HuPerFlow: A Comprehensive Benchmark for Human vs. Machine Motion Estimation Comparison

Yung-Hao Yang[1] Zitang Sun[1] Taiki Fukiage[2] Shin'ya Nishida[1,2*]
[1]Cognitive Informatics Lab, Graduate School of Informatics, Kyoto University, Japan
[2]Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Japan
https://kucognitiveinformaticslab.github.io/Huperflow-Website/

Abstract

As AI models are increasingly integrated into applications involving human interaction, understanding the alignment between human perception and machine vision has become essential. One example is the estimation of visual motion (optical flow) in dynamic applications such as driving assistance. While there are numerous optical flow datasets and benchmarks with ground truth information, human-perceived flow in natural scenes remains underexplored. We introduce HuPerFlow—a benchmark for human-perceived flow, measured at 2,400 locations across ten optical flow datasets, with ∼38,400 response vectors collected through online psychophysical experiments. Our data demonstrate that human-perceived flow aligns with ground truth in spatiotemporally smooth locations while also showing systematic errors influenced by various environmental properties. Additionally, we evaluated several optical flow algorithms against human-perceived flow, uncovering both similarities and unique aspects of human perception in complex natural scenes. HuPerFlow is the first large-scale human-perceived flow benchmark for alignment between computer vision models and human perception, as well as for scientific exploration of human motion perception in natural scenes. The HuPerFlow benchmark is publicly available on the HuPerFlow website.

1. Introduction

Identifying object motion within the natural environment is crucial for humans, as it supports essential tasks such as navigation, tracking, prediction, and pursuit. In computer vision (CV), optical flow—representing an object's movement in pixel coordinates—serves as a foundational spatiotemporal feature for downstream applications, such as action recognition [36, 39], driver assistance [8, 26], robot navigation [6], video processing [43], etc. With the rise of data-driven approaches, the availability of datasets and benchmarks has consistently fueled advancements in optical flow methods. Numerous studies have leveraged physical engines to generate realistic synthetic data [5] or employed specialized laser and depth cameras [13] to capture ground-truth optical flow in real scenes.

Figure 1. Example of HuPerFlow. The red arrows indicate human response to perceived flow, and the green arrows indicate ground truth motion vectors. The circles indicate the magnitudes of endpoint errors, the difference between perceived flow and ground truth.

However, few studies have examined how humans perceive motion in natural scenes. AI models increasingly interact with humans in various application fields, such as image quality assessment [45, 48], video generation [28], and large language model evaluation [29]). Human responses often guide the development of models, which in turn enhances the models' support of human activities. In the context of visual motion perception, estimating human-perceived flow in addition to the ground truth will be use-

ful, for example, in automatic driving [7, 10, 22]. While fully autonomous driving systems can be solely based on the physical ground truth estimation, partially autonomous systems that assist human drivers will be empowered by the ability to predict when, where, and how the drivers are likely to misinterpret motion flow. Another potential use for our dataset is animation synthesis. To control the quality of animation, the animators need to check whether human viewers perceive the movements of the created animation in the way they originally intended. If there are disagreements, the animators should modify the images to minimize the errors. This process can be advanced if human-perceived motion can be automatically predicted. Machine prediction of human-perceived flow is a hard problem for hand-drawn animations since there is no physical ground truth. Even when the animation is rendered from a physically correct dynamic model of the scene, human-perceived motion can be distorted by many factors, including frame rates, image resolution, and the touch of the painting.

Humans possess a sophisticated visual system capable of estimating object movement in complex natural scenes; however, the underlying mechanisms differ significantly from those used in current CV models. Motion inference is inherently an under-constrained problem (such as the aperture problem [30]), and the solution in humans is a perception that unfolds over a spatiotemporal process spanning multiple frames [1]. In contrast, CV methods typically focus on deterministic dense correspondence between consecutive frames. This difference can sometimes result in entirely different motion perceptions between human vision and CV algorithms, as seen in phenomena like the reverse phi [2] and missing fundamental illusions [4] investigated in vision science and psychology.

Despite this significant gap, currently, no large-scale benchmark directly evaluates the relationship between CV model responses, physical ground truth, and human perceptual responses to motion in natural scenes, thus hindering further research within this field. To address this problem, we construct the first large-scale comprehensive benchmark for human-perceived optical flow across several types of naturalistic scenes. Our study introduces several unique features compared to prior work: 1): **Human-centric labeling**: The dataset's labels are directly derived from human perception, gathered through rigorously designed psychophysical experiments. We refer to these human-perceived labels as "Human Perception" and use the term "Ground Truth (GT)" for the physical-correct labels generated by conventional methods (Fig. 1); 2) **Large-scale**: our dataset includes around 38,400 data points (trials) derived from 480 participant sessions, spanning over 240 hours of collection time. The benchmark comprises ten representative computer vision motion datasets—such as Sintel [5], KITTI [26], Spring [25], VIPER [35], etc—each with specifically probed locations and optical flow GT; 3) **A novel experimental paradigm**: Unlike traditional human assessments (e.g., image quality evaluation), collecting motion perception data is complex and time-intensive. Motion data collection requires a strict display setting, including a constant refresh rate, high resolution, and high-speed hardware to ensure them. Our innovative online experimental paradigm, grounded in a psychophysical method developed for a lab experiment [47], ensures stable and reliable data collection even in online settings, making large-scale human-perceived motion data feasible.

In addition to constructing this benchmark dataset, we conducted several initial comparative analyses between several representative CV models and human responses, including spatiotemporal gradient-based methods [11], recurrent refinement models [42], multi-frame methods [37], vision transformers [14], human-inspired approaches [38, 40], etc. Our findings reveal several key differences between model predictions and human perception, as well as unique characteristics of human perception in complex natural scenes, including abrupt direction switches (e.g., VIPER [35]), motion blur and non-rigid motion (e.g., MPI Sintel [5]), object translation and rotation (e.g., Flythings3D [24]), camera motion (e.g., Monkaa [24]), and local versus global motion (e.g., MHOF [34]). While the recent CV models correlate closely with human perception, they often exhibit low partial correlation when controlling the ground-truth effects, suggesting their design inherently fails to capture humans' unique perceptual biases.

In summary, our contributions include: (1) the first large-scale human-perceived optical flow benchmark across ten diverse motion scenarios, (2) a new online psychophysics-based experimental paradigm enabling the large-scale collection of precise human-perceived optical flow in natural scenes, and (3) an evaluation of ten different representative models from both computer vision and human vision fields, revealing significant gaps between human motion interpretation and model predictions. These insights advance computer vision research on human-aligned AI and support the development of technologies for better interaction and assistance with humans.

2. Previous optical flow datasets

2.1. Computer optical flow datasets

There are several optical flow datasets with physical motion ground truth, and we selected some of them based on motion content, image resolution, sequence count, and frames per sequence for online human observers. KITTI flow 2015 [26] is a real-world driving condition, where ground-truth optical flows were captured by stereo camera, laser scanner, and GPS. Virtual KITTI 2 [12] is a photo-realistic synthetic driving dataset featuring different weather conditions

and viewing angles. The Driving subset of the Scene Flow Datasets [24] includes eight synthetic sequences (*slow / fast × 15mm / 300mm focal length × forwards / backward scene*). VIPER [35] and TartanAir [44] provide synthetic video sequences with various speed and direction changes, representing diverse navigation scenarios beyond typical driving benchmarks for distinct object motion and camera motion.

In contrast to automotive navigation scenarios, MPI Sintel [5] and SPRING [25] feature long-range, non-rigid motion with discontinue, large jumps, motion blur, and varying illumination, scene, and material properties. Monkaa, also from Scene Flow Datasets [24], includes significant camera translation and rotation, crucial for revealing motion perception influenced by surrounding objects and background. MHOF [34] captures diverse human motion, featuring both local and global motions across various postures and actions. Flythings3D subset of the Scene Flow Datasets [24], includes multiple geometric shapes and objects translating and rotating along complex 3D trajectories. Since the video presentation required several frames for human observers, we excluded well-known benchmarks such as Middlebury [3] (2-8 frames), FlyingChairs2 [16] (2 frames), ChairsSDHom [15] (2 frames), etc.

2.2. Human motion perception datasets

Previous studies on human visual motion perception generally used well-controlled artificial visual stimuli, such as sine-wave gratings [21], random-dot kinematograms [20], and biological motion [19], to investigate specific processing mechanisms. Some studies used more complex natural-scenes stimuli to explore neural responses for them [27, 33]. While these studies provided valuable insights into human motion perception, only one study [47], to our knowledge, has directly reported human-perceived vectors for nature scenes, and compared them with ground-truth motion vectors and computer vision model predictions. This psychophysical study revealed that human-perceived optical flow shows various perceptual deviations from the ground-truth flow. Note that human-perceived vectors in dynamical movies, which are mainly based on motion energy within a $\sim$100 ms temporal window [1], could be significantly different from the correspondence of salient feature points between adjacent frames (cf. [9]). However, [47] examined only five movies, all selected from the SlowFlow version (1008 frames per second) of MPI Sintel benchmark [18], in which the jump sizes between adjacent frames are substantially smaller than those of the standard version used in the current study. It is also worth noting that the movies in [47] were presented through a circular aperture centered on the probed location.

In the current study, we adopt an approach similar to [47] with significant modifications: (1) using ten optical flow benchmarks that are widely used for training and testing in computer vision fields, (2) choosing a larger number of motion clips with dynamic and diverse spatiotemporal properties, and (3) presenting the entire image on the display, as observers would naturally perceive it, rather than through a central aperture. This approach provides a greater motion context and ecological validity.

3. The HuPerFlow dataset

3.1. Methods

3.1.1. Ethical statement

This study was approved by the Research Ethics Committee of the Graduate School of Informatics at Kyoto University. All participants received online informed consent before the experiments.

3.1.2. Stimuli and Appartatus

The experiment was conducted using PsychoPy [32], a software for building behavioral experiments, and Pavlovia [31], an online platform for running these experiments in a web browser. To account for diverse hardware environments in online experiments, we controlled the visual presentation in both spatial and temporal domains. Spatially, to standardize viewing angles, observers performed a "credit card calibration," adjusting an on-screen virtual credit card to match a physical one. This allowed us to calculate the physical pixel size after observers adjusted their viewing distance to ensure that 50 pixels corresponded to 1° of visual angle, maintaining consistent stimulus pixel sizes across devices. Temporally, visual stimuli were presented at 30 Hz. Since optimal timing control is crucial for online motion perception experiments, we did not use a higher frame rate for which frame drops were more likely to occur under some users' environments. In addition, We recorded frame drops and onset times during online experiments (see Supplementary Appendix 1 for more details). The frame dropping rate was .041%, and the frame onset error was -0.055 ms (SD = 3.664 ms), indicating stable and precise visual stimulus presentation across devices.

This project contained two phases: a preliminary Random Dot Kinematogram (RDK) for training and selection of the observers, and a main session to collect human-perceived flow data using multiple optical flow datasets. In the RDK phase, 5,000 black and white dots moved uniformly within a 600-pixel circular aperture. In the main experiment, we used ten optical flow datasets including KITTI Flow 2015 [26], Virtual KITTI 2 [12], Driving [24], VIPER [35], TartanAir [44], MPI Sintel Flow [5], SPRING [25], Monkaa [24], MHOF [34], and Flythings3D [24]. As most benchmarks include sub-datasets, we present our selection in detail in Tab. 1.

We applied a standard movie-formatting method across datasets with the following exceptions: (1) Most datasets

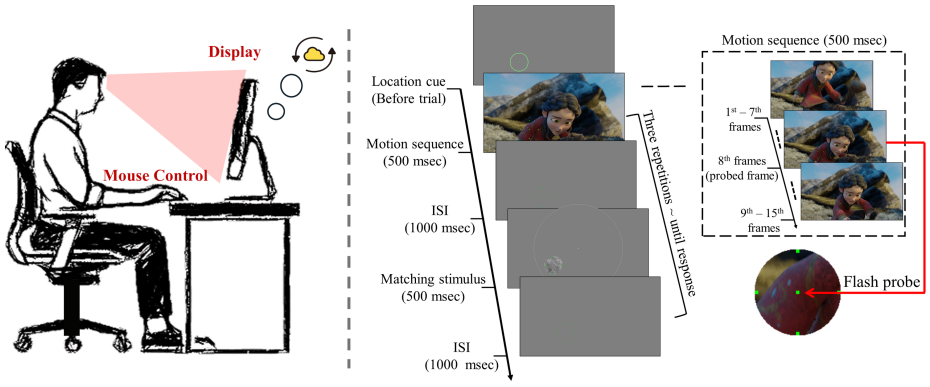

Figure 2. Experimental procedure: At the start of each trial, a green circle appeared to mark the selected area. Next, a motion sequence and a matching stimulus were presented alternately until a response was made. For most optical flow datasets, the motion sequences consisted of 15 frames (500 ms). It was followed by an inter-stimulus interval (ISI) of 1000 ms, then a matching stimulus (500 ms), and a second ISI (1000 ms). A flash probe was shown in the middle frame (the 8th frame) of the motion sequence to indicate the timing and location of the target motion. The green circle location cue, four-dot placeholders, and matching stimulus were presented sequentially at the same location.

Table 1. Summary of the selected optical flow datasets. * indicates the image size is resized by cubic interpolation.

Datasets	Resolution (pixels)	Sequences	Total Frames	Selection
KITTI 2015 [26]	1,242 × 375	200	4,200	Training, left image
Virtual KITTI 2 [12]	1,242 × 375	5	21,260	Clone, camera_0
Driving [24]	960 × 540	8	2,400	Finalpass, left, slow
VIPER [35]	960 × 540*	77	134,097	Val, jpg
TartanAir [44]	640 × 480	16	6,982	image_left
MPI Sintel [5]	1,024 × 436	23	1,064	Training, Final
Spring [25]	960 × 540*	37	5,000	Train_frame_left
Monkaa [24]	960 × 540	24	8,664	Finalpass, left
MHOF [34]	640 × 640	1,238	95,868	Train, masked with segm_EXR
FlyingThings [24]	960 × 540	2,180	21,818	Val, image_clean, left

were presented at their original sizes, except for VIPER [35] and SPRING [25], which were resized from 1920 × 1080 to 960 × 540 for online observers, with the optical flow adjusted accordingly. (2) Frames F-7 to F+7 around a selected testing frame were used as movie sequences, except for MHOF [34] and Flythings3D [24], which were limited to F-4 to F+5 due to each sequence having only ten frames. (3) Virtual KITTI2 [12], SPRING [25] and Scene Flow Datasets [24] (i.e., Driving, Monkaa, and Flythings3D) contain both future (forward) and past (backward) flow directions. However, only future-directed flows were reported as video images were presented sequentially in the forward direction, frame-by-frame. (4) We did not modify the original images, except for MHOF [34] dataset, where frames contained irrelevant background scenes. To standardize the background, we used the "`segm_EXR`" provided by MHOF to render it as a uniform gray background.

3.1.3. Design and Procedure

In both the RDK and main experiments, each session began with six practice trials followed by 80 formal trials. The practice trials familiarized observers with the procedure and helped them understand the differences between the ground truth and their perceived vectors. To support this, we provided visual feedback of both GT and perceived vectors after each response during practice trials. If the angle between perceived vectors and the GT exceeded 30°, or if the endpoint error (EPE, the Euclidean distance between the human responses and GT, Equation 1.) was greater than 10 pixels, the same practice trial would be repeated.

$$\text{EPE} = \sqrt{(u_{\text{Resp}} - u_{\text{GT}})^2 + (v_{\text{Resp}} - v_{\text{GT}})^2} \quad (1)$$

In the RDK session, the direction (0°–360°) and speed (1–20 pixels per frame, PPF) of the dots were varied randomly across trials but kept constant within a trial. During formal trials in this session, feedback was also shown to reinforce training. In the main experimental session, we used ten datasets as previously described, selecting 24 sequences per dataset. Each sequence included ten target motion locations with speeds under 20 PPF at a specific frame. At each probed location, we collected 16 repeated trials (four repetitions per observer × four observers) to obtain a stable measure of perceived flow, totaling 2,400 probed locations across 38,400 trials. In each session, participants viewed two movie sequences with ten target motion locations and four repetitions each (80 trials in total). This yielded 480 sessions across 10 datasets × 12 sessions (2 movie sequences each) × 4 observers. No feedback was provided during formal trials in the main experiments.

Fig. 2 illustrates the display layout and procedure in the main experiment. Both the RDK and main experiments

used a method of adjustment to measure human-perceived flow at the flash-probed spatiotemporal location. Observers reported the perceived flow of a target motion by matching it with a circular Brownian noise pattern ($1/f^2$, 120 pixels in diameter), which has a spectrum similar to natural images and gives the least motion aliasing for a wide range of motion speeds. Using a mouse cursor (a purple dot, 10 pixels in diameter) within a circular control panel (600 pixels in diameter), observers could freely adjust the direction (0°–360°) and speed (1–20 pixels per frame) of the matching Brownian noise. The direction was mapped to angles (θ) in a polar coordinate, and speed was adjusted logarithmically based on the cursor's distance from the center of the panel. This method provides more accurate measurements of perceived motion vectors than subjective reports. The control panel for the matching stimulus was always fixed at the display's center. See Supplementary Video S1 for the animated procedure.

In the RDK session, where direction and speed were uniform within each trial, both the target motion and matching stimulus were presented at the center of the aperture. In the main experiment, a green circle (120 pixels in diameter), presented before each trial, served as a spatial cue indicating the target motion area. Additionally, four additional dots (green dots, 10 pixels in diameter) served as placeholders throughout the trial. They were positioned 60 pixels above, below, on the left, and on the right of the probed dot. For most datasets, a repetition cycle (see below) consisted of the target movie sequence (500 ms, 15 frames), an interstimulus interval (ISI, 1000 ms), and the matching stimulus (500 ms), followed by a second ISI (1000 ms) before the next cycle. For the MHOF [34] and FlyingThings3D [24] datasets, which contain only ten frames (333 ms), the second ISI was adjusted to 1167 ms to keep the cycle duration consistent.

In each trial, a flash-probed dot (a green dot, seven pixels in diameter) was presented at a selected testing frame and location to indicate the timing and position of the target motion. The flash probe appeared in the middle of each movie sequence presentation (typically at the 8$^{\text{th}}$ frame, and at the 5$^{\text{th}}$ frame for MHOF and FlyingThings3D), indicating when observers should respond. The probed target motions were intentionally selected from informative locations, including main objects, the background, and minor objects. In the main experiment, the four-dot placeholders appeared around the probed location, with the matching stimulus displayed directly at that placeholder's location. During each trial, observers could view the movie sequences and match Brownian noise multiple times (repetition cycles), while responses were only accepted after completing three repetition cycles. Since the target motion was shown for just one frame, this adjustment process allowed observers to refine their responses through repeated viewing.

3.2. Results

3.2.1. Data quality control

The study included 55 observers (mean age = 22.18, SD = 2.82, 24 females and 31 males) who participated in a preliminary Random Dot Kinematogram (RDK) experiment. To ensure reliable perceived flow data, we initially selected observers whose correlation coefficient between perceived vectors and ground truth vectors exceeded .90 in the RDK session. Nine observers with low correlation (individual correlation range: .501 to .885) and four who were unable to complete the experiment were excluded. The group-level correlation for the 42 remaining observers was high (Pearson correlation coefficient r = .955, with individual correlations ranging from .917 to .985), demonstrating their ability to reproduce perceived motion accurately. These 42 observers (mean age = 22.19, SD = 2.46; 17 females, 25 males) then proceeded to the main experimental sessions. They participated in 480 sessions since some observers attended multiple different sessions.

3.2.2. Perceived vectors and "flow illusions"

In Fig. 3, we show representative examples of averaged human responses for each dataset (see Supplementary Video S2 for target motions and the corresponding averaged human responses across all probed locations within each dataset). Our findings reveal systematic and stable human-specific biases in motion perception under certain scenarios. We termed these systematic discrepancies between human responses and ground truth motions as "flow illusions". These deviations were not random errors but provided valuable insights into the underlying processing and mechanism of human motion perception. On the other hand, we also observed individual differences, with an average inter-observers correlation of 0.407 (SD: 0.417; range from -0.329 to 0.990) across all probed locations. These Individual variations would be useful for training models to reproduce the distribution of human motion perception. To facilitate further research, we have made all raw individual data publicly available on the HuPerFlow website [46].

In most driving datasets such as KITTI 2015 [26], Virtual KITTI 2 [12], and Driving [24], the direction and speed of optical flows transition smoothly across frames, with perceived flows aligning closely with the ground truth motion. This consistency supports the reliability of our measurement method in capturing observers' perceptions under spatiotemporally smooth scenes. Notably, as optical flow locations deviated from the scene center, observers often perceived the directions to be slightly shifting toward the central region. This finding may reflect observers' attempts to cancel out the image motion components produced by self-motion (camera motion) during driving. In contrast, other navigation datasets like VIPER [35] and TartanAir [44] show abrupt direction changes, particularly during ac-

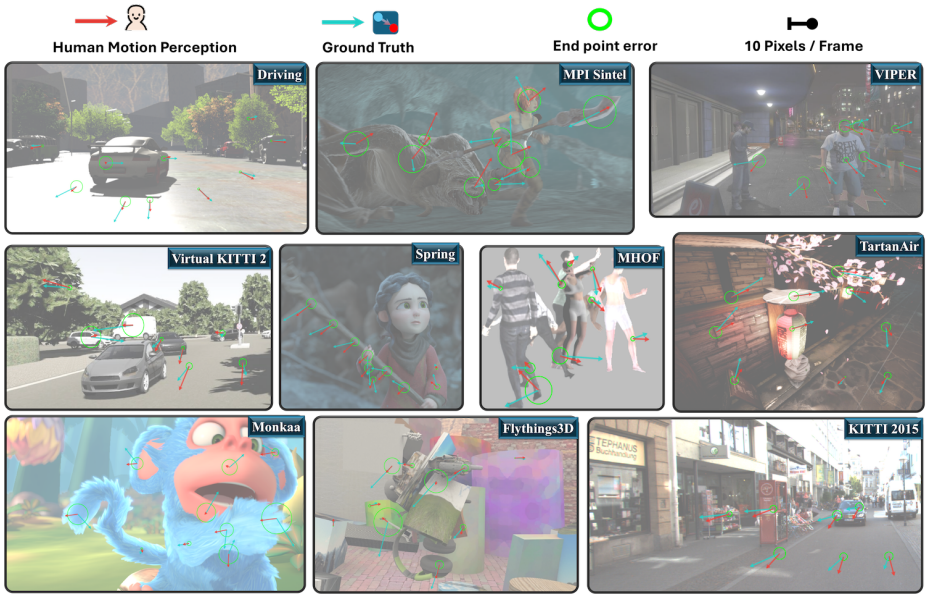

Figure 3. Demonstration of human perceived motion vectors. The results cover ten representative existing optical flow benchmarks.

tions such as turning, lane changing, or flying backward. These involve interactions between image motion and camera motion, with perceived flows often revealing systematic errors primarily driven by camera motion.

In datasets like MPI Sintel [5] and SPRING [25], where images include blurred, non-rigid, discontinuous motion and large jump size (cf. [47] using stimuli with small jumps), observers generally failed to discern individual locations, perceiving them as part of a unified grouping movement instead. Unlike the datasets described in the previous paragraph, which were viewed with smoothly moving cameras, Monkaa [24] dataset involved significant camera rotation around relatively stationary objects. Observers often perceived these objects as moving due to the motion of surrounding objects or backgrounds, illustrating the phenomena of relative and induced motion.

In the MHOF [34] datasets, human figures generally exhibit global body movement accompanied by local motions from the head and limbs. Although the motion structures are less complex than in MPI Sintel and Spring, observers tended to focus on the direction of the global body movement and ignored the local gestures. In the Flythings3D [24], multiple objects float along 3D random trajectories, requiring observers to perform object-ground and object-object segmentation. They were successful in perceiving motion against simple backgrounds and distinct surrounding objects but made more errors in cluttered, messy conditions.

As different datasets revealed various types of errors, we calculated the correlation between human response and ground truth for each dataset separately. In general, the driving datasets (i.e., KITTI 2015 [26], Virtual KITTI 2 [12], and Driving [24]) showed high correlations, but others showed middle to low correlation (see Supplementary Appendix 2 for the scatter plot of each dataset).

3.2.3. EPE in relation to optical flow properties

As shown in Fig. 3, the EPE of human response (indicated by the diameter of the green circle) varied across different positions and scenes. Following [5], we evaluated how EPE changed in relation to various optical flow properties. Fig. 4 shows that EPE increased as GT speed increased, which suggests that faster motions imposed more significant computational challenges. To assess temporal flow dynamics, we compared optical flow from the previous frame (t-1) to the selected probed frame (t-0) and analyzed the temporal gradient changes in uv-vectors over two intervals: from t-1 to t-0, and from t-0 to t+1 (ground truth), normalized by GT speed from the same location. Generally, we observed that EPE increased as the normalized temporal flow gradient increased up to the ratio of 2, and then leveled off.

We also considered spatial image properties by using a Sobel kernel to derive the image gradient, calculating the average edge magnitude within a ± 40-pixel neighborhood around the probed location. The results indicate that EPE decreased as the image gradient increased. Additionally, using instance boundary from the MPI Sintel dataset, we calculated the shortest distance to the boundary and found that EPE also decreased as the distance increased. These findings suggest that edges and boundaries may create ambiguities in motion flow interpretation and interrupt motion perception. Finally, using camera motion data from the MPI Sintel as a proxy for self-motion within a scene, we observed that EPE increased with self-motion, and this effect persisted even when excluding the effect of GT speed (image motion).

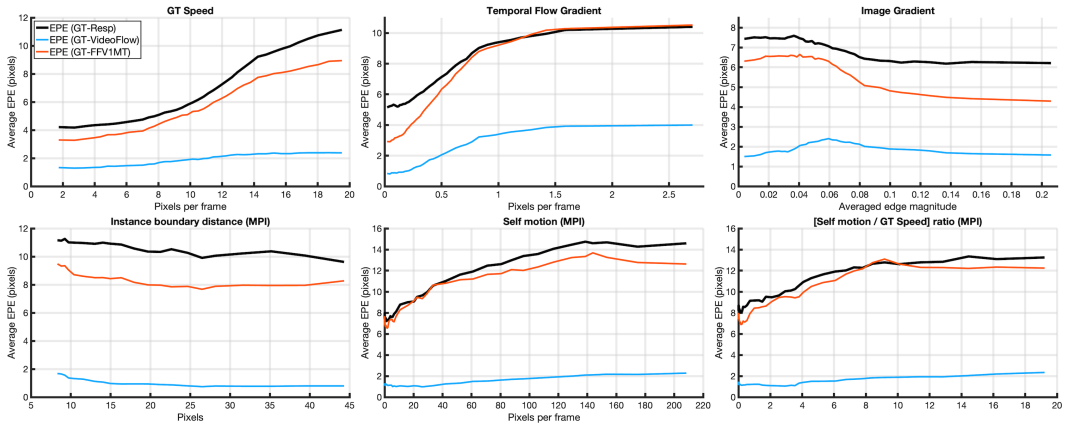

Figure 4. Endpoint errors (EPEs) of human responses and two machine vision models as functions of optical flow properties. The horizontal axis of each panel is the GT speed, normalized temporal flow gradient, spatial image gradient, instance boundary distance, and self-motion with(out) normalized by the image GT speed. The vertical axis represents EPE with different scales for visualization.

In addition to the EPE of human response, Fig. 4 shows the EPEs of two machine vision models, VideoFlow and FFV1MT (see Sec. 4). While there were clear differences in the absolute level of performance, the optical flow properties generally affect EPE in similar ways for human and machine vision models, except for the effects of spatial image gradient effect on VideoFlow.

4. Comparison with optical flow algorithms

We selected a range of representative optical flow algorithms to assess the similarities and differences among human-perceived flows, ground truth motion, and predictions from computer vision models. These included classical methods based on brightness constancy assumptions, such as Farnebäck [11]; biologically inspired approaches like FFV1MT [38]; multiscale inference models, such as FlowNet2 [15]; recurrent refinement models like RAFT [42]; transformer-based models like FlowFormer [14]; graph reasoning methods such as AGFlow [23]; multi-frame approaches like VideoFlow [37]; Motion energy and self-attention based models like V1Attention [40] and DualModel [41]; as well as general-purpose architectures like Perceiver IO [17]. For a fair comparison, we used each model's default configuration or the best pre-trained model fine-tuned on the Sintel dataset. Since unsupervised learning methods approximate the supervised ground truth, we limited our comparison to supervised methods.

For evaluation, we first calculated the average EPE and Pearson correlation coefficients between model predictions and both human responses and ground truth. To further understand the models' alignment with human perception independently of ground truth influence—since most models were trained to fit ground truth and the ground truth was inherently correlated with human responses—we also calculated the partial correlation between model predictions and human responses, controlling for ground truth, as shown in Equation 2. Here, $r_{x,y}$ denotes correlation coefficients between x and y.

$$\rho_{\text{model}} = r_{\text{resp, model}\cdot\text{GT}} = \frac{r_{\text{resp, model}} - r_{\text{resp, GT}} \cdot r_{\text{model, GT}}}{\sqrt{1 - r^2_{\text{resp, GT}}} \cdot \sqrt{1 - r^2_{\text{model, GT}}}}. \tag{2}$$

To complement partial correlation analysis, we also adopted a Response Consistency Index (RCI) from [47]. This index measures the similarity between model predictions and human responses at each probed location, approaching +1 when the model's prediction aligns closely with human flow illusions (see Supplementary Appendix 3 for more details).

Tab. 2 summarizes model predictions across all datasets. Overall, VideoFlow's predictions were the closest to the ground truth. RAFT showed the most closely resembling human responses while also having a strong alignment with the ground truth (second-best among models). When controlling for the ground truth effect, FFV1MT and V1Attention-StageI were the most similar to human flow illusions, as indicated by high partial correlations and average RCI values. A detailed analysis comparing the predictions of optical flow algorithms with both the ground truth and human responses for each dataset is provided in Tables S1-S12 of Supplementary Appendix 4.

Fig. 5 provides examples of model predictions, highlighting examples where predictions aligned closely with the ground truth (e.g., VideoFlow and RAFT) versus those having some resemblance to human response errors (e.g., FFV1MT and V1Attention-StageI). For instance, at Point (A-I) in Fig. 5, both flow illusions and FFV1MT exhibited temporal-sluggish responses but VideoFlow aligned well with GT. Conversely, point (A-II) in Fig. 5 both model predictions were similar to flow illusions, likely due to image signal loss, as the GT indicated a small, dark stone falling. At Point (B-I) and other selected body locations, both human perception and FFV1MT reflected grouping motion in the same direction, regardless of local gesture directions.

Table 2. Predictions of optical flow algorithms versus Human response or ground truth (GT). ρ: Partial correlation; r: Pearson correlation coefficient; EPE: vector end-point error; *uv, dir, spd* represent motion components in Cartesian space, direction, and speed, respectively. RCI is introduced from [47] to represent the model's similarity to human perception (the larger, the more human-aligned) .

Method	ρ			RCI	v.s. Human				v.s. GT			
	ρ_{uv}	ρ_{dir}	ρ_{spd}		r_{uv}	r_{dir}	r_{spd}	EPE	r_{uv}	r_{dir}	r_{spd}	EPE
Ground Truth	NaN	NaN	NaN	NaN	0.66	0.43	0.29	6.96	1.00	1.00	1.00	0.00
Farnebäck [11]	0.17	0.14	0.04	0.13	0.54	0.38	0.15	6.91	0.68	0.65	0.39	5.52
FFV1MT [38]	**0.31**	0.28	**0.06**	0.15	0.63	0.45	0.20	6.58	0.70	0.59	0.54	5.84
RAFT [42]	0.11	0.22	0.01	0.05	0.64	0.47	**0.23**	7.37	0.93	0.85	0.79	2.19
FlowNet2 [15]	0.10	0.26	0.00	0.07	0.59	**0.48**	0.18	7.55	0.82	0.80	0.61	3.24
FlowFormer [14]	0.07	0.19	-0.02	0.04	0.64	0.46	0.21	7.43	0.93	0.87	0.79	2.08
VideoFlow [49]	0.08	0.22	-0.02	0.04	**0.64**	0.47	0.22	7.47	**0.94**	**0.88**	**0.81**	**1.91**
Perceiver IO [17]	0.09	0.17	0.03	0.05	0.62	0.44	0.23	7.75	0.89	0.82	0.74	2.87
AGFlow [23]	0.16	0.24	0.01	0.11	0.54	0.45	0.15	7.94	0.68	0.66	0.46	5.44
V1Attention-StageI [40]	0.23	**0.28**	0.02	**0.18**	0.55	0.45	0.14	**6.36**	0.63	0.54	0.42	7.00
DualModel [41]	0.20	0.25	0.06	0.13	0.57	0.45	0.20	7.31	0.70	0.67	0.53	5.66

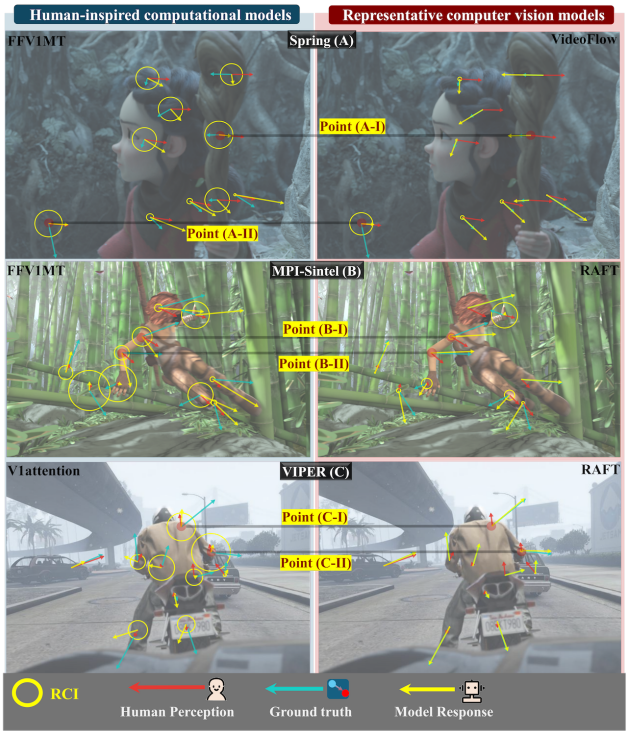

Figure 5. Predicted vectors of optical flow algorithms. Examples are FFV1MT [38], V1Attention [40], VideoFlow [37], and RAFT [42]. We marked some points with large RCI differences between the models in the left and right columns. Generally, the model on the left side shows a more human-aligned response over ground truth, while the right side shows the opposite result.

At Point (C-I), a strong self-motion effect occurred toward a static motorcycle. Despite the strong physical optical flow in the image, human observers and V1Attention-StageI appeared to compensate for self-motion to perceive the motorcycle as stationary. For more examples, please visit the HuPerFlow website [46]. Note, however, that there remains a large gap between human responses and the most similar model predictions, possibly because the purpose of human visual motion processing is not merely to estimate image flow. The primary purpose of making our dataset is to help future studies develop models that can fill this gap.

5. Conclusion

By overcoming technical challenges, we developed a novel online paradigm to collect human motion perception data, introducing the first large-scale benchmark for human-perceived optical flow. Our findings identified distinct perceptual illusions across various scene scenarios, some aligning with and others diverging from the predictions of CV optical flow algorithms. Human motion perception, compared to state-of-the-art CV models, has limited spatial, temporal, and intensity resolutions and struggles with motion estimation in retinal coordinates. However, human motion vision has remarkable capabilities in analyzing hierarchical movement structures and distinguishing object motion from self-motion in world coordinates. Our proposed methods can be extended to collect broader human data from diverse datasets, further bridging gaps between CV, human perception, and ground truth.

While our dataset may not be sufficient to train models from zero, it offers valuable benchmarks for model evaluation and fine-tuning. Physically accurate CV models are crucial, but aligning with human perception is equally important for human-centered applications. Notably, current optical flow models fail to predict various human-perceived motion illusions, highlighting the need to address this gap both for practical applications and for advancing the scientific understanding of human vision.

Acknowledgments

This work was supported in part by JSPS Grants-in-Aid for Scientific Research (KAKENHI), Grant Numbers JP20H00603, JP20H05957, and JP24H00721, and by the Spring Fellowship, Grant Number JPMJFS2123.

References

[1] Edward H Adelson and James R Bergen. Spatiotemporal energy models for the perception of motion. *Josa a*, 2(2): 284–299, 1985. 2, 3

[2] Stuart M Anstis and Brian J Rogers. Illusory continuous motion from oscillating positive-negative patterns: Implications for motion perception. *Perception*, 15(5):627–640, 1986. 2

[3] Simon Baker, Daniel Scharstein, J. P. Lewis, Stefan Roth, Michael J. Black, and Richard Szeliski. A Database and Evaluation Methodology for Optical Flow. *International Journal of Computer Vision*, 92(1):1–31, 2011. 3

[4] Richard O Brown and Sheng He. Visual motion of missing-fundamental patterns: motion energy versus feature correspondence. *Vision Research*, 40(16):2135–2147, 2000. 2

[5] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*, pages 611–625. Springer-Verlag, 2012. 1, 2, 3, 4, 6

[6] Haiyang Chao, Yu Gu, and Marcello Napolitano. A survey of optical flow techniques for robotics navigation applications. *Journal of Intelligent & Robotic Systems*, 73:361–372, 2014. 1

[7] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE international conference on computer vision*, pages 2722–2730, 2015. 2

[8] Jaechan Cho, Yongchul Jung, Dong-Sun Kim, Seongjoo Lee, and Yunho Jung. Moving object detection based on optical flow estimation and a gaussian mixture model for advanced driver assistance systems. *Sensors*, 19(14):3217, 2019. 1

[9] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adrià Recasens, Lucas Smaira, Yusuf Aytar, João Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *arXiv preprint arXiv:2211.03726*, 2022. Published in NeurIPS Datasets and Benchmarks track, 2022. 3

[10] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 2

[11] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the Scandinavian conference on Image analysis*, pages 363–370. Springer, 2003. 2, 7, 8

[12] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3, 4, 5, 6

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 1

[14] Tzehoong Huang, Zhichao Zhang, Wenjie Zeng, Ming Liu, Chunjing Zhang, Hongsheng Li, and Shijian Lu. Flowformer: A transformer architecture for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 7, 8

[15] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2462–2470, 2017. 3, 7, 8

[16] Eddy Ilg, Tonmoy Saikia, Margret Keuper, and Thomas Brox. Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[17] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021. 7, 8

[18] Joel Janai, Fatma Güney, Jonas Wulff, Michael J Black, and Andreas Geiger. Slow flow: Exploiting high-speed cameras for accurate and diverse optical flow reference data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3597–3606, 2017. 3

[19] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, 1973. 3

[20] Bela Julesz. *Foundations of Cyclopean Perception*. University of Chicago Press, 1971. 3

[21] D. H. Kelly. Fourier components of moving gratings. *Behavior Research Methods*, 14(6):435–437, 1982. 3

[22] Monika Lohani, Brennan R Payne, and David L Strayer. A review of psychophysiological measures to assess cognitive states in real-world driving. *Frontiers in human neuroscience*, 13:57, 2019. 2

[23] Ao Luo, Fan Yang, Kunming Luo, Xin Li, Haoqiang Fan, and Shuaicheng Liu. Learning optical flow with adaptive graph reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1890–1898, 2022. 7, 8

[24] 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 *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134. 2, 3, 4, 5, 6

[25] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 4, 6

[26] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference*

on computer vision and pattern recognition, pages 3061–3070, 2015. 1, 2, 3, 4, 5, 6

[27] Shinji Nishimoto and Jack L. Gallant. A three-dimensional spatiotemporal receptive field model explains responses of area mt neurons to naturalistic movies. *Journal of Neuroscience*, 31(41):14551–14564, 2011. 3

[28] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14277–14286, 2023. 1

[29] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 1

[30] Christopher C Pack and Richard T Born. Temporal dynamics of a neural solution to the aperture problem in visual area mt of macaque brain. *Nature*, 409(6823):1040–1042, 2001. 2

[31] Jonathan Peirce and the PsychoPy team. *Pavlovia: Online platform for running experiments*. Open Science Tools, 2018. Accessed: 2024-10-25. 3

[32] Jonathan W. Peirce and Michael MacAskill. *Building Experiments in PsychoPy*. Sage Publications Ltd, London, United Kingdom, 2018. 3

[33] Silvia Pitzalis, Carlo Serra, Valerio Sulpizio, Giovanni d'Avossa, and Gaspare Galati. Neural bases of self- and object-motion in a naturalistic vision. *Human Brain Mapping*, 41(4):1084–1111, 2020. 3

[34] Anurag Ranjan, David T. Hoffmann, Dimitrios Tzionas, Siyu Tang, Javier Romero, and Michael J. Black. Learning multi-human optical flow. *International Journal of Computer Vision (IJCV)*, 2020. 2, 3, 4, 5, 6

[35] Stephan R. Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2232–2241, 2017. 2, 3, 4, 5

[36] Laura Sevilla-Lara, Yiyi Liao, Fatma Güney, Varun Jampani, Andreas Geiger, and Michael J Black. On the integration of optical flow and action recognition. In *Pattern Recognition: 40th German Conference, GCPR 2018, Stuttgart, Germany, October 9-12, 2018, Proceedings 40*, pages 281–297. Springer, 2019. 1

[37] 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. *arXiv preprint arXiv:2303.08340*, 2023. 2, 7, 8

[38] Fabio Solari, Manuela Chessa, Narasimha K. Medathati, and Pierre Kornprobst. What can we expect from a v1-mt feedforward architecture for optical flow estimation? *Signal Processing: Image Communication*, 39:342–354, 2015. 2, 7, 8

[39] Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1390–1399, 2018. 1

[40] Zitang Sun, Yen-Ju Chen, Yung-Hao Yang, and Shin'ya Nishida. Modeling human visual motion processing with trainable motion energy sensing and a self-attention network. In *Advances in Neural Information Processing Systems*, 2023. 2, 7, 8

[41] Zitang Sun, Yen-Ju Chen, Yung-Hao Yang, and Shin'ya Nishida. Acquisition of second-order motion perception by learning to recognize the motion of objects made by non-diffusive materials. *Journal of Vision, Vision Sciences Society Annual Meeting Abstract*, 24(9), 2024. 7, 8

[42] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419, 2020. 2, 7, 8

[43] Longguang Wang, Yulan Guo, Li Liu, Zaiping Lin, Xinpu Deng, and Wei An. Deep video super-resolution using hr optical flow estimation. *IEEE Transactions on Image Processing*, 29:4323–4336, 2020. 1

[44] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. *arXiv preprint arXiv:2003.14338*, 2020. 3, 4, 5

[45] H. Wu, Y. Liu, Z. Zhang, L. Du, L. Wang, Y. Wang, D. Lin, and L. Lin. Q-Instruct: Improving Low-Level Visual Abilities for Multi-Modality Foundation Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25490–25500, Seattle, WA, USA, 2024. IEEE. 1

[46] Yung-Hao Yang, Zitang Sun, Taiki Fukiage, and Shin'ya Nishida. HuPerFlow: A comprehensive benchmark for human vs. machine motion estimation comparison. https://kucognitiveinformaticslab.github.io/Huperflow-Website/. Accessed: March 2025. 5, 8

[47] Yung-Hao Yang, Taiki Fukiage, Zitang Sun, and Shin'ya Nishida. Psychophysical measurement of perceived motion flow of naturalistic scenes. *iScience*, 26(12), 2023. 2, 3, 6, 7, 8

[48] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1

[49] Xizhou Zhu, Yuwen Xiong, Bo Dai, and Dahua Lin. Deep feature flow for video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2349–2358, 2017. 8

HuPerFlow: A Comprehensive Benchmark for Human vs. Machine Motion Estimation Comparison

Supplementary Material

Appendix 1 Frame onset errors. In online psychophysical experiments on motion perception, presenting motion stimuli with precise onset timing is essential. Due to varying display configurations across devices, especially with refresh rates spanning $\tilde{3}0$ to 120 Hz and adaptive refresh rate (ARR) features on some modern devices, adjustments from typical offline experiments were required. To accommodate these differences without imposing strict device requirements, we presented visual stimuli at 30 Hz (33.33 ms per frame) and controlled frame onset timing using the `Clock.getTime` function in PsychoPy. To assess device stability among participants, frame drops and frame onset times were logged. Since each trial included at least three repetitions with varying lengths, we recorded every frame of image presentation only during the third repetition of each trial in the main experiment (see the Procedure section). Fig. S1 shows the distribution of frame onset errors, calculated as deviations from ideal frame onsets (e.g., 33.33 ms, 66.67 ms). The frame drop rate was 0.041%, and the average frame onset error was -0.055 ms (SD = 3.664 ms), indicating consistent and accurate visual stimulus presentation across devices.

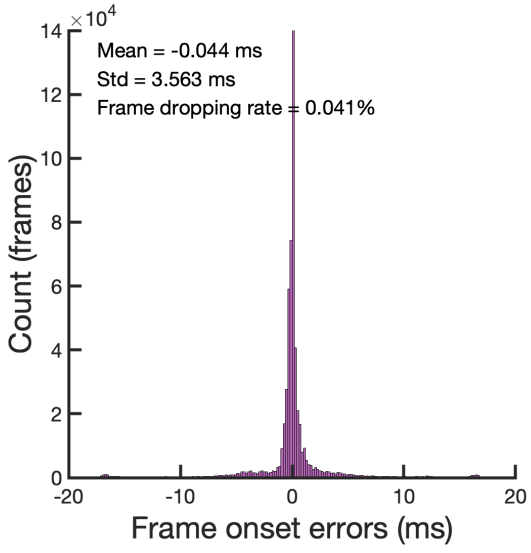

Figure S1. Distribution of frame onset errors

Appendix 2 Correlations between human responses and ground truth. Fig. S2 presents scatter plots comparing human responses to ground truth data across each dataset. Each point represents the average response from the same locations across four repetitions within an observer.

Appendix 3 Response Consistency Index (RCI). Response Consistency Index (RCI) is an index from [47] to evaluate the similarity between model performance and human flow illusions at each probed location.

The RCI is defined as the product of $A \cdot B \cdot C$ in Equations S1, S2, and S3, measuring the relative alignment of ground truth (G), human response (R), model prediction (M), and the origin (O).

- A (Equation S1) quantifies the deviation of human responses from the ground truth.
- B (Equation S2) indicates the directional similarity between the response error vector $\overrightarrow{GR}$ and the model error vector $\overrightarrow{GM}$ relative to the ground truth.
- C (Equation S3) compares the distance between model prediction and ground truth $\|\overrightarrow{GM}\|$ with the distance between model prediction and response $\|\overrightarrow{RM}\|$.

The RCI approaches +1 when the model's prediction aligns closely with human flow illusions and approaches -1 when the prediction diverges in the opposite direction.

$$A = \frac{\|\overrightarrow{GR}\|}{\|\overrightarrow{OG}\| + \|\overrightarrow{OR}\|}. \tag{S1}$$

$$B = \frac{\overrightarrow{GR} \cdot \overrightarrow{GM}}{\|\overrightarrow{GR}\|\|\overrightarrow{GM}\|}. \tag{S2}$$

$$C = 0.5\left(\frac{\|\overrightarrow{GM}\| - \|\overrightarrow{RM}\|}{\|\overrightarrow{GM}\| + \|\overrightarrow{RM}\|} + 1\right) = \frac{\|\overrightarrow{GM}\|}{\|\overrightarrow{GM}\| + \|\overrightarrow{RM}\|}. \tag{S3}$$

Appendix 4 The predictions of optical flow algorithms in each dataset. To assess which datasets align better between optical flow algorithms and human responses, and which show greater discrepancies, we compare algorithm predictions with human responses and ground truth across datasets. Following the approach in Table 2, we calculated indices for each dataset separately and presented them in Tables S1–S12. Overall, the findings align with the trends in Table 2: biologically inspired models (e.g., FFV1MT, V1Attention-StageI) show stronger partial correlations with human responses, while multi-frame models like VideoFlow align more closely with ground truth in dynamic datasets.

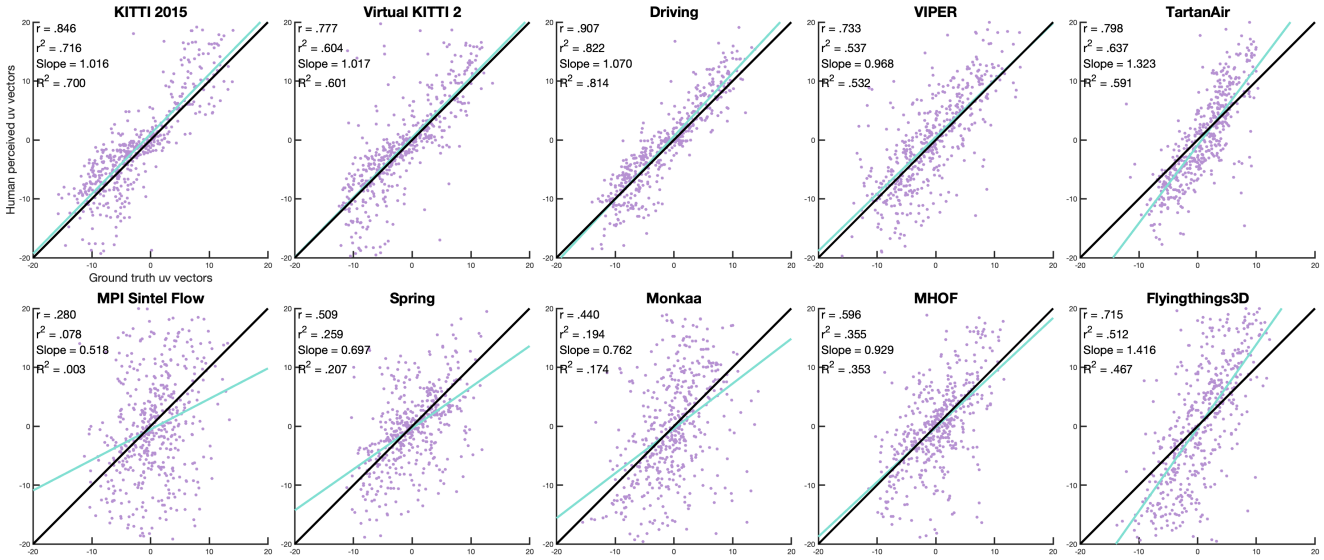

Figure S2. The scatter plots display the relationship between human-perceived UV vectors and ground truth UV vectors from each dataset. In the top-left corner, the Pearson correlation coefficient (r), the squared correlation (r^2), the slope of the linear regression line, and the coefficient of determination (R^2) relative to the ideal response (i.e., the ground truth) are shown.

Video S1 Demonstration of the procedure. This video shows how observers use a mouse to control the direction and speed of circular Brownian noise to match the target motion. After each response, vectors indicating ground truth and observers' responses are displayed during practice trials.

Video S2 Movies of human-perceived flows. These videos display human-perceived flows across the image sequences used in the ten datasets. The video speed (5 Hz) is slower than the actual perceived speed (30Hz) to aid visualization

Table S1. Partial Correlation of uv vectors across Models and Datasets.

Models / Datasets	KITTI 2015	Virtual KITTI 2	Driving	VIPER	TartanAir	MPI Sintel	Spring	Monkaa	MHOF	Flythings3D	All Datasets
Ground truth	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Farnebäck	0.0635	-0.0029	0.1029	0.1806	0.2684	0.1056	0.0440	0.0605	0.2838	0.1504	0.1670
FFV1MT	0.2068	**0.2308**	**0.2246**	**0.3973**	0.3350	0.1568	**0.2839**	**0.2534**	**0.3670**	**0.3168**	**0.3063**
RAFT	-0.0251	0.0957	0.1695	0.2817	0.1287	0.0815	0.1224	0.1632	0.2924	0.1480	0.1090
FlowNet2	0.0003	0.0454	0.1514	0.2712	0.2071	0.0401	0.0235	0.1050	0.3020	0.1538	0.1012
FlowFormer	-0.0229	0.0112	0.2137	0.2811	0.0657	0.0597	0.0698	0.1375	0.2771	0.1153	0.0748
VideoFlow	0.0079	0.0352	0.1798	0.2893	0.0582	0.1390	-0.0014	0.2205	0.2945	0.0847	0.0784
Perceiver IO	0.0402	0.0959	0.1373	0.3256	0.2104	0.0249	-0.0719	0.0121	0.2875	0.1858	0.0860
AGFlow	0.0581	0.1231	0.1954	0.2570	0.1030	0.0554	0.2545	0.1528	0.3603	0.1975	0.1589
V1Attention-Stage1	0.0471	-0.0588	0.1094	0.2835	**0.3661**	0.1646	0.1899	0.1802	0.3507	0.3052	0.2285
DualModel	**0.2677**	-0.0026	0.0578	0.2371	0.3457	**0.1980**	0.0016	0.1612	0.2495	0.2708	0.1973

Table S2. Partial Correlation of Direction across Models and Datasets.

Models / Datasets	KITTI 2015	Virtual KITTI 2	Driving	VIPER	TartanAir	MPI Sintel	Spring	Monkaa	MHOF	Flythings3D	All Datasets
Ground truth	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Farnebäck	-0.0867	0.0625	0.0529	0.0245	0.0206	0.1323	0.1681	0.1537	0.3345	0.1414	0.1424
FFV1MT	0.2775	0.4425	0.0464	**0.4500**	0.2075	0.1959	**0.2863**	0.1635	0.3175	0.2327	0.2756
RAFT	0.2411	0.4133	-0.0667	0.2428	-0.0364	-0.0532	0.1160	0.2310	0.3679	**0.3528**	0.2219
FlowNet2	**0.3628**	0.4406	0.0527	0.1431	0.0625	0.1289	0.2039	0.2164	**0.4359**	0.2249	0.2600
FlowFormer	0.0828	0.2904	0.1116	0.2819	0.0831	-0.1337	0.1933	0.1942	0.2785	0.1884	0.1878
VideoFlow	0.0644	0.2906	0.1597	0.2035	0.0403	0.0331	0.2285	0.2047	0.3675	0.2460	0.2188
Perceiver IO	-0.1114	0.2931	0.1857	0.0971	-0.0912	-0.0480	0.1029	0.1359	0.3593	0.2845	0.1698
AGFlow	0.2389	**0.4662**	**0.2403**	0.4238	0.0157	0.0935	0.0869	0.1815	0.2575	0.0952	0.2417
V1Attention-Stage1	0.1863	0.3324	0.0216	0.3750	**0.2954**	0.1674	0.2747	**0.2562**	0.3980	0.2972	**0.2845**
DualModel	0.3623	0.1425	0.0722	0.3110	0.2883	**0.2721**	0.1341	0.1630	0.3050	0.3391	0.2451

Table S3. Partial Correlation of Speed across Models and Datasets.

Models / Datasets	KITTI 2015	Virtual KITTI 2	Driving	VIPER	TartanAir	MPI Sintel	Spring	Monkaa	MHOF	Flythings3D	All Datasets
Ground truth	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Farnebäck	0.0768	-0.0505	0.1221	0.0382	0.2171	-0.0052	-0.0321	**0.0777**	0.0314	-0.0303	0.0410
FFV1MT	0.0679	-0.0050	0.0643	0.1513	0.2450	0.0194	**0.1046**	-0.1823	0.0908	0.1341	**0.0604**
RAFT	-0.0024	0.2023	0.2298	0.0132	0.1562	0.0879	0.0092	-0.0868	0.0503	-0.0086	0.0128
FlowNet2	0.0044	0.0446	0.2232	0.0373	0.3099	0.0367	0.0232	-0.0058	0.0581	-0.0007	0.0002
FlowFormer	-0.0322	0.0996	**0.2542**	0.0402	0.0029	0.0848	-0.1446	-0.0478	0.0382	-0.0445	-0.0232
VideoFlow	0.0155	0.0997	0.2192	-0.0130	0.0329	**0.1594**	-0.1301	-0.0887	0.0044	-0.0379	-0.0249
Perceiver IO	0.1009	**0.2050**	0.1456	0.0838	0.1999	0.1165	0.0281	-0.0225	0.0755	0.0381	0.0296
AGFlow	0.0026	0.1580	0.2288	0.0735	0.0872	0.0950	0.0750	-0.1143	0.1255	0.1451	0.0139
V1Attention-Stage1	-0.0401	-0.0801	-0.0868	**0.1769**	**0.3956**	-0.0274	-0.0580	-0.0360	**0.1400**	0.0870	0.0202
DualModel	**0.1981**	0.1170	0.0436	0.0697	0.3298	0.0240	-0.1591	-0.1164	0.0561	**0.1613**	0.0589

Table S4. Response Consistent Index (RCI) across Models and Datasets. The RCI is introduced from [47] to represent the model's similarity to human perception (the larger, the more human-aligned).

Models / Datasets	KITTI 2015	Virtual KITTI 2	Driving	VIPER	TartanAir	MPI Sintel	Spring	Monkaa	MHOF	Flythings3D	All Datasets
Ground truth	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Farnebäck	0.0388	0.0725	0.0493	0.0764	0.0722	0.2515	0.1297	0.2342	0.1569	0.1973	0.1279
FFV1MT	0.0706	0.0855	0.0553	0.1472	0.0697	0.2452	0.1749	0.2082	0.2274	0.2152	0.1499
RAFT	0.0121	0.0275	0.0134	0.0250	0.0238	0.0645	0.0731	0.0593	0.1290	0.0480	0.0476
FlowNet2	0.0250	0.0439	0.0190	0.0419	0.0322	0.1289	0.0793	0.0813	0.1544	0.0757	0.0682
FlowFormer	0.0106	0.0308	0.0131	0.0258	0.0189	0.0414	0.0702	0.0605	0.1137	0.0390	0.0424
VideoFlow	0.0128	0.0280	0.0118	0.0242	0.0151	0.0439	0.0553	0.0512	0.1108	0.0270	0.0380
Perceiver IO	0.0233	0.0370	0.0162	0.0293	0.0306	0.0995	0.0629	0.0579	0.1305	0.0477	0.0535
AGFlow	0.0498	0.0597	0.0326	0.0895	0.0439	0.1729	0.1359	0.1315	0.2226	0.1397	0.1078
V1Attention-Stage1	**0.0857**	**0.0927**	**0.0747**	**0.1635**	**0.1530**	**0.3061**	**0.1870**	**0.2647**	**0.2379**	**0.2426**	**0.1808**
DualModel	0.0546	0.0692	0.0441	0.1011	0.0635	0.2226	0.1247	0.2065	0.2265	0.1576	0.1270

Table S5. Correlation of uv vectors across Models and Datasets (v.s. Human).

Models / Datasets	KITTI 2015	Virtual KITTI 2	Driving	VIPER	TartanAir	MPI Sintel	Spring	Monkaa	MHOF	Flythings3D	All Datasets
Ground truth	0.8461	0.7769	0.9068	0.7330	0.7979	0.2802	0.5089	0.4403	0.5962	0.7154	0.6605
Farnebäck	0.7894	0.6119	0.7424	0.6813	0.7861	0.2267	0.3610	0.2925	0.5689	0.4227	0.5394
FFV1MT	0.7976	0.7518	0.8447	0.7343	**0.8155**	0.2620	**0.5330**	0.4472	0.5622	0.5838	0.6263
RAFT	0.8431	**0.7614**	0.9069	0.7576	0.8006	0.2902	0.5098	0.4641	0.6217	0.7086	0.6425
FlowNet2	0.8321	0.7470	0.8999	0.7526	0.8075	0.2083	0.4359	0.3674	0.6166	0.6621	0.5851
FlowFormer	0.8436	0.7276	0.9068	0.7576	0.7948	0.2857	0.4971	0.4564	0.6233	0.7046	0.6362
VideoFlow	0.8428	0.7466	**0.9098**	0.7587	0.7934	0.3030	0.4938	**0.4786**	**0.6261**	**0.7158**	**0.6427**
Perceiver IO	0.8393	0.7500	0.8975	**0.7652**	0.8002	0.2423	0.4361	0.4029	0.6110	0.7031	0.6188
AGFlow	0.7717	0.7237	0.9000	0.7227	0.7854	0.2033	0.5159	0.3310	0.5320	0.5178	0.5372
V1Attention-Stage1	0.7036	0.6271	0.6877	0.6462	0.7523	0.2434	0.4364	0.3676	0.5930	0.5809	0.5478
DualModel	**0.8507**	0.6638	0.6501	0.7127	0.8146	**0.3046**	0.3369	0.3660	0.5025	0.6204	0.5704

Table S6. Correlation of Direction across Models and Datasets (v.s. Human).

Models / Datasets	KITTI 2015	Virtual KITTI 2	Driving	VIPER	TartanAir	MPI Sintel	Spring	Monkaa	MHOF	Flythings3D	All Datasets
Ground truth	0.5358	0.4939	0.4850	0.4927	0.4856	0.1962	0.3040	0.1912	0.2899	0.6483	0.4286
Farnebäck	0.4220	0.3683	0.3635	0.4151	0.4386	0.2057	0.3112	0.2327	0.4141	0.3728	0.3762
FFV1MT	0.4746	0.6095	0.4577	0.5869	0.5179	0.2560	**0.4017**	0.2348	0.3542	0.4515	0.4534
RAFT	0.5693	0.6105	0.4429	0.5359	0.4675	0.1451	0.3079	0.2865	0.4508	**0.7017**	0.4694
FlowNet2	0.6136	0.6243	0.4744	0.4983	0.4794	0.2266	0.3522	0.2848	**0.5079**	0.6314	**0.4841**
FlowFormer	0.5223	0.5404	0.4943	0.5434	0.4796	0.1484	0.3552	0.2682	0.3847	0.6523	0.4560
VideoFlow	0.5127	0.5486	0.5038	0.5234	0.4750	0.1980	0.3567	0.2712	0.4488	0.6738	0.4711
Perceiver IO	0.4872	0.5476	0.4991	0.4947	0.4387	0.1059	0.2996	0.2329	0.4482	0.6628	0.4400
AGFlow	0.5188	**0.6348**	**0.5286**	**0.5961**	0.4577	0.1830	0.2410	0.2555	0.2924	0.5326	0.4468
V1Attention-Stage1	0.4620	0.5392	0.3376	0.5549	0.5334	0.2124	0.3851	**0.3048**	0.4459	0.5100	0.4473
DualModel	**0.6169**	0.4556	0.4106	0.5497	**0.5466**	**0.3281**	0.2775	0.2420	0.3833	0.5656	0.4510

Table S7. Correlation of Speed across Models and Datasets (v.s. Human).

Models/Datasets	KITTI 2015	Virtual KITTI 2	Driving	VIPER	TartanAir	MPI Sintel	Spring	Monkaa	MHOF	Flythings3D	All Datasets
Ground truth	0.3432	0.3674	0.5136	0.2869	0.5685	0.2559	0.3954	0.0888	0.3622	0.2659	0.2876
Farnebäck	0.2860	0.1420	0.3264	0.1726	0.5013	0.0753	0.1426	**0.0961**	0.2261	0.0169	0.1476
FFV1MT	0.2687	0.2547	0.4008	0.2797	0.5750	0.1321	0.2992	-0.1305	0.2161	0.2205	0.2033
RAFT	0.3383	0.4113	0.5500	0.2627	0.5799	0.2627	0.3561	0.0263	**0.3145**	0.2274	**0.2337**
FlowNet2	0.3177	0.3348	0.5441	0.2677	**0.6215**	0.1811	0.3303	0.0203	0.2728	0.1495	0.1760
FlowFormer	0.3350	0.3562	**0.5547**	0.2709	0.5422	0.2673	0.2911	0.0439	0.2964	0.2041	0.2133
VideoFlow	0.3396	0.3732	0.5444	0.2599	0.5464	**0.2949**	0.3314	0.0385	0.2639	0.2434	0.2186
Perceiver IO	0.3552	**0.4115**	0.5254	**0.2947**	0.5796	0.2687	**0.3581**	0.0492	0.2961	0.2365	0.2327
AGFlow	0.2167	0.3680	0.5438	0.2360	0.5280	0.2119	0.2130	-0.0811	0.2048	0.2010	0.1452
V1Attention-Stage1	0.1407	0.1624	0.1529	0.2742	0.5649	0.0522	0.1148	-0.0085	0.2860	0.1539	0.1381
DualModel	**0.3885**	0.3271	0.3070	0.2232	0.6132	0.1480	0.0793	-0.0746	0.1697	**0.2671**	0.2000

Table S8. End-Point Error (v.s. Human).

Models/Datasets	KITTI 2015	Virtual KITTI 2	Driving	VIPER	TartanAir	MPI Sintel	Spring	Monkaa	MHOF	Flythings3D	All Datasets
Ground truth	5.1482	5.5497	3.8906	6.4759	5.8027	10.3036	6.7171	9.7504	6.7639	9.2317	6.9634
Farnebäck	5.5556	6.3728	5.4047	6.3994	5.2625	8.2336	9.0176	7.5378	5.7985	9.5553	6.9138
FFV1MT	5.3739	**5.5053**	4.3577	**6.1055**	5.5255	9.5621	7.3843	8.3927	5.2278	8.3308	6.5766
RAFT	5.1886	5.6502	3.8650	6.2569	5.6337	10.1466	12.3883	9.4856	5.9235	9.1406	7.3679
FlowNet2	5.2660	5.6927	3.9814	6.1794	5.4872	11.0885	12.3818	10.2964	5.6786	9.4780	7.5530
FlowFormer	5.1739	5.8828	3.8653	6.2027	5.7910	10.3444	12.2990	9.4048	6.0709	9.2916	7.4326
VideoFlow	5.1594	5.8054	**3.8246**	6.3017	5.8205	10.2044	12.6882	9.4029	6.2865	9.2238	7.4717
Perceiver IO	5.2073	5.7400	4.2661	6.5664	5.7617	10.4737	13.1479	10.1592	6.1923	10.0008	7.7515
AGFlow	5.2642	6.0433	3.9644	6.8273	5.9579	13.3879	9.2022	13.1371	5.6402	9.9727	7.9397
V1Attention-Stage1	6.1725	6.1028	5.9365	6.4236	**4.7982**	**8.0337**	**6.4705**	**7.4639**	**4.5903**	**7.6485**	**6.3640**
DualModel	**5.0696**	6.1217	6.2542	6.5805	5.7466	9.9290	9.8038	8.6702	5.2384	9.6581	7.3072

Table S9. Correlation of uv vectors across Models and Datasets (v.s. GT).

Models/Datasets	**KITTI 2015**	**Virtual KITTI 2**	**Driving**	**VIPER**	**TartanAir**	**MPI Sintel**	**Spring**	**Monkaa**	**MHOF**	**Flythings3D**	**All Datasets**
Ground Truth	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Farnebäck	0.9170	0.7890	0.7893	0.8381	0.8947	0.4947	0.6531	0.5624	0.6709	0.4604	0.6769
FFV1MT	0.8811	0.8783	0.8824	0.7637	0.9273	0.4576	0.7085	0.6038	0.5213	0.5596	0.6994
RAFT	0.9975	**0.9577**	0.9878	0.9689	**0.9890**	0.9424	0.9214	0.9500	0.8141	0.9420	0.9259
FlowNet2	0.9834	0.9499	0.9775	0.9429	0.9852	0.6377	0.8346	0.6767	0.7793	0.8453	0.8201
FlowFormer	**0.9980**	0.9332	0.9804	0.9672	0.9887	0.9701	0.9350	0.9437	**0.8473**	0.9496	0.9325
VideoFlow	0.9956	0.9522	**0.9949**	**0.9723**	0.9874	**0.9745**	**0.9710**	**0.9580**	0.8274	**0.9877**	**0.9437**
Perceiver IO	0.9880	0.9385	0.9757	0.9671	0.9564	0.8155	0.9080	0.9045	0.7848	0.9060	0.8929
AGFlow	0.8958	0.8852	0.9707	0.8672	0.9634	0.5697	0.7112	0.4780	0.4621	0.5647	0.6811
V1Attention-Stage1	0.8143	0.8335	0.7233	0.6915	0.7646	0.3377	0.6009	0.5214	0.6265	0.5663	0.6270
DualModel	0.9560	0.8555	0.6976	0.8601	0.9166	0.4997	0.6599	0.5585	0.5659	0.6710	0.7045

Table S10. Correlation of direction across Models and Datasets (v.s. GT).

Models/Datasets	KITTI 2015	Virtual KITTI 2	Driving	VIPER	TartanAir	MPI Sintel	Spring	Monkaa	MHOF	Flythings3D	All Datasets
Ground Truth	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Farnebäck	0.8578	0.6635	0.6795	0.8176	0.8859	0.4613	0.6039	0.5672	0.4334	0.4246	0.6497
FFV1MT	0.5095	0.6277	0.9088	0.5053	0.9516	0.4131	0.6109	0.5022	0.1931	0.4528	0.5882
RAFT	0.8877	0.7907	0.9505	0.9090	**0.9768**	0.8702	0.7902	0.8207	0.5113	0.9218	0.8455
FlowNet2	0.8091	**0.8201**	0.9478	0.9024	0.9532	0.6869	0.7066	0.7286	**0.5858**	0.8244	0.8022
FlowFormer	0.9254	0.7648	0.9709	**0.9547**	0.9343	0.9547	0.8699	0.7850	0.5749	0.9190	0.8663
VideoFlow	0.9162	0.8127	**0.9759**	0.9317	0.9570	**0.9674**	**0.9502**	**0.8264**	0.4912	**0.9745**	**0.8804**
Perceiver IO	**0.9590**	0.7970	0.8558	0.9510	0.9531	0.7089	0.7865	0.7916	0.5788	0.8422	0.8237
AGFlow	0.6989	0.6953	0.9185	0.6266	0.9323	0.5386	0.5687	0.5719	0.1709	0.7473	0.6595
V1Attention-Stage1	0.6356	0.6443	0.6671	0.5927	0.7415	0.2787	0.5446	0.3750	0.2747	0.4807	0.5384
DualModel	0.8732	0.7592	0.7622	0.7564	0.8663	0.4785	0.5667	0.5900	0.3989	0.5364	0.6676

Table S11. Correlation of Speed across Models and Datasets (v.s. GT).

Models/Datasets	KITTI 2015	Virtual KITTI 2	Driving	VIPER	TartanAir	MPI Sintel	Spring	Monkaa	MHOF	Flythings3D	All Datasets
Ground Truth	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Farnebäck	0.6792	0.4975	0.4538	0.4901	0.6405	0.3128	0.4281	0.2353	0.5572	0.1720	0.3874
FFV1MT	0.6401	0.7025	0.7040	0.5543	0.7976	0.4508	0.5546	0.4023	0.3805	0.3798	0.5374
RAFT	0.9867	**0.9251**	0.9257	0.8959	0.9486	0.8532	0.8910	0.8337	**0.7886**	0.8707	0.7865
FlowNet2	0.9211	0.8525	0.8875	0.8719	0.9390	0.5962	0.8033	0.2909	0.6382	0.5644	0.6113
FlowFormer	**0.9892**	0.8283	0.9574	0.8804	**0.9526**	0.9162	0.8897	0.8095	0.7537	0.8523	0.7890
VideoFlow	0.9814	0.9133	**0.9679**	**0.9224**	0.9458	**0.9277**	**0.9407**	**0.8887**	0.7207	**0.9560**	**0.8088**
Perceiver IO	0.9432	0.8330	0.9403	0.9127	0.8847	0.7684	0.8742	0.7270	0.6739	0.8083	0.7435
AGFlow	0.6258	0.7275	0.8710	0.6326	0.8656	0.5217	0.3775	0.3071	0.2527	0.2461	0.4640
V1Attention-Stage1	0.5048	0.6036	0.4286	0.4192	0.4968	0.3028	0.4131	0.2899	0.4719	0.2754	0.4192
DualModel	0.8283	0.6707	0.5363	0.5897	0.7798	0.5001	0.5170	0.3714	0.3324	0.4971	0.5288

Table S12. End-Point Error (v.s. GT).

Models/Datasets	KITTI 2015	Virtual KITTI 2	Driving	VIPER	TartanAir	MPI Sintel	Spring	Monkaa	MHOF	Flythings3D	All Datasets
Ground truth	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Farnebäck	2.3448	4.3035	4.0169	3.3315	2.2899	8.0080	7.3517	7.7063	5.2443	10.5800	5.5177
FFV1MT	3.3863	3.5651	2.9159	5.6433	2.3130	9.3565	6.8494	7.7386	7.0360	9.6041	5.8408
RAFT	0.4900	**1.1488**	0.5472	0.8620	0.6431	2.1057	8.1461	2.0466	3.8601	2.0892	2.1939
FlowNet2	1.1011	1.8750	0.9257	1.4860	0.7700	5.8037	8.4985	3.7901	4.4807	3.7035	3.2434
FlowFormer	**0.4159**	1.6521	0.5475	0.8451	0.6086	1.4738	7.9602	1.9914	**3.5032**	1.7794	2.0777
VideoFlow	0.4811	1.3266	**0.4025**	**0.7785**	**0.5140**	**1.4016**	7.7573	**1.7247**	3.6526	**1.0431**	**1.9082**
Perceiver IO	0.9487	1.7568	1.0664	1.2783	1.0709	4.0563	8.5428	2.7422	4.1413	3.0831	2.8687
AGFlow	2.1481	2.9862	1.3214	4.0459	1.6407	10.2528	7.7348	9.6162	7.3885	7.2425	5.4377
V1Attention-Stage1	4.9537	4.9182	5.9380	6.8169	5.0388	10.1230	**6.5684**	8.8315	6.4919	10.3447	7.0025
DualModel	2.2578	3.7184	4.5595	4.3126	2.3973	8.8834	7.8088	7.6007	6.7577	8.2937	5.6590