Optical Flow Estimation for Tiny Objects: New Problem, Specialized Benchmark, and Bioinspired Scheme

Xueyao Ji[1] , **Gang Wang**[1,2] and **Yizheng Wang**[1]
[1]Brain Research Center, Beijing Institute of Basic Medical Sciences, Beijing, China
[2]Chinese Institute for Brain Research (CIBR), Beijing, China
g_wang@foxmail.com

Abstract

Optical flow is pivotal in video-based tasks, yet existing methods mostly focus on medium-/large-size objects, while underperforming when characterizing the motion of tiny objects. To bridge this gap, we introduce the On-off Time-delay with Hassenstein-Reichardt correlator (OTHR), a computationally efficient scheme inspired by the primate visual cortex's direction selectivity mechanism. OTHR kernels, applied across multiple frames, discern bright/dark luminance changes along a specific direction over a time delay, effectively estimating motion of tiny objects amidst noise and static backgrounds. Notably, OTHR integrates seamlessly with leading deep learning flow estimation models such as RAFT and FlowFormer. We also propose refined evaluation metrics for tiny objects and contribute a new dataset featuring such objects to aid algorithm development. Our experiments confirm OTHR's superiority over competing methods, particularly in enhancing state-of-the-art models' performance on tiny object motion estimation at minimal cost. Specifically, for objects less than 100 pixels, OTHR reduces RAFT and FlowFormer's errors by 22.03% and 83.50%, respectively. The codes will be accessible at https://github.com/JaneEliot/OTHR.

1 Introduction

Physiologically, optical flow is interpreted as the motion features formed in the retina caused by moving objects [Gibson, 1951]. For animals, accurate optical flow estimation is critical when perceiving environments and evading predators. Also, with favorable applications such as object tracking and segmentation, optical flow estimation has been intensively studied over the past decades.

Horn and Schunck [1981] laid the foundation by proposing two classical constraints: constant brightness and smooth motion, thereby mathematically formulating optical flow estimation as an optimization problem that minimizes the sum of errors. In recent years, with the advancement of mathematical tools and deep networks, the performance of optical flow methods on large-scale datasets has been greatly improved. For example, FlowNet [Dosovitskiy *et al.*, 2015; Ilg *et al.*, 2016] and RAFT [Teed and Deng, 2020] have obtained favorable results on popular datasets.

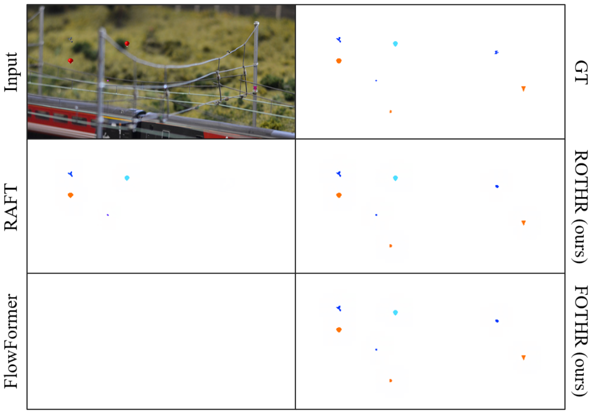


Figure 1. The challenge of tiny object optical flow estimation. The representative optical flow estimation methods underperform seriously when characterizing tiny objects.

Despite the progress outlined above, estimating the motion of tiny objects remains challenging for several reasons. First, tiny objects usually occupy less than 100 pixels and have indistinct edges. Second, they typically have weak textures and may appear as small blobs, blending into complex backgrounds. Third, some models, particularly during sampling and pooling operations, can easily overlook tiny objects. As illustrated in Figure 1, the two deep methods seriously underperform when characterizing the motion of tiny objects.

Intuitively, human visual systems can effortlessly detect flying insects passing through their visual fields; camouflage-covered animals are more likely to be noticed once they move. These examples demonstrate that moving objects are easier to detect [Franconeri and Simons, 2003], even if they are visually non-salient. Motion information facilitates visual recognition [Wexler *et al.*, 2001]. In the primary visual cortex (V1) of the primate visual system, the direction selectivity (DS) mechanism contributes to motion perception. Inspired by the functional properties of the DS, we design a multi-frame motion extraction scheme that integrates the dynamic responses of the On-off cell pathways, the intensity differences among Time-delayed frames, and the renowned

Hassenstein-Reichardt correlator [1956], namely the OTHR method. Instead of estimating the motion direction by solving equations, the OTHR determines local motion through parallel filtering. In addition, the OTHR employs multi-frame spatiotemporal features, thus effectively suppressing the interference of luminance variation and random noise.

The OTHR method mainly simulates the early stage (V1) of biological visual pathways, generating fine-grained sparse motion features with limited receptive fields. However, the mechanisms by which higher-level brain areas obtain detailed motion information remain biologically unclear. To address this issue, we integrate the OTHR with deep models such as RAFT to produce dense optical flow, utilizing OTHR as a cheap plug-and-play module and leveraging deep models to simulate higher-level brain functions. We then build a dataset featuring tiny objects to facilitate algorithm studies and performance evaluation. Using this customized dataset, we compare the OTHR-based methods with competing approaches. The main contributions of this work are as follows:

1) To address the limitations of existing optical flow methods in characterizing tiny objects, we propose a lightweight and non-learning OTHR method inspired by the biological DS mechanism, which can effectively determine motion using multi-frame cues.

2) To facilitate a reliable quantitative evaluation of tiny object optical flow estimation among various methods, we revise the original unitary evaluation metric EPE by distinguishing EPE_{obj}, EPE_{bg}, and $\text{EPE}_{\text{local}}$. In addition, we build a dataset named FlyingTO, featuring flying tiny objects against various real-world backgrounds.

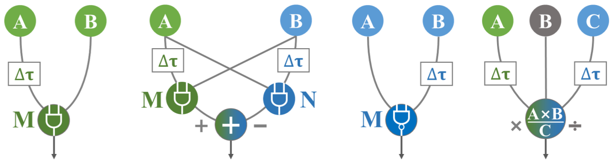
3) To improve the popular deep optical flow methods when characterizing tiny objects, we propose a scheme to integrate the OTHR into deep methods, e.g. RAFT and FlowFormer. To address the imbalance in estimation errors between tiny object areas and background areas in the traditional loss function, we propose to calculate the loss as a weighted sum of the loss_{obj} and the loss_{bg}. The experimental results confirm the efficacy and superiority of our methods.

2 Related Work

2.1 Non-dense Optical Flow

Sparse methods. The LK algorithm [Lucas and Kanade, 1981] is the most classic sparse optical flow method, requiring the motion of all pixels in a window to be consistent to solve the optical flow equations by associating multiple pixels. Its improved version [Bouguet, 2001] introduced a pyramid structure to handle large displacements, primarily addressing large objects.

Bioinspired methods. The motion information extracted by existing bioinspired motion detectors lies between sparse and dense optical flow, usually originating from high-frequency regions of images, such as edges. One of the most classic methods is the Hassenstein-Reichardt (HR) model [1956]. As shown in Fig. 2a, if an object moves from neuron A to neuron B, then both A and B will be activated and successively send signals to neuron M. Due to the delay module set between A and M, the signals sent by A and B will reach M simultaneously, then the multiplicative module in M will produce the enhanced response. This direction of motion in which the DS neuron M can produce the maximum response is defined as its preferred direction. In contrast, if an object moves in the null direction (i.e., B to A), the total response at M is 0. The HR model has been validated in many physiological experiments of flies [Borst and Euler, 2011; Borst and Helmstaedter, 2015], and neurons with discharge patterns like the HR model have also been found in amphibians, rodents, and primates [Borst and Egelhaaf, 1989].



(a) HR (b) HR correlator (c) BL (d) T4

Figure 2. Bioinspired methods.

Figure 2c presents the Barlow-Levick (BL) model [1965]. The delay module is between B and M. The *NAND* module in M will suppress the motion response if an object moves in the null direction. In short, the HR model enhances the response of motion along the preferred direction, while the BL model suppresses the response of motion along the null direction.

Subsequent bioinspired methods are mostly based on the ideas of HR and BL models, such as the HRC in Fig. 2b, and the T4 model in Fig. 2d inspired by the fly [Haag *et al.*, 2016].

2.2 Dense Optical Flow

Classical methods. Dense optical flow is valuable for video segmentation, tracking, etc. Most early proposed dense optical flow methods rely on the two classic constraints, e.g. FarnebackFlow [Farnebäck, 2003], which approximated the neighborhood of each pixel with a polynomial. Some methods treated dense optical flow estimation as an energy minimization problem, e.g. the HS method [1981] mentioned in Sec. 1, and its improved version, TVL1Flow [Pérez *et al.*, 2013], which replaced the L2 norm in the HS energy functional with the L1 norm to enhance robustness against noise. SimpleFlow [Tao *et al.*, 2012] obtained an energy functional based on the probabilistic representation of the motion of each pixel. DeepFlow [Weinzaepfel *et al.*, 2013] constructed an energy functional using the weighted sum of a data term, a smoothness term, and a matching term. The main weakness of these classic constraint-based methods is their reliance on the slowly changing displacement field.

Obtaining sparse matching first and then interpolating, or using a multiscale structure to obtain dense flow, are also common approaches in classical methods, such as the SimpleFlow mentioned before, PCAFlow [Wulff and Black, 2015], EpicFlow [Revaud *et al.*, 2015], DISFlow [Kroeger *et al.*, 2016], and RLOF [Senst *et al.*, 2012; Geistert *et al.*, 2016]. Matching approaches rely on significant image features that tiny objects typically lack, while interpolation operations may result in the loss of tiny objects.

Deep methods. Combining end-to-end CNN to obtain dense optical flow can be traced back to FlowNet [Dosovitskiy *et al.*, 2015], which fed two adjacent frames into CNN to obtain feature maps and utilized transposed convolution and correlation calculation to obtain optical flow. Many then-new methods introduced pyramid structures, such as SpyNet [Ranjan and Black, 2016] and PWCNet [Sun *et al.*, 2018]. This structure is effective in large displacement cases, while it does not have significant advantages for tiny object optical flow estimation. Methods at this stage are still striving to balance the accuracy, calculation speed, model complexity, labeled-data dependence, etc.

RAFT [Teed and Deng, 2020] is a representative deep learning-based method. Based on the features extracted by its encoders, it constructed 4D correlation volumes to compute visual similarity and recurrently updated the estimated flow through the GRU. AnyFlow [Jung *et al.*, 2023] used an implicit neural representation for optical flow estimation and captured small objects in low-resolution inputs. Some methods [Huang *et al.*, 2022; Lu *et al.*, 2023] introduced the attention mechanism into optical flow estimation. VideoFlow [Shi *et al.*, 2023] utilized multiple adjacent frames to estimate optical flow, but with considerable computational cost. Another multiframe method, SplatFlow [Wang *et al.*, 2024], was designed to handle the occlusion problem. AccFlow [Wu *et al.*, 2023] designed a deformable module to recursively backward accumulate local flows, solving the problems of long-range flow estimation. DIFT [Garrepalli *et al.*, 2023] extended an efficient correlation lookup approach from RAFT based on varying cost-volume resolution. MemFlow [Dong and Fu, 2024] embedded a memory module in a RAFT-like framework. Self-supervised or unsupervised optical flow methods [Huang *et al.*, 2023; Yuan *et al.*, 2024] are also a cutting-edge development trend. Among numerous recent studies, the attention paid to tiny objects is insufficient. Although some methods have begun to pursue improvements in optical flow estimation at the fine edges of objects, there is still a gap when it comes to tiny objects below 100 pixels.

3 Methodology

3.1 Primary Biological Visual System Model

Shape of the receptive field. In short, the retina is composed of three layers of cells, which are receptor cells, bipolar cells (BC) and ganglion cells (GC) [Shou, 2010]. Among them, the receptive fields of BC and GC both have a form of concentric circle antagonism (CCA). The CCA receptive field reflects the vital role of feedback in neural information processing, and it is also the neurophysiological basis for shape perception.

Rodieck [1965] proposed a mathematical model of CCA receptive fields, which consists of a small strong-excitatory center and a larger weak-inhibitory periphery. The two components both exhibit Gaussian distributions but have opposite polarities. The Rodieck model is also known as the difference of Gaussians (DoG) model.

Later studies [Cleland and Levick, 1974; Hammond, 1974; Levick and Thibos, 1982; Shou, 2010] show that the shape of the lateral geniculate nucleus (LGN) receptive fields in cats and monkeys is not circular but elliptical, which may be a critical step in the formation of DS in the visual cortex [Shou and Leventhal, 1989]. Leventhal and Schall [1983] believe that the elongated and oriented dendritic field distribution is the anatomical basis for the orientation sensitivity of GC. However, in this work, we still use DoG-style receptive fields to simplify.

Spatial kernels of the filter. In 2D space, DoG is an isotropic filter sensitive to contours but not to orientation. To obtain DS, there should be a displacement between the centers of its positive and negative Gaussian kernels. The asymmetric structure makes it an anisotropic filter sensitive to orientation.

The computational model is inspired by the study on macaque V1 [Chariker *et al.*, 2021; Chariker *et al.*, 2022]. The ON and OFF cells in LGN are simplified as OFF-ON pairs separated by a distance d, as shown in Fig. 3a. The OFF-ON pairs respond asymmetrically to motion stimuli in different directions, thus generating the DS. Specifically, the preferred direction of an OFF-ON pair is from OFF to ON, meaning that the pair will have the maximum response when an object moves in this direction. The receptive fields of both ON and OFF cells are in the form of DoG:

$$S(x,y) = \frac{\alpha}{\pi\sigma_\alpha^2}\cdot e^{-\frac{x^2+y^2}{\sigma_\alpha^2}} - \frac{\beta}{\pi\sigma_\beta^2}\cdot e^{-\frac{x^2+y^2}{\sigma_\beta^2}} \tag{1}$$

where α = 1.0, β = 0.74, σ_α = 0.0894, σ_β = 0.1259 [Zhu *et al.*, 2009]. Technically, we can flexibly adjust the parameters in applications. However, to align with the physiological experimental records, we retain the given parameters.

The differences between the two receptive fields are the opposite polarity and distinct spatial positions. For the OFF-ON pair in Fig. 3a, taking the midpoint of the line connecting the centers of two cells as the coordinate origin, the spatial kernels of OFF and ON cells have the following forms:

$$S_{\text{off}} = -S(x+\frac{d}{2}, y) \tag{2}$$

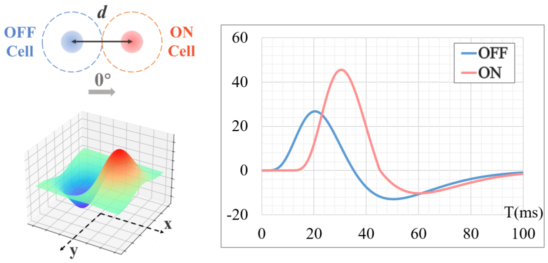
$$S_{\text{on}} = S(x-\frac{d}{2}, y) \tag{3}$$

Temporal kernels of the filter. Considering the time dimension, a filter sensitive to the motion direction should be anisotropic in 3D. Spatial filtering extracts static characteristics, such as contours and textures, while temporal filtering detects changes of this spatial information. An essential step for the OFF-ON pair to become a motion detector is the response delay of the ON cell relative to the OFF cell [Zhu *et al.*, 2009; Reid and Shapley, 2002]:

$$T_{\text{off}}(t) = \frac{t^6}{\tau_0^7}e^{-\frac{t}{\tau_0}} - \frac{t^6}{\tau_1^7}e^{-\frac{t}{\tau_1}} \tag{4}$$

$$T_{\text{on}}(t) = aT^+_{\text{off}}(t-t_0) + bT^-_{\text{off}}(t-t_0) \tag{5}$$

where τ_0 = 3.66 ms, τ_1 = 7.16 ms, and t_0 is the time delay, usually taken 9 $\sim$ 11 ms; $T^+(t) = \max\{0, T(t)\}$ and $T^-(t) = \min\{0, T(t)\}$ represent the positive and negative



(a) Spatial kernels (b) Temporal kernels

Figure 3. The spatial and temporal views of the designed On-off Time-delay kernel inspired by the primate DS mechanism.

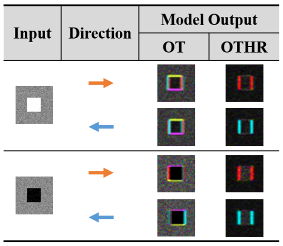


Figure 4. The proposed OTHR model obtains more reasonable and accurate motion characteristics compared to the OT model.

parts of $T(t)$, respectively. According to Reid and Shapley [2002], the distributions of a and b are shown in Tab. 1. Figure 3b gives an example with a = 1.7 and b = 0.8.

A spatiotemporal kernel with DS can be obtained by:

$$K_{\text{DS}} = S_{\text{off}}T_{\text{off}} + S_{\text{on}}T_{\text{on}} \tag{6}$$

However, the anisotropy of such a kernel not only contributes to motion detection but also leads to a problem. ON cells are sensitive to an increase in brightness, whereas OFF cells are the opposite. Picture a white square moving to the right on a gray background, of which the leading and trailing edges give rise to light-on and light-off responses, respectively. The kernel K_{DS} acts inconsistently on these two edges, which is not the desired result of the direction estimation.

Combination with HR correlator. To solve the above problem, we propose to combine this On-off Time-delay (OT) model with the Hassenstein-Reichardt correlator (HRC):

$$\begin{aligned} R_{\text{DS}} =& K_\tau^A \times K^B - K^A \times K_\tau^B \\ =& (S_{\text{off}}T_{\text{on}}) \times (S_{\text{on}}T_{\text{off}}) - (S_{\text{off}}T_{\text{off}}) \times (S_{\text{on}}T_{\text{on}}) \end{aligned} \tag{7}$$

where K^A and K^B represent the spatiotemporal filters of cells A and B; τ marks a filter with time delay. The structure of HRC is shown in Fig. 2b. It combines the preferred-direction enhancement and the null-direction suppression. Here is a simple explanation of the mechanism of HRC. Use value 1 to indicate a positive response. When the HRC perceives a stimulus moving along its preferred direction (that is, from A to B), its total response is equal to 1 (1 minus 0), and vice versa is -1 (0 minus 1); when there is no stimulus, its total response is 0. This On-off Time-delay method that combines HRC is called OTHR. As shown in Fig. 4, the OTHR obtains the consistent motion saliency on the light-on and light-off edges that move in the same direction. To better display, the areas with optical flow values of 0 are visualized as black. In Sec. 4, the experimental results will verify the advantages of the scheme in Fig. 2b over that of Fig. 2d.

(a, b)	(1.7, 0.8)	(1.6, 0.7)	(1.1, 0.5)	(1.0, 0.4)
Probability	10%	30%	30%	30%

Table 1. Distribution of parameters a and b in the Eq. (5).

3.2 High-level Visual Model

The bioinspired method in Sec. 3.1 mainly simulates the functions of the retina and V1, which belong to the early stage of the biological visual system. The neuroscience field has confirmed that motion information is processed and transmitted among different levels of the visual system [Dai *et al.*, 2025], and full semantic motion cognition occurs at the higher level of the brain areas, such as the medial superior temporal cortex. Consequently, the motion characteristics yielded by the OTHR are considered low-level and should be further processed. How higher-level brain areas perceive motion remains unknown; nevertheless, deep models in CV fields may be able to functionally simulate this process. In this work, we employ the widely used deep optical flow models, e.g. RAFT and FlowFormer, to assist the OTHR yield semantic optical flow results. Instead of studying all SOTA models, we select these two representative CNN- or Transformer-based models mainly because we intend to assess how the deep models perform in characterizing tiny objects and to explore how the cheap bioinspired OTHR benefits deep models.

The method of combining our cheap plug-and-play module with deep models is simple. The primary motion information extracted by OTHR is concatenated with the input images in the channel dimension, and the images after channel expansion are fed into deep models to extract full semantic motion information. The architecture is shown in Fig. 5.

Additionally, we notice that the loss function commonly used by deep optical flow models equally calculates the estimation errors at each pixel, thus tending to focus on large objects. However, when characterizing tiny objects, the traditional loss function is far from reasonable. Due to the small proportion of tiny object areas in the entire image, the fluctuation of estimation errors in these areas is often too slight to be noticeable. Therefore, such a loss function is less capable of guiding the model to be optimized correctly. The improved version distinguishes between background loss and object loss and then assigns different weights:

$$\mathcal{L}_w = \sum_{i=1}^{N} \gamma^{N-i} \|\alpha(\mathbf{f}_{\text{gt}}^{\text{bg}} - \mathbf{f}_i^{\text{bg}}) + \beta(\mathbf{f}_{\text{gt}}^{\text{obj}} - \mathbf{f}_i^{\text{obj}})\|_1 \tag{8}$$

where N is the number of iterations of the recurrent unit in deep models; $\gamma < 1$, assigning different weights to different iterations; generally, $\alpha < \beta$ and α is set to 1 by default. Subsequent experiments show that the improved weighted loss function effectively reduces the EPE in the estimation of optical flow for tiny objects.

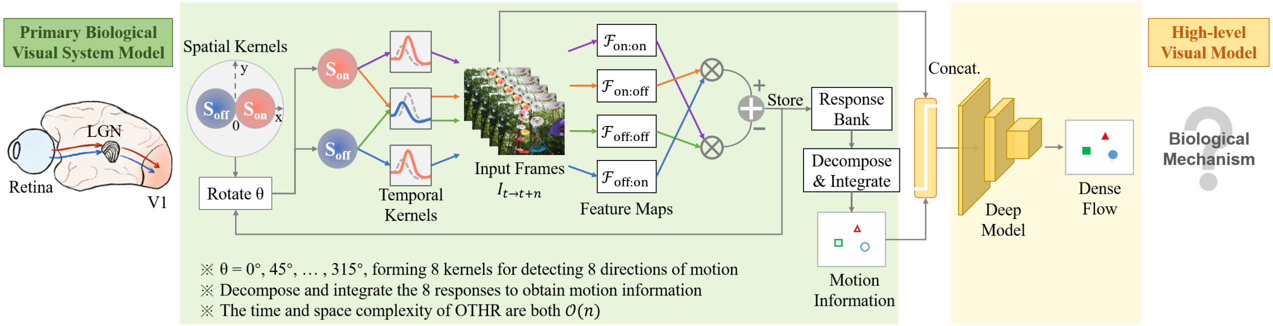


Figure 5. Illustration of the proposed architecture that combines the deep model with the bioinspired OTHR.

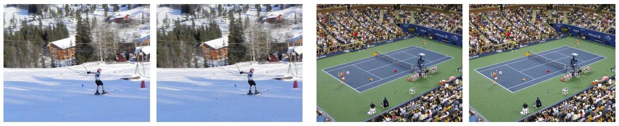
4 Experiments

4.1 Dataset Construction

Given the scarcity of optical flow datasets for tiny objects, we build FlyingTO, a dataset that contains tiny objects flying randomly against complex backgrounds, as shown in Fig. 6. The FlyingTO contains 320 videos for training and 100 videos for testing. The frame resolution is 512×640 Px. Each video has a duration of 2.4s with a frame rate of 25 FPS. Based on the clean FlyingTO videos, a noisy version is established by adding Gaussian noise, salt-pepper noise, and a change of brightness, which can be used for evaluating the algorithms' robustness against noise. In FlyingTO, 50% of objects have sizes less than 100 pixels, while the sizes of the other 50% range from 100 to 400 pixels. In addition, flying objects have diverse shapes, including polygons, circles, stars, and irregular real-world object appearances (e.g., birds, balloons, drones). The number of objects in each video is randomly selected between 1 and 9. The backgrounds are selected from the COCO [Lin *et al.*, 2014] and AntiUAV [Jiang *et al.*, 2021] datasets to simulate real-world scenarios. To demonstrate our method's feasibility, a synthetic dataset suffices. Building a costly real-world dataset is planned for the future.

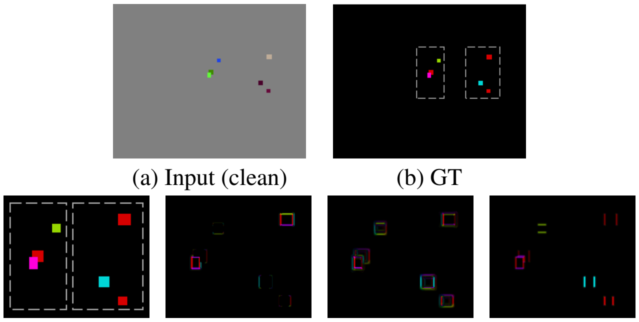
4.2 Evaluation Metrics

Similarly, we extend the end-point-error (EPE) to EPE, EPE_{obj}, EPE_{bg}, and $\text{EPE}_{\text{local}}$, following the idea of optimizing the loss function. The former three represent the overall, object, and background EPE, respectively. When calculating the $\text{EPE}_{\text{local}}$, we determine a dilated bounding box for each tiny object and subsequently calculate the EPE within all boxes as the $\text{EPE}_{\text{local}}$. The side length of each dilated bounding box is approximately 20 $\sim$ 50 pixels.



(a) Example 1 (clean) (b) Example 1 (noisy) (c) Example 2 (clean) (d) Example 2 (noisy)

Figure 6. Example frames in the FlyingTO dataset. The flying object sizes in (a) and (b) are less than 100 pixels, while their sizes in (c) and (d) are between 100 $\sim$ 400 pixels.



(a) Input (clean) (b) GT

(c) Enlarged GT (d) BL (e) T4 (f) OTHR (ours)

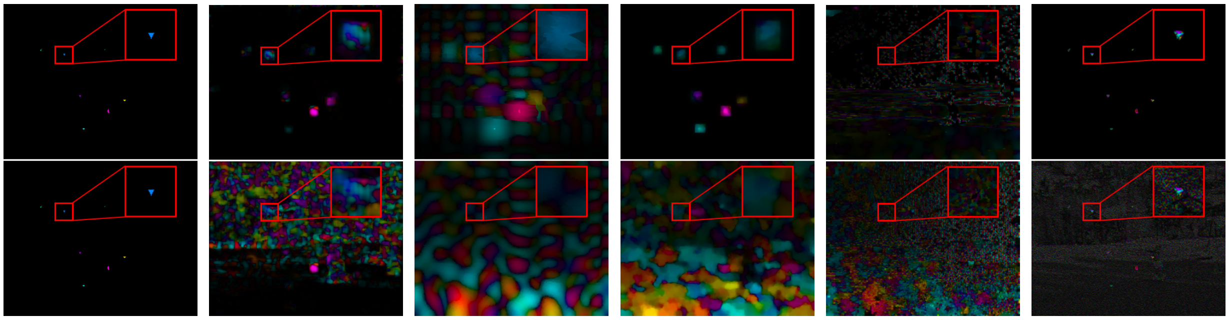
Figure 7. Bioinspired methods vs. OTHR. The gray dashed boxes in (c) represent the cropped areas corresponding to (b).

4.3 Bioinspired Methods vs. OTHR

As illustrated in the previous section, bioinspired methods usually perform well at object edges that are perpendicular to the direction of motion. Therefore, we first compare the OTHR with the aforementioned BL and T4 models on a video that contains moving rectangles, as in Fig. 7a. Compared to BL and T4, the OTHR achieves consistent motion saliency on the leading and trailing edges, as shown in Fig. 7f, obtaining more reasonable results than competing methods. For the BL model, the difference in contrast between objects and backgrounds affects its motion extraction. On the right side of Fig. 7c, there are two rectangles with red optical flow; in Fig. 7a, one of these two rectangles is lighter than the background while the other is darker, leading to inconsistent results of edge motion extraction, as shown in Fig. 7d. In the result of T4, as shown in Fig. 7e, there are many blurry shadows around the objects, indicating errors in the estimation of motion near the edges.

4.4 Classical Methods vs. OTHR

We compare the OTHR with some classical optical flow methods that are widely used. The quantitative and qualitative comparisons obtained on the FlyingTO dataset are shown in Tab. 2 and Fig. 8, respectively. OTHR is proposed to extract sparse motion information from edges, while many classical methods were designed for dense optical flow. Therefore, it can be anticipated that OTHR underperforms compared to these classical methods in terms of EPE_{obj}. How-



(a) GT (b) FarnebackFlow (c) PCAFlow (d) DISFlow (e) RLOF (f) OTHR (ours)

Figure 8. Classicial methods vs. OTHR. The first row: the GT and the estimated optical flow results yielded by different methods on the clean frame 6a; the second row: the results obtained on the noisy frame 6b. Please refer to the digital version for better visualization.

ever, it can be observed in Fig. 8 that some classical methods show limited performance in estimating the optical flow of tiny objects. Among the results of the competing methods, concurrent estimation errors occur in the areas around tiny objects, especially the PCAFlow, yielding significant optical flow estimation errors in the background, whereas the OTHR can accurately characterize the motion. This is also certified by quantitative comparisons, as the OTHR obtains the smallest $\text{EPE}_{\text{local}}$ among all the methods tested.

4.5 Deep Models vs. Versions with OTHR

This work aims to demonstrate the feasibility of our method for tiny object flow estimation, not to minimize EPE. Hence, we choose widely tested representative deep models, i.e., CNN-based RAFT and Transformer-based FlowFormer. We revise the two models by integrating the OTHR into their networks, thus having ROTHR (RAFT+OTHR) and FOTHR (FlowFormer+OTHR). RAFT, ROTHR, FlowFormer and FOTHR are pretrained on the FlyingChairs [Dosovitskiy *et al.*, 2015] and FlyingThings [Mayer *et al.*, 2016] datasets to obtain baseline performance. Subsequently, the four methods are finetuned by training on the FlyingTO. To evaluate the contributions of the weighted loss $\mathcal{L}_w$ and OTHR method, respectively, we perform different groups of ablation experiments. The results are reported in Tab. 2 and are visually shown in Fig. 9.

Although the original RAFT and FlowFormer can hardly describe the tiny moving objects, the $\mathcal{L}_w$ and OTHR contribute to the performance improvement of these deep models. The complete combination of $\mathcal{L}_w$, OTHR, and deep models can achieve the best performance. As shown in Fig. 9, $\mathcal{L}_w$ and OTHR help deep models estimate the optical flow of moving edges more accurately and reduce estimation errors for tiny objects. More specifically, from top to bottom of Fig. 9b, the second (RAFT+$\mathcal{L}_w$) and the third (RAFT+OTHR) results can both characterize more tiny objects than the first (original

Model	Clean				Noise			
	EPE	EPE_{obj}	EPE_{bg}	$\text{EPE}_{\text{local}}$	EPE	EPE_{obj}	EPE_{bg}	$\text{EPE}_{\text{local}}$
FarnebackFlow	0.00224	0.68513	0.00196	0.15667	0.06738	0.70529	0.06711	0.18904
PCAFlow	0.00703	0.71198	0.00674	0.12259	0.26570	0.74971	0.26550	0.23830
DISFlow	0.00182	0.67248	0.00154	0.14482	0.32115	0.68456	0.32100	0.27445
RLOF	0.12162	0.83850	0.12132	0.16236	0.52452	0.89321	0.52437	0.44707
OTHR (ours)	0.00070	0.82754	0.00036	0.06577	0.06821	0.82403	0.06790	0.12099
RAFT△	0.02533	0.74099	0.02503	0.06982	0.08573	0.76772	0.08545	0.10330
RAFT	0.00020	0.25072	0.00010	0.01836	0.00022	0.27382	0.00011	0.01962
RAFT (β=2)	0.00020	0.23293	0.00011	0.01838	0.00023	0.25013	0.00013	0.01946
ROTHR	0.00017	0.20934	0.00009	0.01584	0.00022	0.26405	0.00011	0.01911
ROTHR (β=2)	0.00017	0.19548	0.00009	0.01573	0.00022	0.24610	0.00012	0.01897
FlowFormer	0.00048	0.77245	0.00016	0.04582	0.00048	0.77245	0.00016	0.04582
FlowFormer (β=2)	**0.00012**	0.13592	0.00007	0.01142	0.00014	**0.14927**	0.00009	0.01246
FOTHR	**0.00012**	0.12936	**0.00006**	**0.01062**	**0.00013**	0.15105	**0.00007**	**0.01207**
FOTHR (β=2)	**0.00012**	**0.12742**	0.00007	0.01088	0.00014	0.14959	0.00008	0.01241

Table 2. The evaluation results of different methods obtained on the FlyingTO dataset (object size < 100 pixels). The symbol △ indicates that the method is tested using the officially provided weights without finetuning on our dataset. The best results are marked in bold.

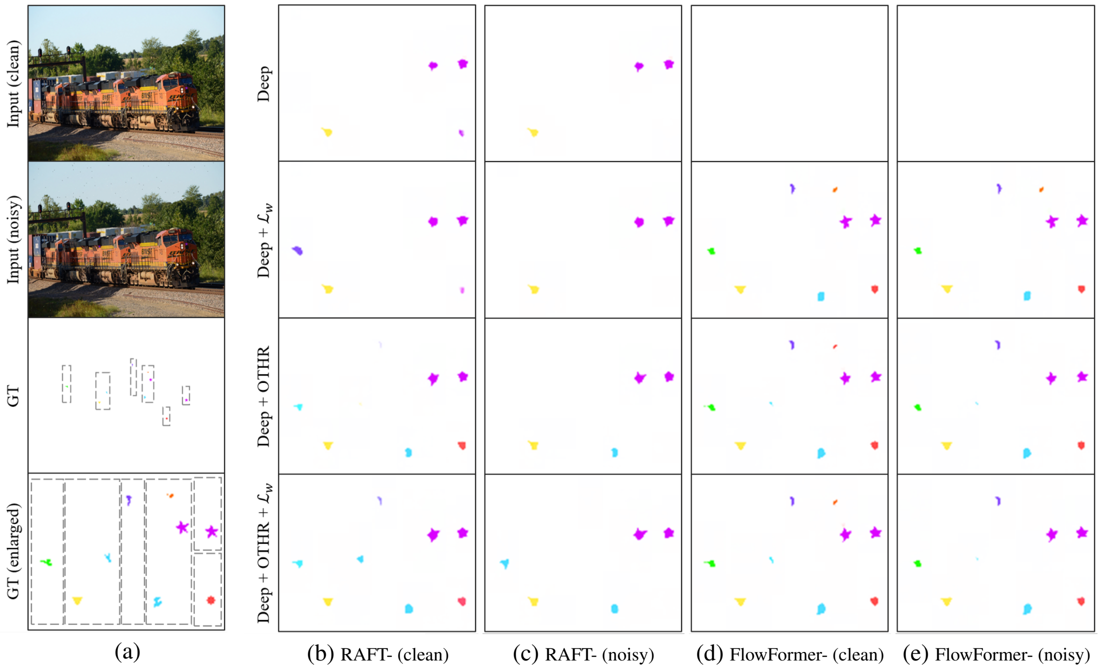


Figure 9. The estimated optical flow obtained by two popular deep methods and the versions integrated with the weighted loss $\mathcal{L}_w$ or the OTHR model. The original deep RAFT and FlowFormer have limitations in tackling tiny objects and noise. Benefiting from the $\mathcal{L}_w$ and OTHR, the performance of these two deep methods in characterizing tiny moving objects is improved, even in noisy frames. (a) Inputs and GT in which the object sizes are less than 100 pixels. (b) $\sim$ (e): The zoomed-in version of the optical flow results yielded by the RAFT-based and FlowFormer-based methods on clean and noisy frames. Gray dashed boxes represent the locally zoomed-in cropped areas from the GT. Areas with optical flow values of 0 are set as white.

RAFT). The bottom one (ROTHR+$\mathcal{L}_w$) can yield the optical flow of all tiny objects, except for the loss of an extremely small object in the upper right corner and the inaccurate estimation of the optical flow of the leftmost object (there is a slight difference compared to GT). In Fig. 9c, due to noise interference, the results are not as good as in Fig. 9b, but our method performs much better than the original. According to Fig. 9d and Fig. 9e, we surprisingly find that although the original FlowFormer fails to estimate the optical flow for tiny objects, our schemes, including the $\mathcal{L}_w$ and OTHR, help the FlowFormer accurately characterize tiny moving objects.

For the quantitative results in Tab. 2, one may doubt that the EPE- values of many tested methods are rather low, while the corresponding visual results are unsatisfactory. This is plausible for the following reason. Compared with large objects and their large displacements, tiny objects in FlyingTO generally have sparse distributions, tiny sizes, and low velocities. Thus, we might obtain rather small EPE values even if tiny objects were neglected. We highly suggest evaluating performance by distinguishing EPE_{obj}, $\text{EPE}_{\text{local}}$, etc.

The results in Tab. 2 reveal that the classical methods and original deep models struggle to yield accurate optical flow for tiny objects, as they all result in comparatively larger EPE_{obj}. With the help of the FlyingTO dataset and the proposed $\mathcal{L}_w$, the performance of RAFT and FlowFormer has been significantly improved. For example, after finetuning, the EPE_{obj} (clean) of RAFT decreases by 66.2%. Furthermore, the proposed cheap OTHR method can benefit both the deep models, reducing EPE_{obj} greatly. More details and experimental results are reported in the supplementary material.

5 Conclusions

Inspired by the direction selectivity mechanism in primate visual systems [Chariker *et al.*, 2021; Chariker *et al.*, 2022], we have designed a robust spatiotemporal OTHR method to characterize motion information for tiny objects. A new dataset has been built to facilitate the design and evaluation of optical flow methods for tiny objects. We have revised traditional optical flow evaluation metrics to be more reasonable by jointly considering object areas and local regions. To obtain semantic dense optical flow, we have designed schemes of characterizing tiny moving objects by combining the cheap bioinspired OTHR with the popular deep RAFT and FlowFormer. We have proposed to use the weighted sum of background loss and object loss to calculate the final loss. The experimental results have validated the superiority of our proposed method over competing approaches. In particular, popular deep methods can improve their performance in characterizing tiny moving objects with benefits from the cheap OTHR model, even in noisy frames.

Nevertheless, as aforementioned, our OTHR method functionally simulates the early motion processing stage of primate visual systems. This bioinspired model is far from the full motion-processing mechanism of primate brains. We hope to further develop the framework with the help of neuroscience under the NeuroAI paradigm.

Acknowledgements

The authors would like to thank Prof. Dajun Xing (BNU, China), Prof. Jiayi Zhang (FUDAN, China), Dr. Liang Li (BIBMS, China) for their constructive suggestions. This work is sponsored by Beijing Nova Program (2022038, 20240484703).

Contribution Statement

Xueyao Ji and Gang Wang contributed equally to this work (methodology, software, validation, writing, visualization, review, editing). Gang Wang and Yizheng Wang are the co-corresponding authors (supervision, funding acquisition).

References

[Barlow and Levick, 1965] H. B. Barlow and W. R. Levick. The mechanism of directionally selective units in rabbit's retina. *The Journal of Physiology*, 178(3):477–504, 1965.

[Borst and Egelhaaf, 1989] Alexander Borst and Martin Egelhaaf. Principles of visual motion detection. *Trends in Neurosciences*, 12(8):297–306, 1989.

[Borst and Euler, 2011] Alexander Borst and Thomas Euler. Seeing things in motion: models, circuits, and mechanisms. *Neuron*, 71(6):974–994, 2011.

[Borst and Helmstaedter, 2015] Alexander Borst and Moritz Helmstaedter. Common circuit design in fly and mammalian motion vision. *Nature Neuroscience*, 18(8):1067–1076, 2015.

[Bouguet, 2001] Jean-Yves Bouguet. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel Corporation*, 2001.

[Chariker *et al.*, 2021] Logan Chariker, Robert Shapley, Michael Hawken, and Lai-Sang Young. A theory of direction selectivity for macaque primary visual cortex. *Proceedings of the National Academy of Sciences*, 118(32):e2105062118, 2021.

[Chariker *et al.*, 2022] Logan Chariker, Robert Shapley, Michael Hawken, and Lai-Sang Young. A computational model of direction selectivity in macaque v1 cortex based on dynamic differences between on and off pathways. *Journal of Neuroscience*, 42(16):3365–3380, 2022.

[Cleland and Levick, 1974] B. G. Cleland and W. R. Levick. Brisk and sluggish concentrically organized ganglion cells in the cat's retina. *The Journal of Physiology*, 240(2):421–456, 1974.

[Dai *et al.*, 2025] Weifeng Dai, Tian Wang, Yang Li, Yi Yang, Yange Zhang, Yujie Wu, Tingting Zhou, Hongbo Yu, Liang Li, Yizheng Wang, Gang Wang, and Dajun Xing. Cortical direction selectivity increases from the input to the output layers of visual cortex. *PLOS Biology*, 23(1):e3002947, 2025.

[Dong and Fu, 2024] Qiaole Dong and Yanwei Fu. Memflow: Optical flow estimation and prediction with memory. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19068–19078, 2024.

[Dosovitskiy *et al.*, 2015] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: learning optical flow with convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015.

[Farnebäck, 2003] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In Josef Bigun and Tomas Gustavsson, editors, *Image Analysis*, pages 363–370. Springer Berlin Heidelberg, 2003.

[Franconeri and Simons, 2003] Steven L. Franconeri and Daniel J. Simons. Moving and looming stimuli capture attention. *Perception & Psychophysics*, 65:999–1010, 2003.

[Garrepalli *et al.*, 2023] Risheek Garrepalli, Jisoo Jeong, Rajeswaran C Ravindran, Jamie Menjay Lin, and Fatih Porikli. Dift: Dynamic iterative field transforms for memory efficient optical flow, 2023.

[Geistert *et al.*, 2016] Jonas Geistert, Tobias Senst, and Thomas Sikora. Robust local optical flow: dense motion vector field interpolation. In *2016 Picture Coding Symposium (PCS)*, pages 1–5, 2016.

[Gibson, 1951] James J. Gibson. The perception of the visual world. *The American Journal of Psychology*, 64(3):440–444, 1951.

[Haag *et al.*, 2016] Juergen Haag, Alexander Arenz, Etienne Serbe, Fabrizio Gabbiani, and Alexander Borst. Complementary mechanisms create direction selectivity in the fly. *eLife*, 5:e17421, 2016.

[Hammond, 1974] P. Hammond. Cat retinal ganglion cells: size and shape of receptive field centres. *The Journal of Physiology*, 242(1):99–118, 1974.

[Hassenstein and Reichardt, 1956] B. Hassenstein and W. Reichardt. Systemtheoretische analyse der zeit-, reihenfolgen- und vorzeichenauswertung bei der bewegungsperzeption des rüsselkäfers chlorophanus. *Zeitschrift für Naturforschung B*, 11(9-10):513–524, 1956.

[Horn and Schunck, 1981] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1):185–203, 1981.

[Huang *et al.*, 2022] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: a transformer architecture for optical flow, 2022.

[Huang *et al.*, 2023] Hsin-Ping Huang, Charles Herrmann, Junhwa Hur, Erika Lu, Kyle Sargent, Austin Stone, Ming-Hsuan Yang, and Deqing Sun. Self-supervised autoflow, 2023.

[Ilg *et al.*, 2016] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: evolution of optical flow estimation with deep networks, 2016.

[Jiang *et al.*, 2021] Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Jian

Zhao, Guodong Guo, and Zhenjun Han. Anti-uav: A large multi-modal benchmark for uav tracking, 2021.

[Jung *et al.*, 2023] Hyunyoung Jung, Zhuo Hui, Lei Luo, Haitao Yang, Feng Liu, Sungjoo Yoo, Rakesh Ranjan, and Denis Demandolx. Anyflow: Arbitrary scale optical flow with implicit neural representation, 2023.

[Kroeger *et al.*, 2016] Till Kroeger, Radu Timofte, Dengxin Dai, and Luc Van Gool. Fast optical flow using dense inverse search, 2016.

[Leventhal and Schall, 1983] Audie G. Leventhal and Jeffrey D. Schall. Structural basis of orientation sensitivity of cat retinal ganglion cells. *Journal of Comparative Neurology*, 220(4):465–475, 1983.

[Levick and Thibos, 1982] W. R. Levick and L. N. Thibos. Analysis of orientation bias in cat retina. *The Journal of Physiology*, 329(1):243–261, 1982.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[Lu *et al.*, 2023] Yawen Lu, Qifan Wang, Siqi Ma, Tong Geng, Yingjie Victor Chen, Huaijin Chen, and Dongfang Liu. Transflow: Transformer as flow learner, 2023.

[Lucas and Kanade, 1981] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, pages 674–679. Morgan Kaufmann Publishers Inc., 1981.

[Mayer *et al.*, 2016] Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016.

[Pérez *et al.*, 2013] Javier Sánchez Pérez, Enric Meinhardt-Llopis, and Gabriele Facciolo. Tv-l1 optical flow estimation. *Image Processing On Line*, 3:137–150, 2013.

[Ranjan and Black, 2016] Anurag Ranjan and Michael J. Black. Optical flow estimation using a spatial pyramid network, 2016.

[Reid and Shapley, 2002] R. Clay Reid and Robert M. Shapley. Space and time maps of cone photoreceptor signals in macaque lateral geniculate nucleus. *Journal of Neuroscience*, 22(14):6158–6175, 2002.

[Revaud *et al.*, 2015] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Epicflow: edge-preserving interpolation of correspondences for optical flow, 2015.

[Rodieck, 1965] R. W. Rodieck. Quantitative analysis of cat retinal ganglion cell response to visual stimuli. *Vision Research*, 5(12):583–601, 1965.

[Senst *et al.*, 2012] Tobias Senst, Volker Eiselein, and Thomas Sikora. Robust local optical flow for feature tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(9):1377–1387, 2012.

[Shi *et al.*, 2023] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation, 2023.

[Shou and Leventhal, 1989] Tiande Shou and A. G. Leventhal. Organized arrangement of orientation-sensitive relay cells in the cat's dorsal lateral geniculate nucleus. *Journal of Neuroscience*, 9(12):4287–4302, 1989.

[Shou, 2010] Tiande Shou. *Brain Mechanisms of Visual Information Processing*. USTC Press, 2010.

[Sun *et al.*, 2018] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume, 2018.

[Tao *et al.*, 2012] Michael Tao, Jiamin Bai, Pushmeet Kohli, and Sylvain Paris. Simpleflow: a non-iterative, sublinear optical flow algorithm. *Computer Graphics Forum*, 31(2pt1):345–353, 2012.

[Teed and Deng, 2020] Zachary Teed and Jia Deng. Raft: recurrent all-pairs field transforms for optical flow, 2020.

[Wang *et al.*, 2024] Bo Wang, Yifan Zhang, Jian Li, Yang Yu, Zhenping Sun, Li Liu, and Dewen Hu. Splatflow: Learning multi-frame optical flow via splatting. *International Journal of Computer Vision*, 132(8):3023–3045, 2024.

[Weinzaepfel *et al.*, 2013] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. Deepflow: large displacement dptical flow with deep matching. In *2013 IEEE International Conference on Computer Vision*, pages 1385–1392, 2013.

[Wexler *et al.*, 2001] Mark Wexler, Francesco Panerai, Ivan Lamouret, and Jacques Droulez. Self-motion and the perception of stationary objects. *Nature*, 409(6816):85–88, 2001.

[Wu *et al.*, 2023] Guangyang Wu, Xiaohong Liu, Kunming Luo, Xi Liu, Qingqing Zheng, Shuaicheng Liu, Xinyang Jiang, Guangtao Zhai, and Wenyi Wang. Accflow: Backward accumulation for long-range optical flow, 2023.

[Wulff and Black, 2015] Jonas Wulff and Michael J. Black. Efficient sparse-to-dense optical flow estimation using a learned basis and layers. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 120–130, 2015.

[Yuan *et al.*, 2024] Shuai Yuan, Lei Luo, Zhuo Hui, Can Pu, Xiaoyu Xiang, Rakesh Ranjan, and Denis Demandolx. Unsamflow: Unsupervised optical flow guided by segment anything model, 2024.

[Zhu *et al.*, 2009] Wei Zhu, Michael Shelley, and Robert Shapley. A neuronal network model of primary visual cortex explains spatial frequency selectivity. *Journal of Computational Neuroscience*, 26(2):271–287, 2009.