magnitudes but tuned to different spatial frequencies $F_0$, orientations $\theta_0$, and temporal frequencies $f_{t_0}$. An example of such a filter can be found in Fig. 6 E, and its 2D spatiotemporal representation is shown in Fig. 7. A possible explanation is that these filters are sensitive to occlusion, as discussed in Section 5. Lastly, we find filters that appear noisy and are hard to interpret given the limitations of our methodology (further discussed in Section 5). Such an example can be seen in Fig. 6D.

### 4.3 Temporal Bandwidth

For orientation $\theta$ and temporal frequency $f_t$, the bandwidth is defined as the width of the filter which provides an output above half the maximum response. This leads to a bandwidth in degrees $\Delta\theta_{1/2}$ and cycles per frame $\Delta f_{t_{1/2}}$ for orientation and temporal frequency respectively

$$\Delta f_{t_{1/2}} = f_{t_{\max}} - f_{t_{\min}} \tag{7}$$

$$\Delta\theta_{1/2} = \theta_{\max} - \theta_{\min}. \tag{8}$$

For spatial frequency $F$, the bandwidth is defined in terms of octaves as follows:

$$\Delta F_{1/2} = \log_2\left(F_{\max}/F_{\min}\right). \tag{9}$$

Although we estimate the Gabor parameters of the active c6 filters in the fitting process, the apparent bandwidth of these filter differs due to the non-linear transform in Eq. (5). The bandwidth is therefore measured based on the fitted Gabor filter response. In Fig. 8, the bandwidth of $F$, $\theta$, and $f_t$ can be seen. As shown, the interquartile range for spatial frequency bandwidth is between 1 and 2 octaves and the median orientation bandwidth is approximately $50^\circ$. Lastly, the temporal frequency bandwidth is of large extent with a median of approximately 0.27 cycles per frame.

We note that the network is able to narrow the extent of the filter response in the temporal domain using the non-linear transform in Eq. (5). An illustration of this mechanism can be seen in Fig. 9. As shown, the extent of the half-magnitude profile is wider if the non-linear transformation is not employed. This figure also shows what happens when more frames are added to the input and the other parameters are kept the same (see Fig. 9, bottom). This suggests that an even narrower extent could be reached by feeding the network with more images over time than just the two

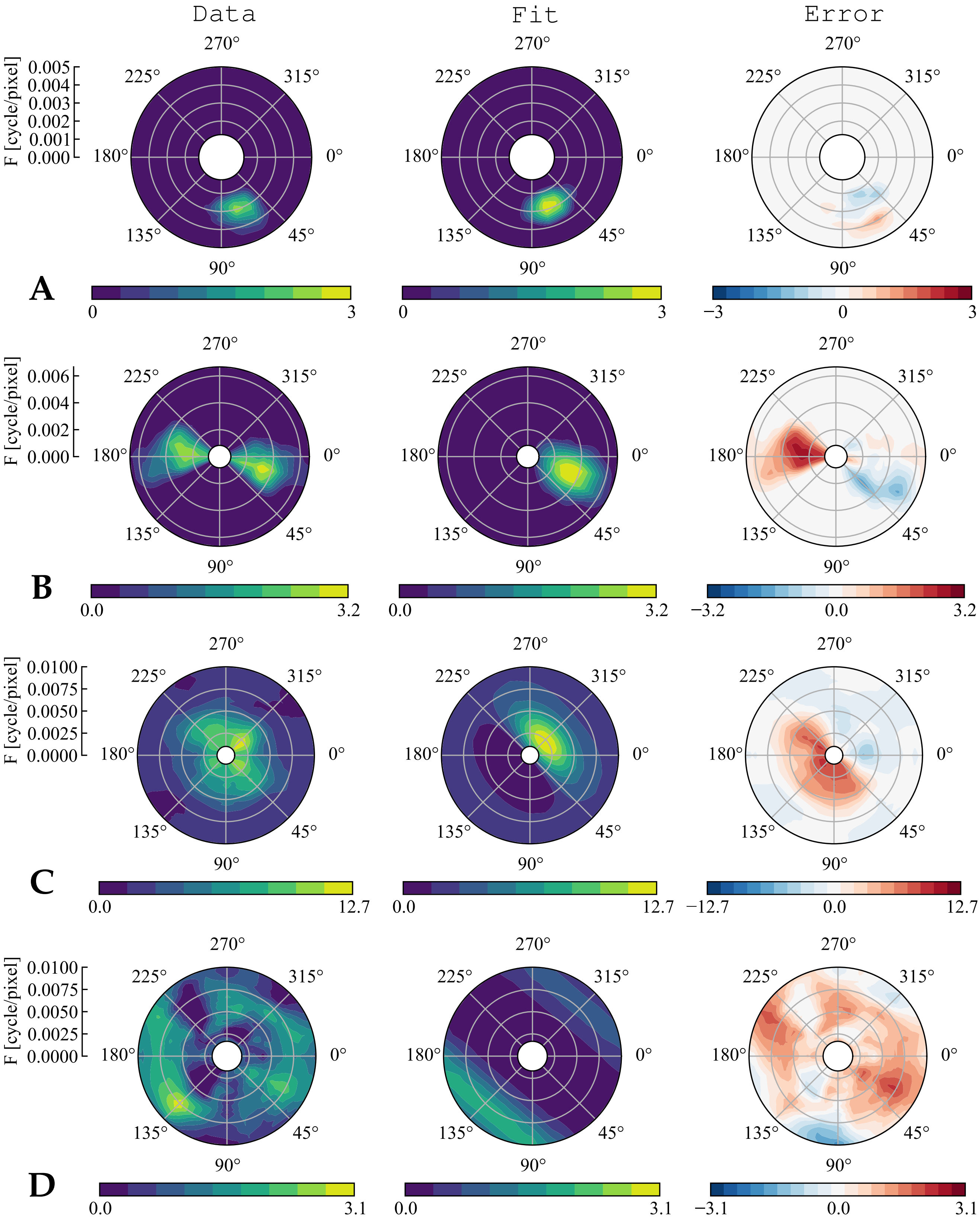


Fig. 6. Qualitative results of the error patterns of the spectral Gabor fitting process. The spectral response profiles are shown as a function of spatial frequency $F$ and orientation $\theta$. Data shows to the measured response of a c6 filter, Fit is the response of the corresponding fitted Gabor filter, and Error shows their difference. Evaluations are with respect to $f_{t_0}$ and $\varphi_0$. (A) c6 filter whose response profile is accurately captured by the Gabor model. (B) Red c6 filter from Fig. 4, which activates on opposite spatial frequencies. (C) c6 filter with a very weak directional bias. (D) Noisy c6 filter pattern (further discussed in Section 5). (E) For this c6 filter, the spectral response profile for three different temporal frequency $f_t$ values is visualized. Two different Gaussian peak responses at opposite orientation can be observed at $f_t = 0.3$ and $f_t = 0.5$ cycles per frame. The blue and red lines correspond to the axes of the 2D representation of this filter shown in Fig. 7.

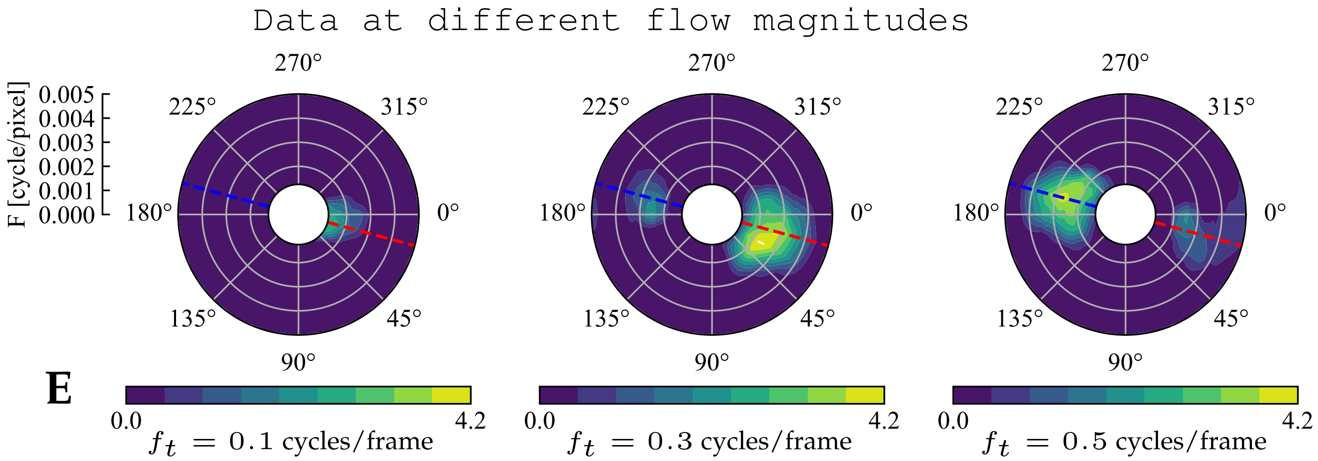


Fig. 6. (Continued).

subsequent images used in FlowNetS. A higher resolution in the frequency domain is beneficial as it allows for a more precise measurement of the flow.

## 5 Network Response to Dilation & Rotation

In this section, the sensitivity of c6 filters to dilation and rotation is analyzed. First, we explain the limitations of the spectral Gabor response profile fitting process and why we are not able to discern filters activating on translation, dilation, rotation, and occlusion with this methodology. Second, the theory used to identify filters sensitive to dilation and rotation is presented. Lastly, our results are discussed.

Note that Gabor translation filters [14] and occlusion filters [59] already have an analytical description in both the space-time and frequency domain. Such a description of dilation and rotation is, to the best of the authors' knowledge, missing. Therefore, fitting c6 filters to a dilation and rotation motion filter model requires a novel mathematical foundation which is outside of the scope of this work.

### 5.1 Limitations of the Spectral Response Profile Fitting

In the first part of the spectral response fitting process, a gridsearch is performed to find the peak response. In the subsequent fitting process, three response lines are generated by varying either $F$, $f_t$, or $\theta$, whilst keeping $\varphi$ constant. This method only allows the measurement of the relative attenuation in amplitude with respect to the peak response $\hat{r}_0$. This is sufficient for translation, which can be defined as a single constant phase Gaussian in the 3D frequency spectrum and thus produces a Gaussian in response. However, it is insufficient for other more complex motion types.

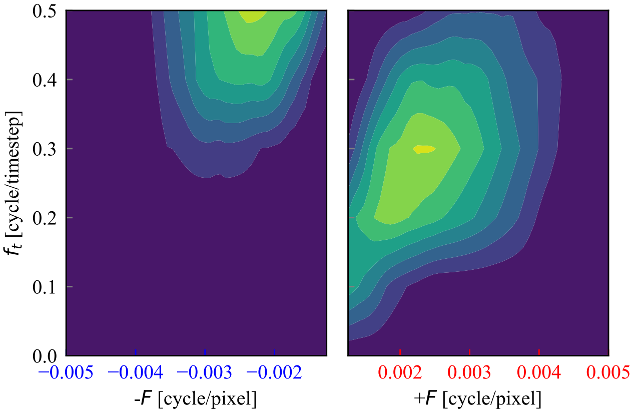


Fig. 7. Spatiotemporal frequency representation of the measured filter response in Fig. 6 E. The positive and negative $F$-axes correspond to the blue and red lines in Fig. 6 E.

Due to the ReLU activation function, the dot product of two translating plane waves at the same frequency, which are more than or equal to 90 degrees out-of-phase, is zero. Note that a convolution in the space-time domain equals to multiplication in the frequency domain according to the convolution theorem[46]. Because we evaluate the convolution response only at discrete frequencies of $k$ integer multiples along the $f_x$, $f_y$, and $f_t$ axis, only a single frequency component of the Fourier-transformed translating plane wave $S$ will contain power.[4] Then, if we define the $k$th frequency component of $S$ as the complex vector $\mathbf{p}$, and the $k$th frequency component of the Fourier transformation of the filter to be analyzed as $\mathbf{q}$, the phase difference between these two complex vectors is defined as the angle $\psi$ and given by:

$$\psi = \cos^{-1}\left(\frac{\mathbf{p}\cdot\mathbf{q}}{|\mathbf{p}||\mathbf{q}|}\right), \tag{10}$$

where the maximum value of $\psi$ is $\pi$, and values of $\psi \geq \pi/2$ result in a zero response due to the ReLU in Eq. (5).

*Convolution Response: Dilation & Rotation Filters.* To determine which frequency components of dilation, rotation, and occlusion are more than 90 degrees out of phase, the Discrete Fourier Transform (DFT) [46] is used to transform a simulated space-time signal to a representation in the frequency domain. Fig. 10 shows the convolution response of a dilation filter $dw$ with a translating plane wave $s$. From this figure, it can be observed that a diamond-like pattern emerges in the response, due to the immeasurable out-of-phase components of $dw$ and $s$. Because we evaluate the responses along lines orthogonal to the peak response, the pattern perceived is indicated by the dashed lines in the

[4] Not taking into account the complex conjugate component.

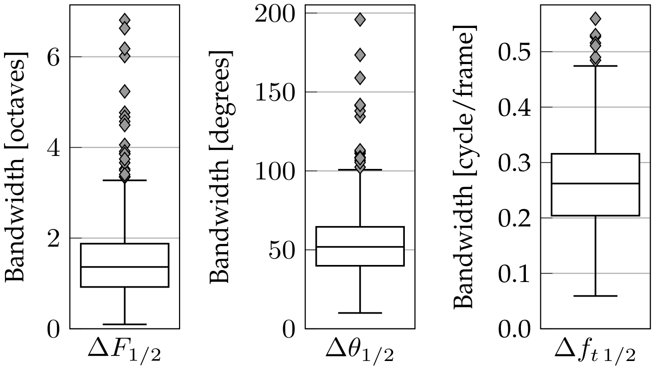


Fig. 8. Bandwidth of spatial frequency $F$, orientation $\theta$, and temporal frequency $f_t$ of the fitted Gabor filters of the 75 percent active c6 filters with the lowest $\mathcal{L}_{norm}$.

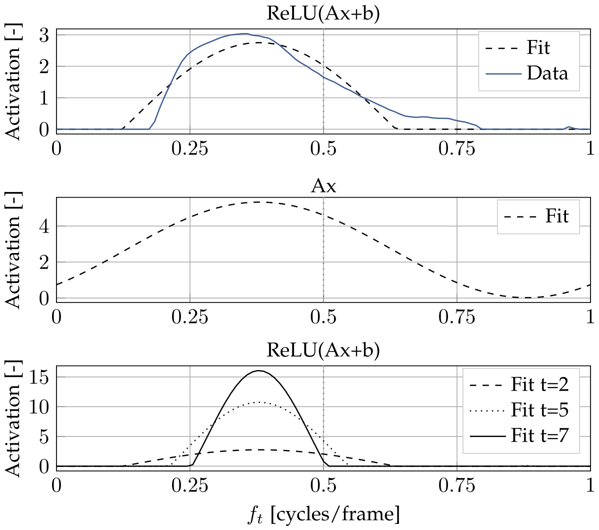


Fig. 9. Illustration of how the network is able to decrease the extent of the filter response in the temporal domain. *Top:* Fit and measured data for the median c6 filter (see Fig. 4). *Middle:* The response of the fitted Gabor filter without the bias term and ReLU non-linearity. *Bottom:* Response of the fitted Gabor filter when the number of frames is increased.

right-most plot of this figure, which correspond to the colored linear patterns in Fig. 2. Thus, a Gaussian will be perceived along the spatial and the temporal frequency ranges. Hence, we are not able to discern between dilation and translation filters.

Similarly, Fig. 11 shows the convolution response of a rotation filter $cw$ with $s$. Note that the 3D power spectrum of $cw$ is different from a Gaussian. At high temporal frequencies (i.e., $\pm 0.2$ cycles per frame), the frequency components of $cw$ and $s$ are out-of-phase. Thus, these frequency components will not be detected. The pattern perceived along the varying $\theta$ (also shown in Fig. 2) is two Gaussian lobes at opposite frequency. This pattern is similar to the convolution response of a cosine Gabor filter tuned to stationary patterns (i.e., zero temporal frequency). Therefore, our methodology is also not able to detect rotation filters.

*Convolution Response: Occlusion Filters.* Furthermore, we convolve an occlusion filter, using the description of *Beauchemin et al.* [59], with translating plane waves $s$. Occlusion in the spatiotemporal domain can be described as the combination of a Gaussian, a Heaviside step function, and two translating plane waves translating with different frequencies, as shown in Fig. 12. The power spectrum of the Fourier-transformed filter can be described as two Gaussian filter pairs with *tails* due to the Heaviside step function. The angle $\psi$ demonstrates that these tails have a large phase difference. Consequently, only the pattern above the dashed line is detected using our methodology, which corresponds to two different Gaussian lobes tuned to different frequencies. This pattern resembles that of Figs. 6 E and 7, thus making it likely that the filter represented in these figures is responsive to occlusion. However, it should be noted that we are not able to discern such a pattern from the superposition of two regular Gabor filter pairs tuned to different frequencies.

### 5.2 Methodology

In order to still assess the sensitivity of the c6 filters to dilation and rotation, we come up with a different methodology in which two gridsearches are performed. We assess the locations of the peak responses for filters which have a higher response to dilation or rotation than to translation. We do not classify a filter as either a rotation or dilation filter, since a filter can be sensitive to a composition of these respective motions.

*Dilation Parametrization.* As in [11], a dilating wave $d$ is given by:

$$d(x, y, t) = \cos\left(2\pi F_0(x_r - \alpha x_r t) + \varphi_0\right), \tag{11}$$

where $\alpha$ denotes the dilation factor. The training dataset used to train FlowNetS, i.e., FlyingChairs [17], defines scaling motion in terms of the affine scaling factor $h$. Because

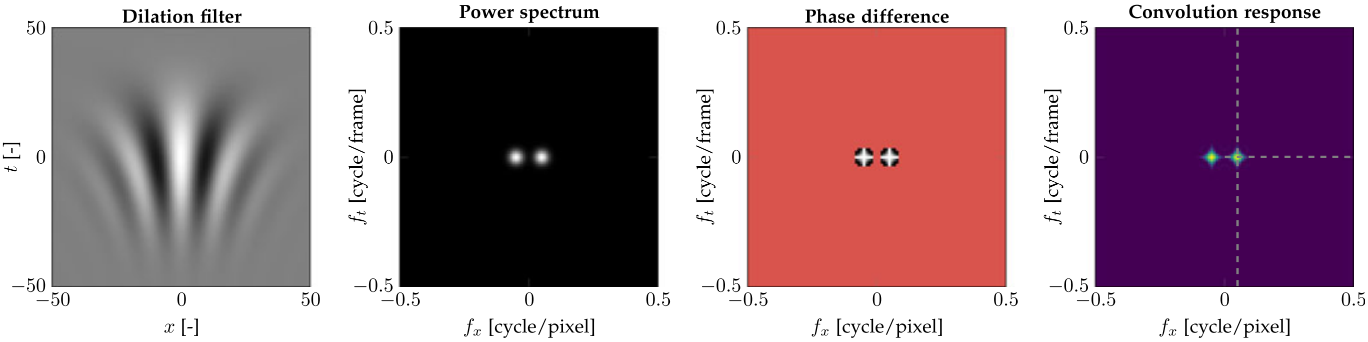


Fig. 10. Convolution response of a dilation filter $dw$ with a translating plane wave $s$ evaluated with spatiotemporal frequencies at $k$ integer multiples of the fundamental frequency. In the $\psi$ plot, a larger phase difference corresponds to a darker color with black being equal to or greater than $\pi/2$. A red mask is applied to frequency components with low power. The dashed lines indicate the Gaussian pattern perceived by the spectral fitting procedure.

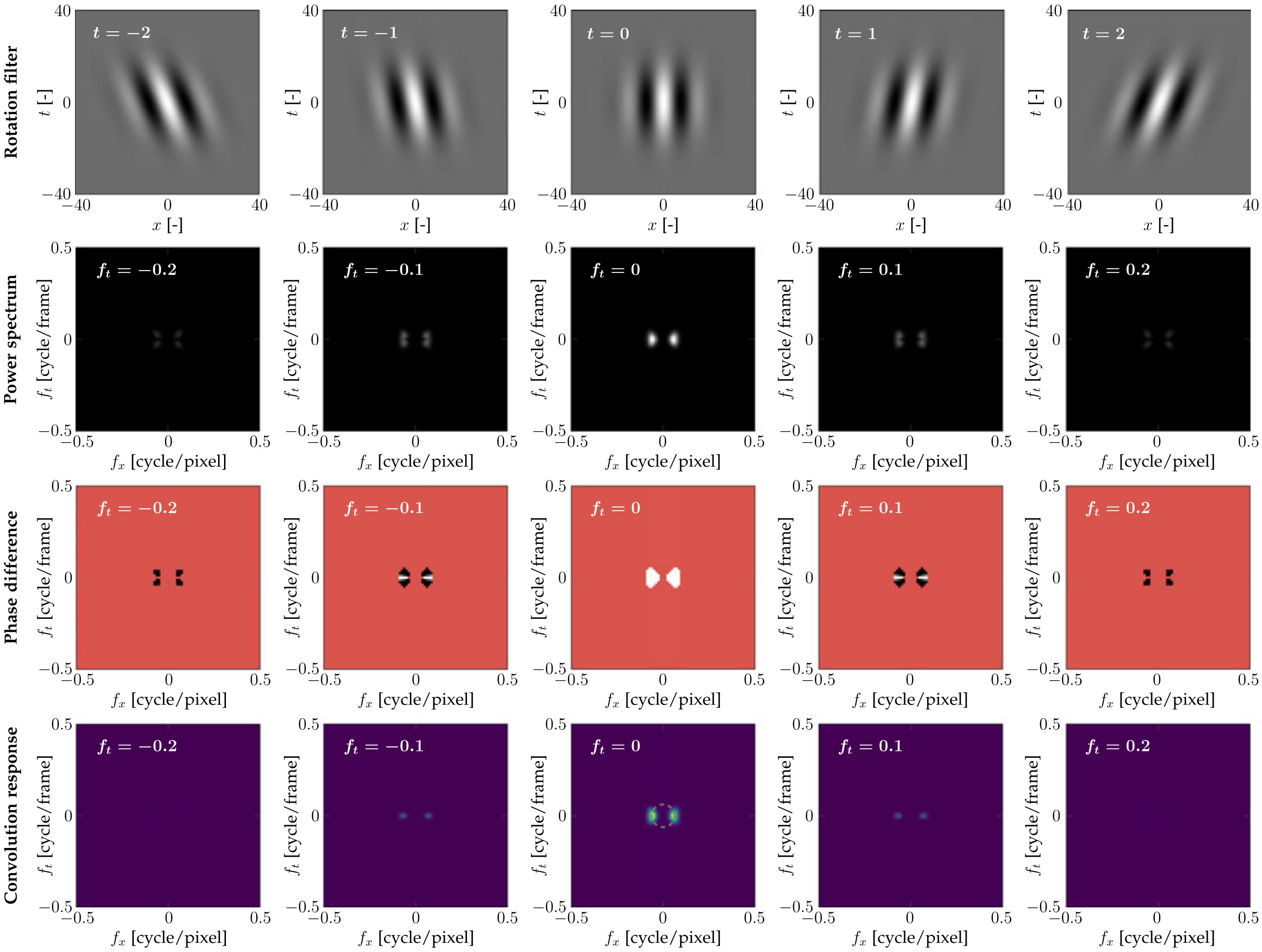


Fig. 11. Convolution response of a rotation filter $cw$ with a translating plane wave $s$ evaluated with spatiotemporal frequencies at $k$ integer multiples of the fundamental frequency. In the $\psi$ plot, a larger phase difference corresponds to a darker color with black being equal to or greater than $\pi/2$. A red mask is applied to frequency components with low power. The dashed circle indicates the double lobe Gaussian pattern perceived by the spectral fitting procedure.

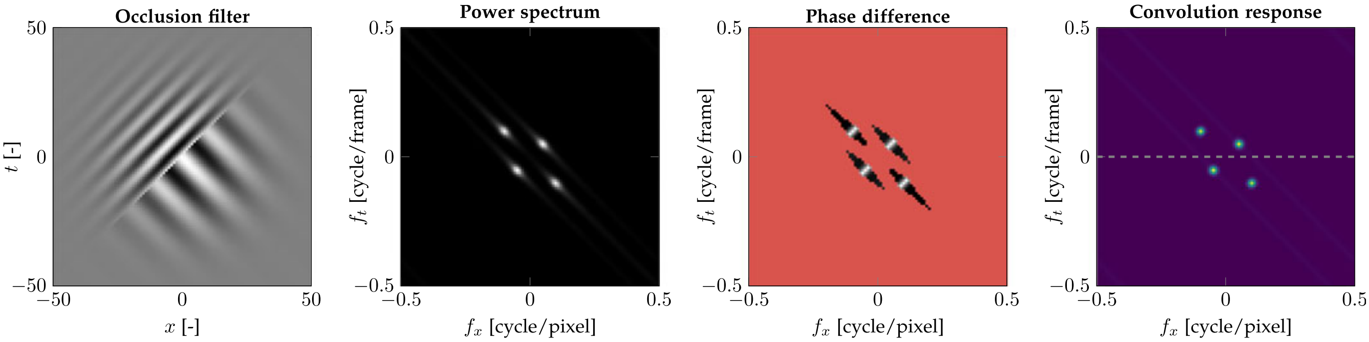
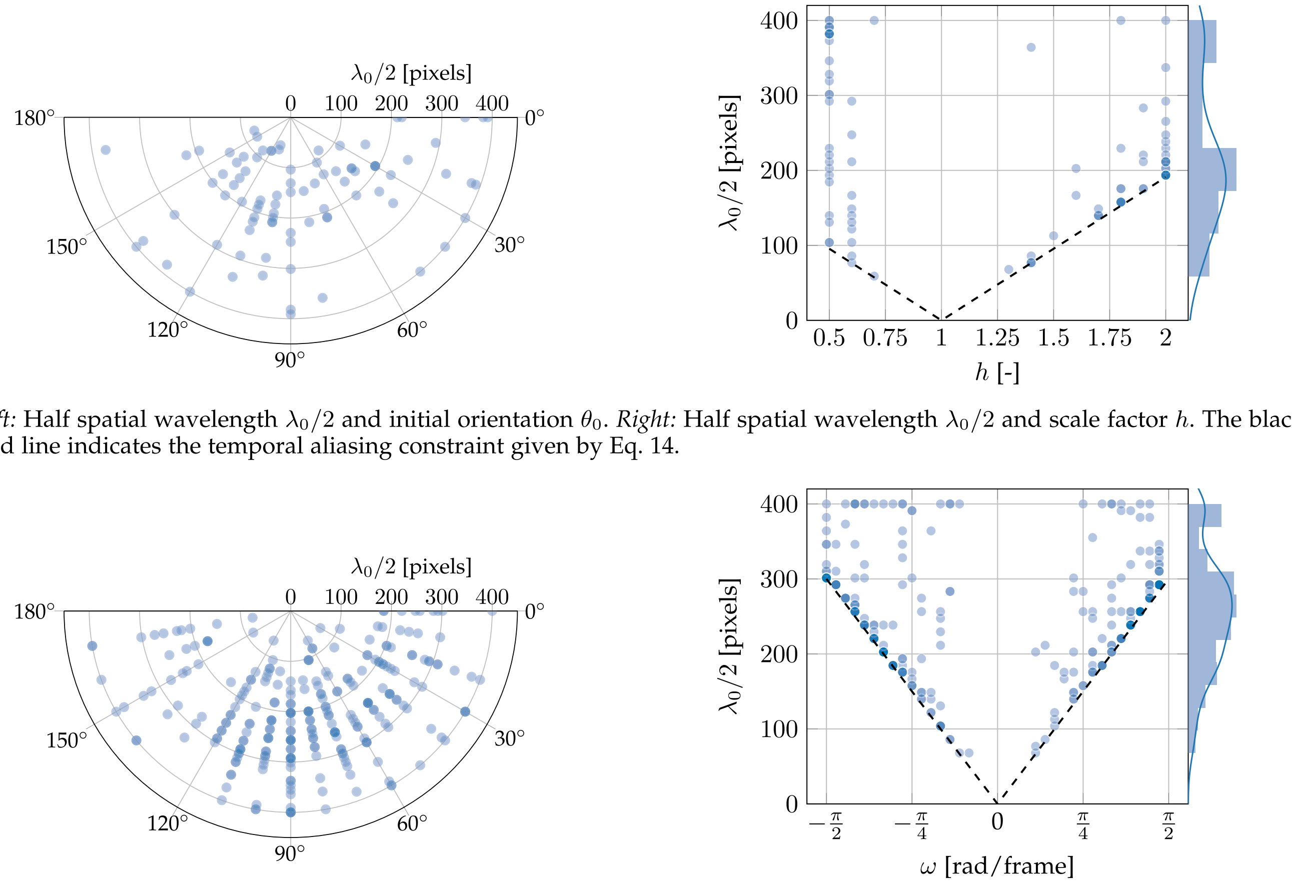


Fig. 12. Convolution response of an occlusion filter with a translating plane wave $s$ evaluated with spatiotemporal frequencies at $k$ integer multiples of the fundamental frequency. *Left:* Example occlusion signal following the description of *Beauchemin et al.* [59]. *Middle left:* The power spectrum of the Fourier-transformed occlusion filter. *Middle right:* The angle $\psi$ indicating the phase difference between the Fourier components of the occlusion filter and $s$. A larger phase difference corresponds to a darker color with black being equal to or greater than $\pi/2$. A red mask is applied to frequency components with low power. *Right:* Convolution response between the occlusion filter and $s$. The pattern above the dashed gray line resembles that of Figs. 6 E and 7.

the network only takes two frames as input, we define the relation between $h$ and $\alpha$ as follows:

$$h = \frac{1}{1-\alpha}. \tag{12}$$

The gridsearch is performed for the [0.5,2.0] range of $h$, as it encapsulates the values encountered during training. More details about this search space can be found in Appendix B, available in the online supplemental material. In order to mitigate the effect of temporal aliasing, the search



(a) *Left:* Half spatial wavelength $\lambda_0/2$ and initial orientation $\theta_0$. *Right:* Half spatial wavelength $\lambda_0/2$ and scale factor $h$. The black dashed line indicates the temporal aliasing constraint given by Eq. 14.

(b) *Left:* Half spatial wavelength $\lambda_0/2$ and initial orientation $\theta_0$. *Right:* Half spatial wavelength $\lambda_0/2$ and angular temporal frequency $\omega$. The black dashed line indicates the temporal aliasing constraint given by Eq. 17.

Fig. 13. Location of peak response $\hat{r}_0$ per c6 filter in the spatiotemporal frequency domain in response to dilating (top) and rotating waves (bottom). Only filters whose peak response $\hat{r}_0$ was higher than the maximum found in the translation gridsearch are shown.

space is constrained so that the velocity of a point is not more than half its spatial wavelength $\lambda_0/2$. For a dilating wave, this velocity is given by:

$$v = \left(\frac{1}{1-\alpha} - 1\right)x = (h-1)x. \tag{13}$$

Then, the temporal aliasing constraint for dilating waves is given by:

$$(h-1)x \leq \frac{1}{2}\lambda_0. \tag{14}$$

*Rotation Parametrization.* A rotation wave $c$ is given by:

$$c(x, y, t) = \cos\left(2\pi F_0 x_r(t) + \varphi_0\right), \tag{15}$$

where $x_r(t)$ varies with time, and is defined as

$$x_r(t) = x\cos\left(\theta_0 + \omega t\right) + y\sin\left(\theta_0 + \omega t\right), \tag{16}$$

where $\omega$ denotes the angular velocity in radians per frame.

The search space for the rotation gridsearch can be found in Appendix B, available in the online supplemental material. A constraint was also added to limit the effect of temporal aliasing. $\omega$ can be related to a point at distance $m$ from the center of rotation by $v = \omega m$. The maximum distance from the center of rotation to the edge is equal to half the receptive field size, which is 383 pixels in the c6 layer of our FlowNetS. As the wave rotates around the center pixel, the velocity at this point should thus be lower than half the spatial wavelength. The constraint is given by the following relation:

$$\omega m_{\max} \leq \frac{1}{2}\lambda_0. \tag{17}$$

### 5.3 Results

The peak responses of c6 filters which have a higher activation to dilation than to translation (i.e., approximately 15 percent of the active filters) are shown in Fig. 13a. These filters show a radially dispersed pattern along the $\theta$-axis, and a peak in the distribution of half spatial wavelengths near 200 pixels. Lastly, peak responses are often close to the temporal aliasing limit and the maximum scaling value of the gridsearch. This is similar to the temporal peak response location for the translation gridsearch (see Fig. 3).

In Fig. 13b, the peak responses of the c6 filters for the rotation gridsearch are shown. It can be observed that most filters are active near the temporal translation and temporal rotational aliasing limit. Also, a peak in the distribution of half spatial wavelengths can be identified around 250 pixels, which is slightly higher than expected. A possible explanation for this discrepancy is that rotation is actually a 3D

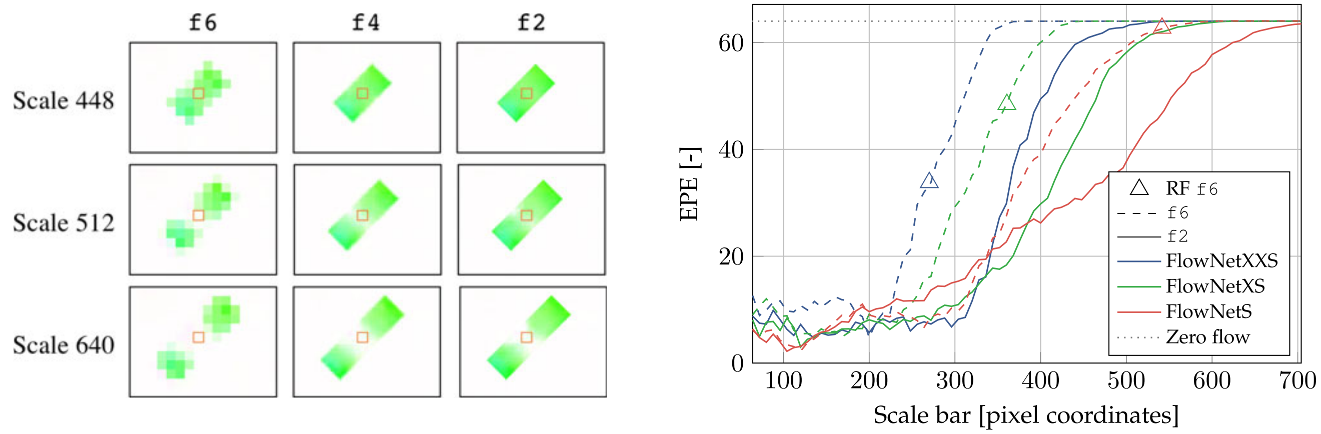


Fig. 14. Response of our FlowNetS and its two variations, FlowNetXS and FlowNetXXS, to diagonally translating bars with motion magnitude $|\mathbf{u}| = 64$ pixels. *Left:* f6, f4 and f2 FlowNetS flow maps in response to downward-left diagonally translating bars of different scales, using the color-coding scheme from [60]. The red squares highlight the output region used for evaluating the error. *Right:* EPE versus scale of the bar in pixel coordinates. RF f6 indicates the diagonal receptive field size in pixel coordinates corresponding to the f6 flow map.

motion and thus the scale should also be limited along its radial axis. Approximately 45 percent of the active c6 filters activate more on rotation than on translation, which could be due to the fact that we do not limit the wavelength along the axis of rotation. The points in the motion field at the far end of the receptive field then move with a very high velocity, and therefore, the response of the filters is higher.

## 6 Solving the Aperture Problem

### 6.1 Methodology

In order to determine until what scale of input stimuli FlowNetS can resolve the aperture problem, three different versions of this network are trained under the same circumstances with varying receptive field sizes. The receptive field size is defined as the region in the input images which affects the value of the feature map at a particular layer and feature map location. Therefore, we modify the filter size of the convolutional kernels in c6, which is actually composed of two layers: c6_0 and c6_1. The original (and our) FlowNetS uses 3x3 kernels in these layers, which leads to a receptive field size of 383 pixels in the f6 flow map. We train two additional models with kernels sizes (1x1, 3x3) and (1x1, 1x1) for c6_0 and c6_1, which we name FlowNetXS and FlowNetXXS, and whose f6 receptive field size is 255 pixels and 191 pixels, respectively. For the three of these networks, the receptive field size increases in the expanding part of the architecture due to the upconvolutional layers.

As input, we use a diagonally translating bar of different scales with motion magnitude $|\mathbf{u}| = 64$ pixels. We determine the error at the center of the bar, and at three flow maps of different resolutions: f6, f4, and f2 (see Fig. 1).

### 6.2 Results

In Fig. 14 (left), the FlowNetS response to a downward left translating bar of varying scale is shown. First, the flow becomes more and more refined in the expanding part of the architecture. Second, the network is able to extrapolate motion cues from the edges of the bar towards the center, but only to an extent determined by the scale of the bar.

Fig. 14 (right) shows the average End-Point-Error (EPE) of FlowNetS, FlowNetXS, and FlowNetXXS in response to two translating bars of different scales moving upward right and downward left, respectively. As shown, the network's robustness to the aperture problem is related to the receptive field size, and networks with larger receptive fields are able to resolve the aperture problem at larger scales.

## 7 Discussion and Future Work

### 7.1 Impact on Computer Vision

Our results show that the neural responses in the deepest encoding layer of FlowNetS, c6, are well captured by Gabor-like filters. This finding provides insight into the limits and robustness of the approach. Given this core mechanism for estimating optical flow, it is to be expected that the network generalizes quite well to out-of-training-set samples. However, it also raises some concerns, since traditional Gabor filters for optical flow estimation had certain disadvantages. They deal badly with deviations from translation, varying contrast due to changing lighting conditions, and are subject to the uncertainty relation, which corresponds to the balance between localization of the stimuli in the spatial domain and resolution in the frequency domain.

FlowNetS successfully copes with all of these issues. We have shown that deviations from translations are dealt with by additional filters that are sensitive to more complex motion types. Moreover, *Mayer et al.* [61] showed that FlowNet is able to cope with varying contrast over time due to changing lighting conditions. Lastly, we have demonstrated that FlowNetS is able to achieve a better spatial localization of motion cues in the expanding part of the network, thus coping with the uncertainty relation.

In terms of accuracy, FlowNetS did not reach the levels of state-of-the-art methods. For example, it has poor performance on sub-pixel flow [19]. One reason for this might be the large number of strides utilized before the initial flow prediction is made. Also, our analysis shows that a Gabor filter based on two frames results in a large temporal frequency bandwidth, and hence limited performance concerning flow velocity estimation. This is narrowed somewhat by the non-linear transformations due to the ReLU activation function and bias term. However, our analysis indicates that this could be further improved by using more frames and thus providing more temporal information to the network. Please note that there is an increasing number of multi-frame methods for deep optical flow estimation, e.g., [62], [63], [64], [65]. As remarked in [62], most of these methods use multiple images in order to track flow to future frames and track flow back to the past, in order to enhance consistency of the flow. Methods such as StarFlow [65] additionally pass the flow and extracted features from the previous image pairs as input to the deep net, while other methods make use of LSTMs [64]. However, the basic matching still happens between two frames with FlowNetC-like neural correlation blocks. What we propose here is to enter multiple images directly into a FlowNetS-like network in order to reduce the temporal bandwidth, something which to our knowledge has not been investigated yet.

The Gabor-like nature of the neural filters in c6 may also be a reason for less accuracy; These responses are mapped to coarse flow in a linear way by pf6. This means that optical flow velocities that are higher than the filter's tuned velocity, actually lead to an underestimation of the optical flow (due to the bell-shape of the response, see, e.g., Fig. 4). The network likely copes with this in the following ways. First, it can narrow the response bandwidth with the nonlinear activation function. To see why this helps, think of the extreme in which a neuron would respond in a Dirac-like way to a very specific optical flow velocity. Of course, such a narrow response would then require a very large number of neurons to cover all optical flow velocities. This brings us to the second coping mechanism; The final flow is mostly determined by the neurons in the neural filter bank that are tuned closer to the true optical flow velocity, as they will react more intensely. Finally, the biases in the network can be set in a way to deal with this problem, which is biased since it mostly involves underestimation. Still, it may be worth investigating if different mechanisms would lead to a better accuracy, for instance by introducing a winner-take-all mechanism.

We observed that only 592 of the 1024 c6 filters have an activation larger than zero. However, the high similarity of the active filters to the Gabor model already suggests that it would also be worth studying a hybrid FlowNetS network, in which there is a fixed Gabor filter bank (extended with rotation and dilation features) followed by a convolutional multi-layer loss flow refinement. This would greatly reduce training time, and, most probably, improve the generalizability of the network.

Finally, our findings for FlowNetS may also be relevant to "PoseNets" (e.g., [66], [67]) that take as input subsequent images and output the relative pose, i.e., an estimate of the translation and rotation between them. Typically, for such relative pose estimation networks a simple encoder structure is used, which is very similar to FlowNetS's structure up to and including c6. We expect that optical flow plays a large role in the estimation of translation and rotation between subsequent images, and - given the similar network structure - it is possible that PoseNets also implicitly determine flow with Gabor-like filters before synthesizing the information into a translation and rotation estimate.

### 7.2 Impact on Biology

We have used and extended methods from neuropsychology for determining the types of motion filters represented by neurons in the deep c6 layer of FlowNetS. The analysis gave very similar results to those on neurons in the mammalian visual cortex. First, many filter responses fit very accurately with Gabor filters that capture translational motion. Second, the spatial and orientation bandwidth statistics show similarity to bandwidths of neurons found in the mammalian visual cortex. We report a median spatial frequency bandwidth of 1.36 octaves, while *De Valois et al.* [68] report 1.4 octaves for the macaque visual cortex. Similarly, we find a median orientation bandwidth of 52 degrees, while *De Valois et al.* [69] find 65 degrees. These similarities may be due to similar optical flow statistics being perceived both by the network and the animals. Third, as in neuropsychological experiments [47], we observed that some filters respond poorly to translating plane waves. Our analysis shows that such poor response may be due to the filters being sensitive to more complex motions such as dilation and rotation. Indeed, in the human brain, channels sensitive to dilation have been found [70]. However, this did not provide conclusive evidence of neurons sensitive to dilation. Our analysis and results suggest that it is worth looking for dilation- and rotation-sensitive neurons in animal brains. In fact, one could even extend the analysis to also check for shear, as this forms an additional basis for the flow field derivatives [71].

## 8 Conclusion

We have employed a spectral response fitting approach from neuropsychology to demonstrate that the deepest layer of FlowNetS essentially encodes a bank of spatiotemporal Gabor filters. Although accurate fits were obtained, the spectral response fitting approach is limited, since it is not able to identify the exact motion pattern causing the maximum activation of a filter. In this work, we have already shown that the network also contains a large number of filters that are more sensitive to dilation and rotation than to translation, but more complex motion filters may be present. Finally, we have studied how FlowNetS tackles the aperture problem. Our results suggest that, on the one hand, the receptive field size is highly correlated to the scale at which the network can resolve the aperture problem. On the other hand, the expanding part of the network allows to solve the aperture problem at slightly larger scales by performing a filling-in function similar to that in mammal vision systems.

Future work could: (i) perform a similar analysis on SpyNet [34], (ii) study the neural response to more complex motion patterns like compositions of affine and 3D motion, as present in more realistic synthetic training datasets (e.g., FlyingThings [72]), (iii) attempt to improve FlowNetS' performance by using smaller strides or more input images, and (iv) employ our extended spectral response fitting method to investigate if animal brains have dilation- and rotation-sensitive neurons as well.

## REFERENCES

[1] J. J. Gibson, *The Perception of the Visual World*. L. Carmichael, Ed. Boston, MA, USA: Houghton Mifflin, 1950.

[2] J. Feng, *Computational Neuroscience: A Comprehensive Approach*. London, U.K./Boca Raton, FL, USA: Chapman and Hall/CRC, 2003.

[3] A. Borst, J. Haag, and D. F. Reiff, "Fly motion vision," *Annu. Rev. Neurosci.*, vol. 33, no. 1, pp. 49–70, 2010.

[4] A. Borst and M. Helmstaedter, "Common circuit design in fly and mammalian motion vision," *Nat. Neurosci.*, vol. 18, no. 8, pp. 1067–1076, 2015.

[5] I. Kajo, A. S. Malik, and N. Kamel, "An evaluation of optical flow algorithms for crowd analytics in surveillance system," in *Proc. Int. Conf. Intell. Adv. Syst.*, 2017.

[6] G. C. H. E. de Croon, "Monocular distance estimation with optical flow maneuvers and efference copies: A stability-based strategy," *Bioinspiration Biomimetics*, vol. 11, no. 1, 2016, Art. no. 016004.

[7] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM Comput. Surv.*, vol. 27, no. 3, pp. 433–466, 1995.

[8] P. Anandan, "A computational framework and an algorithm for the measurement of visual motion," *Int. J. Comput. Vis.*, vol. 2, no. 3, pp. 283–310, 1989.

[9] A. Singh, *Optic Flow Computation: A Unified Perspective*. Los Alamitos, CA, USA: IEEE Computer Society Press, 1991.

[10] D. J. Heeger, "Optical flow using spatiotemporal filters," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 279–302, 1988.

[11] D. Fleet and A. Jepson, "Computation of normal velocity from local phase information," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 1989, pp. 379–386.

[12] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, no. 1–3, pp. 185–203, 1981.

[13] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. Joint Conf. Artif. Intell.*, 1981, pp. 674–679.

[14] D. Gabor, "Theory of communication," *J. Institution Elect. Engineers Part I: Gen.*, vol. 94, no. 73, pp. 58–58, 1945.

[15] S. Ullman, *The Interpretation of Visual Motion*. Cambridge, MA, USA: MIT Press, 1979.

[16] H. Zimmer, A. Bruhn, and J. Weickert, "Optic flow in harmony," *Int. J. Comput. Vis.*, vol. 93, no. 3, pp. 368–388, 2011.

[17] A. Dosovitskiy *et al.*, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2758–2766.

[18] Z. Tu *et al.*, "A survey of variational and CNN-based optical flow techniques," *Signal Process., Image Commun.*, vol. 72, pp. 9–24, 2019.

[19] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1647–1655.

[20] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Models matter, so does training: An empirical study of CNNs for optical flow estimation," pp. 1–15, 2018. [Online] Available: https://arxiv.org/abs/1809.05571.

[21] T.-W. Hui, X. Tang, and C. C. Loy, "A lightweight optical flow CNN – Revisiting data fidelity and regularization," pp. 1–13, 2019. [Online] Available: https://arxiv.org/abs/1903.07414.

[22] J. P. Jones, A. Stepnoski, and L. A. Palmer, "The two-dimensional spectral structure of simple receptive fields in cat striate cortex," *J. Neurophysiol.*, vol. 58, no. 6, pp. 1212–1232, 1987.

[23] D. G. Albrecht, R. L. De Valois, and L. G. Thorell, "Visual cortical neurons: Are bars or gratings the optimal stimuli?," *Science*, vol. 207, pp. 88–90, 1980.

[24] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex," *J. Neurophysiol.*, vol. 58, no. 6, pp. 1233–1258, 1987.

[25] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman, "Receptive-field dynamics in the central visual pathways," *Trends Neurosci.*, vol. 18, no. 10, pp. 451–458, 1995.

[26] J. H. Van Hateren and D. L. Ruderman, "Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex," *Proc. Roy. Soc. London Ser. B Biol. Sci.*, vol. 265, no. 1412, 1998, pp. 2315–2320.

[27] B. A. Olshausen, "Learning sparse, overcomplete representations of time-varying natural images," in *Proc. IEEE Int. Conf. Image Process.*, 2003, pp. 41–44.

[28] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 25–36.

[29] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, Mar. 2011.

[30] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "EpicFlow: Edge-preserving interpolation of correspondences for optical flow," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1164–1172.

[31] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "DeepFlow: Large displacement optical flow with deep matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1385–1392.

[32] S. Zweig and L. Wolf, "InterpoNet, A brain inspired neural network for optical flow dense interpolation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6363–6372.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Interv.*, 2015, pp. 234–241.

[34] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2720–2729.

[35] D. Teney and M. Hebert, "Learning to extract motion from videos in convolutional neural networks," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 412–428.

[36] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5574–5584.

[37] E. Ilg *et al.*, "Uncertainty estimates and multi-hypotheses networks for optical flow," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 652–667.

[38] A. Ranjan, J. Janai, A. Geiger, and M. J. Black, "Attacking optical flow," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2404–2413.

[39] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014. [Online]. Available: https://arxiv.org/abs/1412.6806

[40] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[41] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *Univ. Montreal*, vol. 1341, no. 3, 2009, Art. no. 1.

[42] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, 2017. [Online]. Available: https://distill.pub/2017/feature-visualization

[43] A. Mordvintsev, "Inceptionism: Going deeper into neural networks," 2015. [Online]. Available: https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

[44] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 3395–3403.

[45] D. Wei, B. Zhou, A. Torrabla, and W. Freeman, "Understanding intra-class knowledge inside CNN," 2015. [Online]. Available: https://arxiv.org/abs/1507.02379

[46] R. N. Bracewell and R. N. Bracewell, *The Fourier Transform and its Applications*. New York, NY, USA: McGraw-Hill, 1986.

[47] G. C. Deangelis, I. Ohzawa, and R. D. Freeman, "Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development," *J. Neurophysiol.*, vol. 69, no. 4, pp. 1091–117, 1993.

[48] N. Petkov and E. Subramanian, "Motion detection, noise reduction, texture suppression, and contour enhancement by spatiotemporal Gabor filters with surround inhibition," *Biol. Cybern.*, vol. 97, no. 5/6, pp. 423–439, 2007.

[49] L. A. Palmer and T. L. David, "Receptive-field structure in cat striate cortex," *J. Neurophysiol.*, vol. 46, no. 2, pp. 260–276, 1981.

[50] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman, "Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. II. Linearity of temporal and spatial summation," *J. Neurophysiol.*, vol. 69, no. 4, pp. 1118–1135, 1993.

[51] J. P. Jones and L. A. Palmer, "The two-dimensional spatial structure of simple receptive fields in cat striate cortex," *J. Neurophysiol.*, vol. 58, no. 6, pp. 1187–1211, 1987.

[52] H. Komatsu, "The neural mechanisms of perceptual filling-in," *Nat. Rev. Neurosci.*, vol. 7, no. 3, pp. 220–231, 2006.

[53] R. Von Der Heydt, H. S. Friedman, and H. Zhou, *Filling-In: From Perceptual Completion to Cortical Reorganization*. London, U.K.: Oxford Univ. Press, 2003, pp. 106–127.

[54] J. Poort, F. Raudies, A. Wannig, V. A. Lamme, H. Neumann, and P. R. Roelfsema, "The role of attention in figure-ground segregation in areas V1 and V4 of the visual cortex," *Neuron*, vol. 75, no. 1, pp. 143–156, 2012.

[55] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3367–3375.

[56] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 611–625.

[57] J. O. V. Touryan, *Nonlinear Analysis of Complex Cells in Primary Visual Cortex*. Berkeley, CA, USA: Univ. California Press, 2004.

[58] Y. X. Yuan, "A review of trust region algorithms for optimization," in *Proc. Int. Congr. Ind. Appl. Math.*, 2000, pp. 271–282.

[59] S. S. Beauchemin and J. L. Barron, "The frequency structure of one-dimensional occluding image signals," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 2, pp. 200–206, Feb. 2000.

[60] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *Int. J. Comput. Vis.*, vol. 92, no. 1, pp. 1–31, 2011.

[61] N. Mayer *et al.*, "What makes good synthetic training data for learning disparity and optical flow estimation?" *Int. J. Comput. Vis.*, vol. 126, pp. 942–960, 2018.

[62] M. Neoral, J. Šochman, and J. Matas, "Continual occlusion and optical flow estimation," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 159–174.

[63] P. Liu, M. Lyu, I. King, and J. Xu, "SelFlow: Self-supervised learning of optical flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4571–4580.

[64] S. Guan, H. Li, and W.-S. Zheng, "Unsupervised learning for optical flow estimation using pyramid convolution LSTM," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2019, pp. 181–186.

[65] P. Godet, A. Boulch, A. Plyer, and G. L. Besnerais, "StarFlow: A spatiotemporal recurrent cell for lightweight multi-frame optical flow estimation," 2020, *arXiv: 2007.05481*.

[66] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1851–1858.

[67] Y. Chen, C. Schmid, and C. Sminchisescu, "Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7063–7072.

[68] R. L. De Valois, D. G. Albrecht, and L. G. Thorell, "Spatial frequency selectivity of cells in macaque visual cortex," *Vis. Res.*, vol. 22, no. 5, pp. 545–559, 1982.

[69] R. L. De Valois, E. WilliamYund, and N. Hepler, "The orientation and direction selectivity of cells in macaque visual cortex," *Vis. Res.*, vol. 22, no. 5, pp. 531–544, 1982.

[70] D. Regan and K. I. Beverley, "Looming detectors in the human visual pathway," *Vis. Res.*, vol. 18, no. 4, pp. 415–421, 1978.

[71] H. C. Longuet-Higgins and K. Prazdny, "The interpretation of a moving retinal image," *Proc. Roy. Soc. London Ser. B Biol. Sci.*, vol. 208, no. 1173, pp. 385–397, 1980.

[72] N. Mayer *et al.*, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4040–4048.

**David B. de Jong** received the BSc degree in aerospace engineering from the Delft University of Technology, Delft, The Netherlands, in 2016, and the MSc degree in aerospace engineering from the Control and Simulation Department, Delft University of Technology, Delft, in 2020. His research interests include the intersection of machine learning, neuroscience, and computer vision.

**Federico Paredes-Vallés** received the BSc degree in aerospace engineering from the Polytechnic University of Valencia, Valencia, Spain, in 2015, and the MSc degree from the Delft University of Technology, Delft, The Netherlands, in 2018. He is currently working toward the PhD degree at the Micro Air Vehicle Laboratory, Delft University of Technology, Delft, The Netherlands. His research interests include the intersection of machine learning, neuroscience, computer vision, and robotics.

**Guido C. H. E. de Croon** (Member, IEEE) received the MSc and PhD degrees in artificial intelligence from Maastricht University, Maastricht, The Netherlands. His research interest include computationally efficient algorithms for robot autonomy, with an emphasis on computer vision. Since 2008, he has worked on algorithms for achieving autonomous flight with small and light-weight flying robots, such as the DelFly flapping wing MAV. In 2011-2012, he was a research fellow with the Advanced Concepts Team of the European Space Agency, where he studied topics such as optical flow based control algorithms for extraterrestrial landing scenarios. Currently, he is full professor with the Delft University of Technology, The Netherlands, where he is the scientific lead of the Micro Air Vehicle Laboratory.