# Learning Omnidirectional Flow in 360° Video via Siamese Representation

Keshav Bhandari[1], Bin Duan[2], Gaowen Liu[3], Hugo Latapie[3], Ziliang Zong[1], and Yan Yan[2⋆]

[1]Texas State University [2]Illinois Institute of Technology [3]Cisco Research

**Abstract.** Optical flow estimation in omnidirectional videos faces two significant issues: the lack of benchmark datasets and the challenge of adapting perspective video-based methods to accommodate the omnidirectional nature. This paper proposes the first perceptually natural-synthetic omnidirectional benchmark dataset with a 360° field of view, FLOW360, with 40 different videos and 4,000 video frames. We conduct comprehensive characteristic analysis and comparisons between our dataset and existing optical flow datasets, which manifest perceptual realism, uniqueness, and diversity. To accommodate the omnidirectional nature, we present a novel Siamese representation Learning framework for Omnidirectional Flow (SLOF). We train our network in a contrastive manner with a hybrid loss function that combines contrastive loss and optical flow loss. Extensive experiments verify the proposed framework's effectiveness and show up to 40% performance improvement over the state-of-the-art approaches. Our FLOW360 dataset and code are available at https://siamlof.github.io/.

**Keywords:** 360° Optical Flow Dataset, Siamese Flow Estimation

## 1 Introduction

Optical flow estimation, as a fundamental problem in computer vision, has been studied over decades by early works [44,34] dated back to 80*s*. Before the era of modern deep learning, traditional optical flow estimation methods relied on hand-crafted features based optimizations [49,5,17], energy-based optimizations [33,14,56] and variational approaches [15,28,66]. Although deep learning-based approaches [41,61,70,59,39,35] have shown great advantages over these classical approaches, most of them are specially tailored for perspective videos. The availability of perspective optical flow datasets [16,29,30,48,6] heavily supports the advancement of these modern deep learning-based approaches. The optical flow datasets are difficult to obtain and requires the generation of naturalistic synthetic dataset like Sintel [16]. As these datasets mark the foundation for optical flow estimation research, the availability of reliable omnidirectional

⋆ Corresponding author.

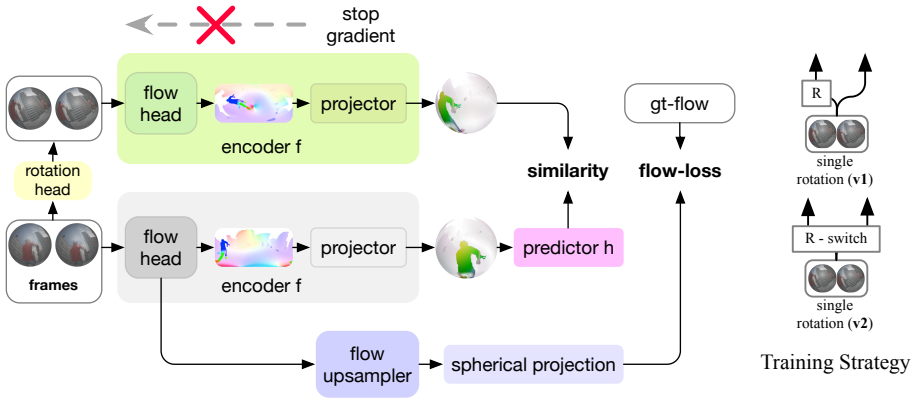



**Fig. 1. Siamese Representation Learning for Omnidirectional Flow (SLOF).** Pairs of frame sequence (w/ and w/o random rotation) are passed as inputs to encoder $f$ (RAFT as a flow head backbone and a standard convolutional projector layer). A predictor layer $h$ is an MLP layer. The entire framework is trained by fusing the pre-training and fine-tuning stage to combine the similarity and flow-loss in a single stage. The model maximizes the similarity between latent representations of flow information from two streams and minimizes the flow loss. **Training Strategy (right):** Here two different arrows$(left, right)$ represent siamese streams or input pathways to our model. **v1** and **v2** (either of the stream is subjected to rotational augmentation) are similar strategies achieving overall better performance.

datasets is equally important to advance the omnidirectional flow estimation research. The need for the datasets brings up the first challenge: there is no such reliable (perceptually natural and complex) 360° or omnidirectional video dataset in the literature collected for omnidirectional optical flow estimation. Another challenge of omnidirectional optical flow estimation is that current perspective video-based deep networks fail to accommodate the nature of 360° videos. These perspective optical flow estimation methods inevitably require fine-tuning due to the presence of radial distortion [4] on 360° videos. This fine-tuning task is effort-intensive and requires several transformation techniques to adapt the distortion [58,23]. An intuitive solution is to fine-tune perspective-based deep networks under omnidirectional supervised data. However, this brute-force migration of perspective-based networks often requires enormous supervision and still leads to significant performance degradation [9].

We address the first challenge of reliable benchmark dataset shortage by proposing a new dataset named FLOW360. To the best of our knowledge, this is the first perceptually natural-synthetic 360° video dataset collected for omnidirectional flow estimation. Currently, existing omnidirectional datasets face two significant issues i.e., lack of full 360° FOV (field of view) and lack of perceptual realism. Specifically, OmniFlow[52] dataset only has 180° FOV failing to address the omnidirectional nature, while the dataset proposed in OmniFlowNet[3] lacks perceptual realism in scene and motion. Meanwhile, perspective optical flow

datasets such as [6,16,29] have facilitated researchers in investigating perspective optical flow estimation methods [35,22,39,70,61], where the availability of such omnidirectional videos dataset is essential to advance this particular field. It is worth noting that FLOW360 dataset can be used in various other areas such as continuous flow estimation in 3-frame settings with forward and backward consistency [43,46,38], depth [71,25] and normal map estimation [65].

The accommodation to the omnidirectional nature generally requires modification of convolution layers and further refinements on the target dataset due to the presence of radial distortions [9], which is caused by projecting 360° videos (spherical) to an equirectangular plane. Existing works design various convolution layers to address the distortion problem, such as spherical convolution [58,11,20,57], spectral convolution [19,24] and tangent convolution [23]. Although these methods can achieve better performance than classical CNNs, they require immense effort with layer-wise architecture design, which is impractical for high-demanding deployment in the real-world setting.

Instead of adding new convolution layers, we design a novel SLOF (**S**iamese representation **L**earning for **O**mnidirectional **F**low) framework (Fig. 1), which leverages the rotation-invariant property of omnidirectional videos to address the radial distortion problem. The term rotation-invariant here implies that 360° videos are rotated in a random projection such that the reverse rotation of such projection is equal to the original projection. This rotation-invariant property ensures that omnidirectional videos can be projected to a planar representation with infinite projections by rotating the spherical videos on three different axis $(X, Y, Z)$, namely "pitch", "roll" and "yaw" operations preserving overall information. Specifically, we design a siamese representation learning framework for learning omnidirectional flow from a pair of consecutive frames and their rotated counterparts, assuming that the representations of these two cases are similar enough to generate nearly identical optical flow in the spherical domain. Besides, we design and compare different combinations of rotational augmentation and derive guidelines for selecting the most effective augmentation scheme.

To summarize, we make three major contributions in this paper: (**i**) we introduce FLOW360, a new optical flow dataset for omnidirectional videos, to fill the dataset's need to advance the omnidirectional flow estimation field. (**ii**) We propose SLOF, a novel framework for optical flow estimation in omnidirectional videos, to mitigate the cumbersome framework adjustments for omnidirectional flow estimation. (**iii**) We demonstrate a new distortion-aware error measure for performance analysis that incorporates the relative error measure based on distortion. Finally, we compare our method with existing omnidirectional flow estimation techniques via kernel transformation [58] to address radial distortions. The FLOW360 dataset, the SLOF framework, and our experimental results provide a solid foundation for future exploration in this important field.

## 2 Related Work

**Optical Flow Datasets.** Perspective datasets such as [7,6,45,50,42,27] comprise synthetic image sequences along with synthetic and hand-crafted optical flow. However, these datasets fall short in terms of perceptual realism and complexities. Even though several optical flow datasets have been published recently in [47,29,30,48], they are primarily used in automotive driving scenarios. The other relevant dataset in the literature was Sintel [16], which provided a bridge to contemporary optical flow estimation and synthetic datasets that can be used in real-world situations.

All datasets, as mentioned earlier, are introduced for perspective videos thus cannot be used for omnidirectional flow estimation. So to address this problem, LiteFlowNet360 [9] on omnidirectional flow estimation was released to augment the Sintel dataset by introducing distortion artifacts for the domain adaptation task. Nevertheless, these augmented datasets are discontinuous around the edges and violate the 360° nature of omnidirectional videos. The closest datasets to ours are OmniFlow [52] and OmniFlowNet [3]. OmniFlow introduced a synthetic 180° FOV dataset, which is limited to indoor scenes and lacks full 360° FOV. Similarly, OmniFlowNet introduced a full 360° FOV dataset. However, both datasets lack complexities and evidence for perceptual realism. We show a detailed comparison of FLOW360, OmniFlow, and OmniFlowNet in Fig. 5. Compared to existing datasets in the literature, FLOW360 is the first perceptually natural benchmark 360° dataset and fills the void in current research.

**Optical Flow Estimation.** Advancements in optical flow estimation techniques largely rely on the success of data-driven deep learning frameworks. Flownet [22] marked one of the initial adoption of CNN- based deep learning frameworks for optical flow estimation. Several other works [39,35,68,63,2,40,62,51] followed the footsteps with improved results. Generally, these networks adopt an encoder-decoder framework to learn optical flow in a coarse-to-fine manner. The current framework RAFT [61] has shown improvements with correlation learning.

The methods mentioned above are insufficient on omnidirectional flow field estimation as they are designed and trained for perspective datasets. One of the initial work [53] on omnidirectional flow estimation was presented as flow estimation by back-projecting image points to the virtually curved retina, thus called back-projection flow. It showed an improvement over classical algorithms. Similarly, another classical approach [21] relyed on spherical wavelet to compute optical flow on omnidirectional videos. However, these methods are limited to classical approaches as they are not relevant in existing deep learning-based approaches. One of the recent works, LiteFlowNet360 [9] tried to compute optical flow on omnidirectional videos using domain adaptation. This method utilized the kernel transformer technique (KTN [58]) to adapt convolution layers on LiteFlowNet [35] and learn correct convolution mapping on spherical data. Similarly, OmniFlowNet [3] proposed a deep learning-based optical flow estimation technique for omnidirectional videos. The major drawback of these methods is the requirement to adapt convolution layers, which takes a substantial amount of time and makes portability a significant issue. For example, in LiteFlowNet360,

each convolution layer in LiteFlowNet was transformed using KTN with additional training and adjustments. Similar to OmniFlowNet, every convolution layer in LiteFlowNet2 [36] was transformed using kernel mapping [26] based on different locations of the spherical image. These techniques incur computational overheads and limit the use of existing architectures. Such approaches demand explicit adaptation of convolution layers, which is hard to maintain when more up-to-date methods are published constantly. Contrary to these methods, we propose a Siamese Representation Learning for Omnidirectional Flow (SLOF) method to learn omnidirectional flow by exploiting existing architectures with designed representation learning objectives, significantly reducing the unnecessary effort of transforming or redesigning the convolution layer.

**Siamese Representation Learning.** Representation learning is a powerful approach in unsupervised learning. Siamese networks have shown great success in different vision-related tasks such as verification [12,60,13] and tracking[8]. A recent approach [18] in siamese representation learning showed impressive results in unsupervised visual representation learning via exploiting different augmentation views of the same data. They presented their work in pre-training and fine-tuning stages, where the former being the unsupervised representation learning. We use the representation learning scheme on omnidirectional data via rotational augmentations, maximizing the similarity for latent representations and minimizing the flow loss.

## 3 FLOW360 Dataset

FLOW360 is an optical flow dataset tailored for 360° videos using Blender [10]. This dataset contains naturalistic 360° videos, forward and backward optical flow, and dynamic depth information. The dataset comprises 40 different videos extracted from huge 3D-World 'The Room', 'Modern', 'Alien Planet', and 'City Rush'. Due to their size, this 3D-World cannot be rendered at once in a single video. We render several parts of this 3D-World, which provides enough qualitative variation in motion and visual perception like 3D-assets, textures, and illuminations. The nature of this large and diverse animated world provides relatively enough diversity to qualify for a standard benchmark dataset. The Fig. 3 shows some of the examples of motion and scene diversity of FLOW360. Similarly, samples from the dataset of different 3D-World are shown in Fig. 2. We build these 3D-World using publicly available 3D models [32,67,37] and 3D animated characters [64,55,1]. Meanwhile, we adopt Blender [10] for additional rigging and animation for the dataset.

FLOW360 contains 40 video clips extracted from different parts of huge 3D-World, 'The Room', 'Alien Planet', 'City Rush', and Modern'. The datasets also contain other information like depth maps and normal fields extracted from the 3D-World. The FLOW360 dataset has 4,000 video frames, 4,000 depth maps, and 3,960 flow fields. We divide the video frames into 2700/1300 train/test split. We render the video frames with the dimension of $(512, 1024)$ to save the rendering

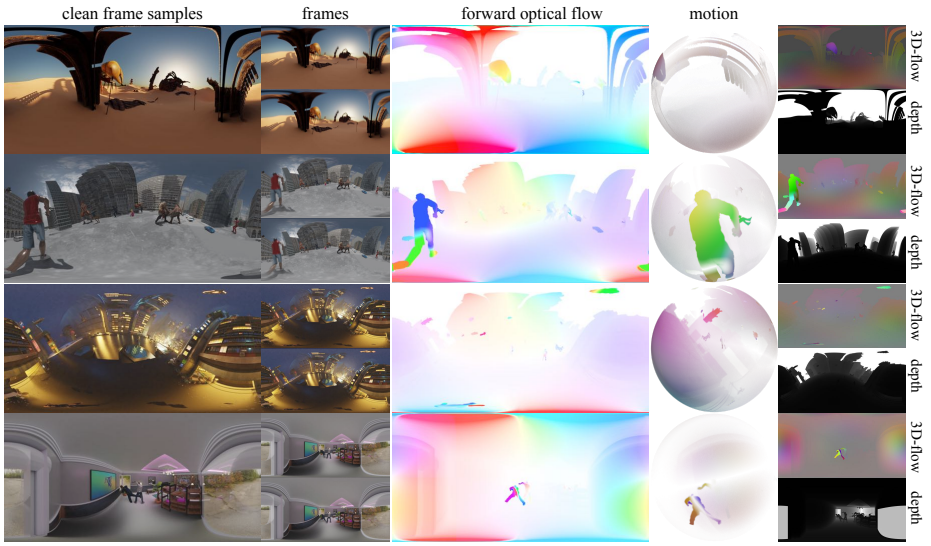



**Fig. 2. The FLOW360 Dataset**. Sample frames (first and second column, respectively) from some of the videos with corresponding forward optical flow and dynamic depth information. Motion in 3D Sphere (fourth column) is computed by transforming the motion vectors from Equirectangular plane $(\theta, \phi)$ to unit sphere $f(x, y, z)$. Motion in the sphere is represented in RGBA color notation. RGB color representation (as suggested in Middlebury [6]) is encoded using $(x, y)$ components, and the alpha color is encoded from $z$ of a unit sphere. RGB encoding (fifth column) is an RGB color map of flow in 3D space. **Note**: flow fields are clipped for better visualization.

time. However, FLOW360 can be rendered with higher resolution, as 3D models and Blender add-ons (provided in supplementary material) will also be public.

**Diversity.** We design FLOW360 datasets to include a diverse situation that resembles the real world scenario as much as possible. The statistical validity of the datasets in terms of perceptual realism of scene and motion is presented in Fig. 5. The datasets contain a wide range of motion complexity from smaller to larger displacement, occlusion, motion blur, and similar complexities on the scene using camera focus-defocus, shadow, reflections, and several distortion combinations. As these complexities are quite common in natural videos, the FLOW360 provides similar complexities. Similarly, the datasets cover diverse scenarios like environmental effects, textures, 3D assets, and diverse illuminations. The qualitative presentation of these diversities and complexities are presented in Fig. 3 and Fig. 4 respectively.

**Fairness.** The FLOW360 dataset contains custom-tailored animated 360 videos. We plan to release the dataset with the 3D models and our custom Blender add-ons to provide researchers a platform to create their custom optical flow datasets for all kinds of environments (perspective, 180° and 360° FOV). However, the release of 3D world scenes can raise questions regarding fairness. To mitigate

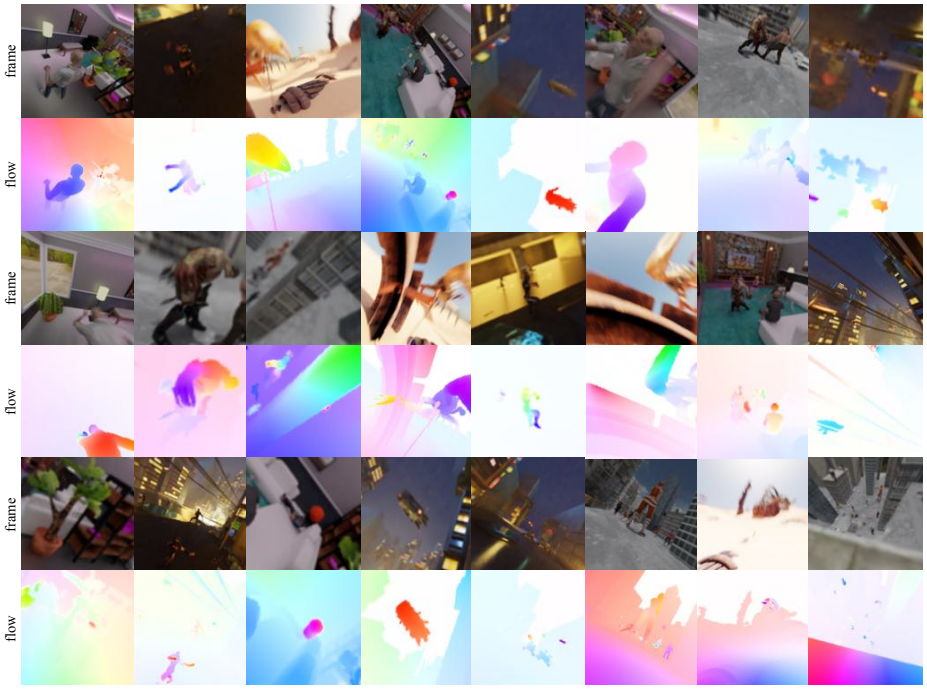
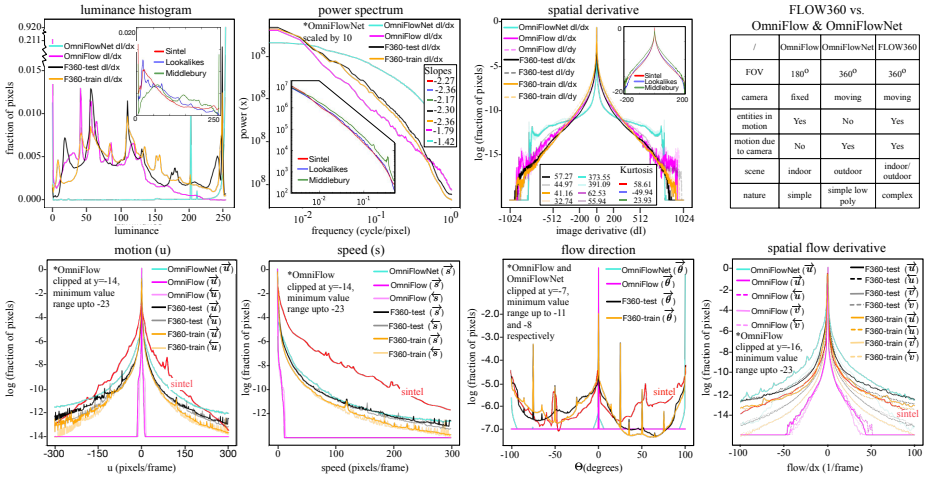



**Fig. 3. Motion and Scene Diversity**. Samples from FLOW360 Dataset with random projection (pitch, roll, yaw, fov) showing scene and motion diversity. The FLOW360 dataset has a vast scene consisting of several lighting scenarios, textures, diverse 3D assets, and motion complexity in different regions.


**Fig. 4. Complexity of FLOW360 Dataset**. Final frames in FLOW360 Dataset include complex characteristics like camera focus/defocus, motion blur, lens distortion, shadow, and reflections. Our dataset provides ambiance occlusion and environmental effects for a realistic visual appearance.

this issue, we will perturb certain parts of 3D world scenes and not release any camera information related to the test set.




| / | OmniFlow | OmniFlowNet | FLOW360 |
|---|---|---|---|
| FOV | $180^0$ | $360^0$ | $360^0$ |
| camera | fixed | moving | moving |
| entities in motion | Yes | No | Yes |
| motion due to camera | No | Yes | Yes |
| scene | indoor | outdoor | indoor/ outdoor |
| nature | simple | simple low poly | complex |

**Fig. 5. Comparision of frames and flow statistics**. Top row represents the frames statistics and comparison with Sintel, Lookalikes, Middlebury, OmniFlow [52] and OmniFlowNet [3]. Bottom row represents flow statistics and comparison with Sintel (red), OmniFlow (magenta) and OmniFlowNet (turquoise). The table on the top-right shows a brief comparision of OmniFlow & OmniFlowNet with FLOW360 dataset. **Note**: ($\rightarrow$, $\leftarrow$) represents forward and backward flow fields, respectively.

**Render Passes.** We exploit several modern features from Blender-v2.92 like advanced ray-tracing as a render engine along with render passes like vector, normal, depth, mist, and so on to produce realistic 3D scenes. Additionally, we incorporate features like ambient occlusion, motion blur, camera focus/defocus, smooth shading, specular reflection, shadow, and camera distortion to introduce naturalistic complexity (shown in Fig. 4) in our dataset. Besides optical flow information, the FLOW360 3D-world may be used to collect several other helpful information like depth, normal maps, and semantic segmentation.

**Dataset Statistics.** We conduct a comprehensive analysis and compare our dataset with Sintel [16], Lookalikes (presented in the original Sintel paper to compare the image statistics with the simulated dataset), Middlebury [6], OmniFlow [52] and OmniFlowNet [3]. The analysis shown in Fig. 5 shows the image and motion statistics in the top and bottom rows, respectively.

Based on analysis from Sintel, we present frame statistics with three different analysis: luminance histogram, power spectrum, and spatial derivative. For luminance statistics, we convert the frames to gray-scale, $I(x, y) \in [0, 255]$ then we compute histograms of gray-scale images across all pixels in the entire dataset. The luminance statistics show the FLOW360 has a similar distribution with the peak in the range between $[0-100]$ and decreasing luminosity beyond that range. Similarly, we estimate power spectra from the 2D FFT of the $512 \times 512$ in the center of each frame. We compute the average of these power spectra across all the datasets. We present power spectra analysis separately for the training and

test set in this analysis. The power spectra analysis closely resembles the Sintel, Lookalikes, and Middlebury datasets. Based on [27,54], the real-world movies exhibit a characteristic of a power spectrum slope around -2, which is equivalent to a $1/f^2$ falloff. FLOW360 with the slope $(-2.30, -2.36)$ on test and training split shows such characteristics. We do not claim that FLOW360 is realistic, but it certainly exhibits perceptual similarity with natural movies. The spatial and temporal derivative analysis additionally supports this characteristic. The Kurtosis of frames spatial derivatives range from 32.74 to 57.27, peaked at zero. This characteristic shows that FLOW360 has a resemblance to natural scenes [27].

Regarding the flow field analysis we directly compare the distribution of motion $u(x, y)$, speed defined as $s(x, y)=\sqrt{u(x, y)^2 + v(x, y)^2}$, flow direction $\Theta(x, y)=\tan^{-1}(v(x, y)/u(x, y))$ and spatial flow derivative of $u$ and $v$. The close resemblance of the flow field statistics between Sintel and FLOW360 suggests motion field resemblance with natural movies. Based on these comparisons, FLOW360 exhibits sufficient properties evident enough for its perceptual realism and complexities.

**Comparison with OmniFlow and OmniFlowNet.** OmniFlow [52] presents an omnidirectional flow dataset that is roughly similar to FLOW360. However, the major distinction between these datasets is the FOV. FLOW360 provides immersive 360° FOV, whereas OmniFlow provides only 180° FOV showing FLOW360 compared to OmniFlow is the true omnidirectional dataset. Similarly, OmniFlowNet [3] presents synthetic omnidirectional flow dataset with 360° FOV. However, this dataset contains low poly unnatural scenes, which can be explained by relatively larger kurtosis $(373.55, 391.09)$, characteristic of a power spectrum and luminance distribution (peaked at 255). The overall statistical analysis reveals FLOW360's better perceptual realism and diversity.

**Applications.** As we mentioned, the FLOW360 dataset contains frames and forward flow field and includes backward flow field, depth maps, and 3D-FLOW360 worlds, providing potential for applications like continuous flow-field estimation in 3 frames setting. Besides optical flow estimation, the FLOW360 dataset can be used in other applications such as depth and normal field estimation. Moreover, given 3D-FLOW360 animation data, the researcher can create as many optical flow datasets as needed.

## 4 SLOF

SLOF, as shown in Fig. 1, is inspired by the recent work on Siamese representation learning [18]. Since the method we rely on acts as a hub between several methods like contrastive learning, clustering, and siamese networks, it exhibits two special properties required for our case. First, this method has non-collapsing behavior. Here, the term collapsing refers to a situation where an optimizer finds possible minimum -1 similarity loss resulting degenerate solution (characterized by zero *std* of $l_2$-normalized output $z/||z||_2$ for each channel) while training without stop-gradient operations. Stop-gradient yields *std* value near $\frac{1}{\sqrt{d}}$ across each channel for all samples preventing such behaviour [18]. Sec-

ond, it is useful when we have only positive discriminative cases. SLOF does not consider radial distortion mitigation via changing/transforming the convolution layers rather learns the equivariant properties of 360 videos via siamese representation. We claim that such transformation is trivial, based on the following fact. First, the omnidirectional videos are projected in angular domain, *w.r.t.* **polar**($\theta$), **azimuthal**($\phi$); $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2}), \phi \in (-\pi, \pi)$, so we can learn flow fields in these domains and convert these flow fields to spherical domain using planar to spherical transformations as shown in Eq. 1 and Eq. 2. Second, the intent of a convolution operator in optical flow architecture is relatively different from other applications like classification, detection, or segmentation network, where other tasks require convolution to learn relevant features (spatially consistent), the relevance of these features should stay consistent (strictly for better performance) throughout any spatial location of the images/videos. However, the convolution operation is dedicated to computing the pixel-wise displacement regardless of spatial inconsistency in the distorted region via equivariant representation learning [18]. Another important consideration of such a design is to make this method portable to any existing optical flow architecture. This eliminates the architecture re-adjustments tasks and make it powerful and portable.

**Mapping Flow Field to Unit Sphere.** Input to our model are equirectangular images projected in angular domain **polar**($\theta$), **azimuthal**($\phi$), where these angles are defined in radian as $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2}), \phi \in (-\pi, \pi)$, thus the predicted optical flow is in $(\theta, \phi)$. These flow fields can be converted to unit sphere using planar to spherical co-ordinate transformation as shown below:

$$(x_s, y_s, z_s) = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta). \tag{1}$$

We can compute sphere to catadioptric plane [31] projections to express the flow field in Cartesian co-ordinates as:

$$(x, y) = (\frac{x_s}{1 - z_s}, \frac{y_s}{1 - z_s}) = (\cot\frac{\theta}{2}\cos\phi, \cot\frac{\theta}{2}\sin\phi). \tag{2}$$

**Design.** Given a pair of input image sequence $X_1 = (x_1, x_2)$, the rotation head ($R$) computes augmented view of this sequence as $X_2 = (x_1', x_2')$ with rotation $r$ using a random combination of "pitch", "yaw" and "roll" operations. These two augmented views are passed as an input to an encoder network $f$, defined as $f = P(R'(\Theta(E(R(X, r)))))$ where $E$ is a flow prediction module, RAFT [61] in our case, $\Theta$ is a mapping of 2D flow to unit sphere, $R'$ is a reverse rotation operation and $P$ is a convolution based down-sampling head. A prediction head presented as $h$ (an MLP head), transforms the output from the encoder $f$ from one stream to match the other stream. The illustration of this process shown in Equation. 3 as maximization of cosine similarity two views from siamese stream:

$$D(p^{left}, z^{right}) = -\frac{p^{left}}{||p^{left}||_2} \cdot \frac{z^{right}}{||z^{right}||_2}. \tag{3}$$

Here, $p^{left} \triangleq h(f^{left}(X_1))$ and $z^{right} \triangleq f^{right}(X_2)$ denotes the output vectors to match from two different streams($f^{left}, f^{right}$). This maximization problem can

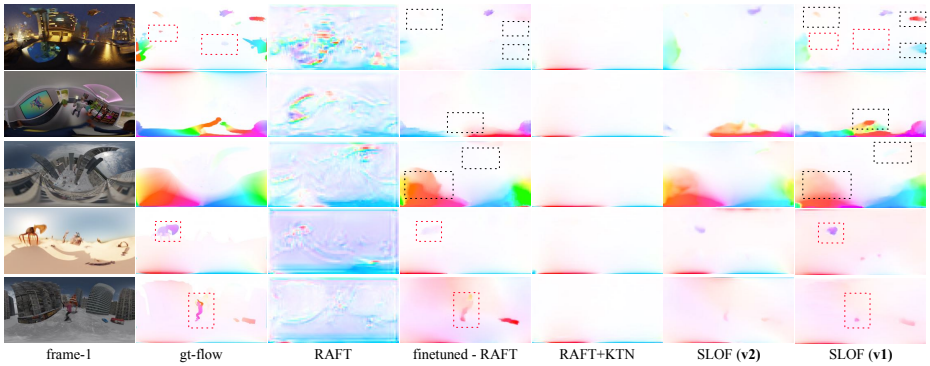



**Fig. 6. Qualitative results on FLOW360 test set**. Qualitative results show our best model SLOF(**v1**) shows better results compared to fine-tuned RAFT trained with policy explained in [61]. The dotted (**black**) rectangle indicates the comparative improvements of our model over fine-tuned RAFT. RAFT+KTN method fails to predict flow-field correctly; instead, it only predicts shallow flow fields from camera motion. The weakness of our model can be seen on dotted (red) rectangle where smaller motion segments are missing. **Note:** Flows information is clipped for better visualization.

be viewed from another direction, with $(p^{right}, z^{left})$ as the second matching pair from siamese stream $(f^{right}, f^{left})$ respectively. Given two matching pairs, we can use following (Eq. 4) symmetrized similarity loss function $L_{sim}$ (note that $z^{left}$ and $z^{right}$ are treated as a constant term using stop-grad operations to prevent a degenerate solution due to model collapse [18]). Similarly, the optical flow loss $L_{flow}$ is computed as a sequence loss [61] over predicted flow field and ground truth. This loss ($l_1$ distance over predicted and ground truth flow $f_{gt}$) is computed and averaged over sequence of predictions iteraterively generated for the same pair of input frames $\{f_1, f_2, ..., f_n\}$=$E(R(X, r))$ as shown in Equation. 4, where $\gamma$=$0.8^{n-i-1}$ served as weights over sequence loss. Note that $(n, i)$ denotes number of prediction($n$) in sequence and prediction id($i$) in predicted flow sequences. The design of the weighted schemes ensures different levels of confidence on predicted flows over time.

$$L_{sim} = \frac{1}{2}D(p^{left}, z^{right}) + \frac{1}{2}D(p^{right}, z^{left}),\ L_{flow} = \sum_{i=1}^{n} \gamma ||R(f_{gt}, r) - f_i||. \quad (4)$$

Given similarity loss($L_{sim}$) and flow loss($L_{flow}$) we implement a hybrid loss function $L$=$L_{sim}$+$L_{flow}$. The overall objective of this loss function is to maximize the similarity between latent representation of flow information while minimizing the loss between ground truth and predicted optical flow.

## 5 Experiments

We evaluate SLOF on the FLOW360 test set. We use pre-trained RAFT on Sintel [16] and fine-tune on FLOW360 as a comparison baseline. The fine-tuning

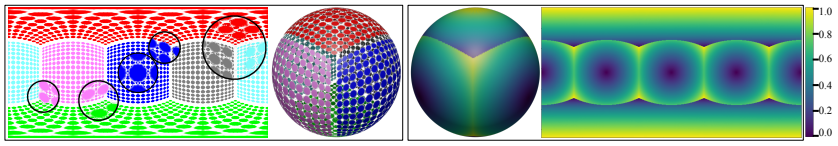



**Fig. 7. Distortion density map**. Illustrating different distortion intensity due to equirectangular projections. Left: upper (red) and lower (green) part of projections shows higher distortion in central part where as the equatorial region (cyan, pink, blue, gray) exhibit higher distortion rate away from the center of tangential plane. Right: shows the distortion density from $(0, 1)$. This distortion density map is used to evaluate the distortion aware EPE ($\text{EPE}_d$). **Note**: Each circle patch in left spherical projection have same area.

process is done using training protocols suggested in [61]. Moreover, to make a fair comparison with traditional methods, we transform RAFT (pre-trained) to adapt spherical convolution using KTN [58]. KTN transforms the convolution kernel to mitigate the radial distortions via estimating the spherical convolution function. Additionally, we run ablation studies on different training strategies and propose a distortion-aware evaluation. We will present details of the training procedure in the supplemental material.

**Scope.** The scope of our experiments are two folds: First, create a baseline for future researchers to explore novel methodologies. Second, address the validity of our method based on the fair comparison with a flow network designed for a spherical dataset. We formulate our baseline experiment on perspective optical flow network RAFT and modified version of RAFT with KTN [58] to compare the performance. The RAFT+KTN architecture simulates a domain adaptation similar to approaches like [9,3]. We choose KTN because of its success over alternative approaches like [19,24,69,20,57]. It is worth noting that the design of omnidirectional flow estimation can be extended to several techniques involving mitigation of radial distortions, making it practically impossible to cover all.

**Augmentation Strategy.** Given the nature of SLOF, we can train it using two different training strategies (**v1**,**v2**) as shown in Fig. 1(right). These strategies can be achieved by performing different rotational augmentation on the input sequences. The first strategy (**v1**) can be achieved by using set of inputs $(R(X_1, r_1), R(X_2, r_2))$ where $r_1{=}(0, 0, 0)$, i.e., $X_1$ does not have any rotational augmentation, whereas $r_2{\neq}(0, 0, 0)$ has rotation defined with random combinations of "pitch", "roll", and "yaw" operations. This setting is kept consistent throughout the training process. Alternatively, identical augmentation can be achieved by flipping this augmentation protocols. The second rotational scheme (**v2**) can be achieved by randomly switching rotation such that when $r_1$ is none, the $r_2$ is some random rotational augmentation and vice versa. This approach performs on par with **v1**.

$$\text{AE} = \arccos\left(\frac{u_e u_r + v_e v_r + 1}{\sqrt{u_r^2 + v_r^2 + 1}\sqrt{u_e^2 + v_e^2 + 1}}\right). \tag{5}$$

**Table 1. Quantitave results on FLOW360 test set.** $*$ denotes that we use $\text{EPE}_d$/$\text{AE}_d$ as the metrics; otherwise, the normal EPE and AE. Compared to baseline, SLOF achieves lower end-point-error and angular error on both distortion aware ($\text{EPE}_d$ and $\text{AE}_d$) and normal scheme. In terms of end-point-error (lower the better) our model (**v1**,**v2**) outperforms all the baseline. Similarly in terms of angular error (lower the better) our models (**v1**, **v2**) perform comparatively similar and outperform all the baseline. Though RAFT+KTN achieves comparable normal EPE, the distortion aware (Weighted) metrics ($\text{EPE}_d$ and $\text{AE}_d$) are significantly larger. **Note**: metrics in range (all, less than (5, 10, 20) and greater than 20) is computed as an average, based on the speed ($\mathbf{s}(x,y)=\sqrt{u(x,y)^2+v(x,y)^2}$) only in the respective pixel regions.

| Mehtod | Version | Metric | Weighted $\mathbf{s}\geq 0^*$ | $\mathbf{s}\geq 0$ | $\mathbf{s}<5$ | $\mathbf{s}<10$ | $\mathbf{s}<20$ | $\mathbf{s}\geq 20$ |
|---|---|---|---|---|---|---|---|---|
| Baselines | RAFT [61] | EPE | 3.344 | 2.058 | 0.558 | 0.682 | 0.838 | 71.736 |
| | | AE | 1.120 | 0.820 | 0.825 | 0.821 | 0.819 | 0.868 |
| | Finetuned RAFT [61] | EPE | 2.635 | 1.624 | 0.314 | 0.393 | 0.509 | 65.340 |
| | | AE | 0.745 | 0.522 | 0.527 | 0.522 | 0.520 | 0.647 |
| | RAFT + KTN [58] | EPE | 3.899 | 2.222 | 0.598 | 0.742 | 0.924 | 76.426 |
| | | AE | 2.020 | 0.912 | 0.912 | 0.910 | 0.911 | 1.0114 |
| SLOF | Switch rotation (**v2**) | EPE | 2.626 | 1.615 | 0.326 | 0.401 | 0.512 | 64.678 |
| | | AE | **0.691** | **0.485** | **0.489** | **0.484** | **0.482** | 0.659 |
| | Single rotation (**v1**) | EPE | **2.548** | **1.568** | **0.309** | **0.387** | **0.502** | 62.476 |
| | | AE | 0.708 | 0.497 | 0.501 | 0.497 | 0.495 | **0.607** |

**Evaluation Strategy.** We evaluate our method based on 2D-raw flow. Besides, using EPE (End Point Error in Eq. 6), i.e., Euclidean distance between the predicted flow and ground truth flow, as a single evaluation metric, we incorporate AE (Angular Error) as shown in Eq. 5 as the second measure. To explain the error in the omnidirectional setting, we introduce a distortion-aware measure called $\text{EPE}_d$ as in Eq. 6. This metric penalizes the error in the distorted area based on the distortion density map.

$$\text{EPE} = \frac{1}{N}\sum_{i}^{N} ||f_{pred} - f_{gt}||_2, \;\; \text{EPE}_d = \frac{1}{N}\sum_{i}^{N} \frac{||f_{pred} - f_{gt}||_2}{1-d}. \tag{6}$$

As $\text{EPE}_d$, $\text{AE}_d$ is calculated as $\frac{1}{N}\sum_{i}^{N}\frac{\text{AE}}{1-d}$ where, $d$ represents the distortion density map illustrated in Fig. 7, $f_{pred}=(u_e, v_e)$ represents predicted flow, and $f_{gt}=(u_r, v_r)$ represents ground truth flow. Note that, to maintain lower metrics scale the distortion density is mapped between $[0.500, 1.000)$ from $(0.0, 1.0]$. Please refer to supplemental for additional details on distortion density map.

**Results.** Fig. 6, Fig. 8 and Table 1 summarize our experimental results. The overall summary of qualitative results is presented in Fig. 6. SLOF performs better than baseline RAFT and kernel transformed RAFT+KTN methods. This result is evident enough to show that siamese representation learning can exploit the rotational properties of 360° videos to learn omnidirectional optical flow regardless of explicit architecture adjustments.

Our methods, SLOF (**v1**,**v2**) perform better than presented baselines. Among these methods **v1** has the best EPE score whereas, **v2** has better AE score. However, AE on both **v1** and **v2** are relatively similar, suggesting **v1** as our best method. This is clearly visible in qualitative results shown in Fig. 6.

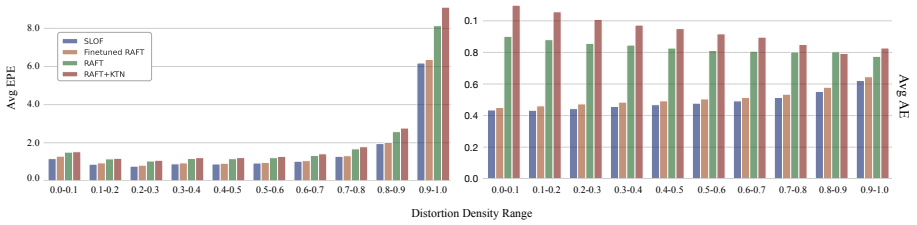



**Fig. 8. Error distribution plot**. Illustrating error (EPE and AE) in different distortion density ranges. SLOF relatively performs better in all distortion density ranges.

By investigating distortion-aware EPE, we can see that RAFT with KTN achieves significantly higher EPE regardless of comparable normal EPE with the other methods. This clearly explains why RAFT+KTN methods could not predict the motion around the distorted area; instead, it predicts shallow flow fields due to camera motion only. Moreover, comparing qualitative results in Fig. 6 and EPE measure in different distortion ranges in Fig. 8, we can see that our best method can predict smoother flow fields compared to baseline methods. These fields in the polar region are comparatively better and have better motion consistency in the edge region. However, our model might fail to predict relatively smaller motion regions in some cases, which leaves room for future improvements based on the proposed method. This concludes that RAFT+KTN requires additional re-engineering and domain adaptation, which is out of the scope of current work.

## 6 Conclusion

Omnidirectional flow estimation remains in its infancy because of the shortage of reliable benchmark datasets and tedious tasks dealing with inescapable radial distortions. This paper proposes the first perceptually natural-synthetic benchmark dataset, FLOW360, to close the gap, where comprehensive analysis shows excellent advantages over other datasets. Our dataset can be extended for other non-motion applications like segmentation and normal estimation task as well. Moreover, we introduce a siamese representation learning approach for omnidirectional flow (SLOF) instead of redesigning the convolution layer to adapt omnidirectional nature. Our method leverages the invariant rotation property of 360° videos to learn similar flow representation on various video augmentations. Meanwhile, we study the effect of different rotations on the final flow estimation, which provides a guideline for future work. Overall, the elimination of network redesigns aids researchers in exploiting existing architectures without significant modification leading faster deployment in real world setting.

**Acknowledgements.** This research was partially supported by NSF CNS-1908658, NeTS-2109982 and the gift donation from Cisco. This article solely reflects the opinions and conclusions of its authors and not the funding agents.

## References


1. Adobe: Mixamo, https://www.mixamo.com/ 5
2. Ahmadi, A., Patras, I.: Unsupervised convolutional neural networks for motion estimation. In: ICIP (2016) 4
3. Artizzu, C.O., Zhang, H., Allibert, G., Demonceaux, C.: Omniflownet: a perspective neural network adaptation for optical flow estimation in omnidirectional images. In: ICPR (2021) 2, 4, 8, 9, 12
4. Azevedo, R., Birkbeck, N., Simone, F., Janatra, I., Adsumilli, B., Frossard, P.: Visual distortions in 360-degree videos. TCSVT **2019**(8), 2524–2537 (2020) 2
5. Bailer, C., Taetz, B., Stricker, D.: Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation. In: ICCV (2015) 1
6. Baker, S., Roth, S., Scharstein, D., Black, M.J., Lewis, J., Szeliski, R.: A database and evaluation methodology for optical flow. In: ICCV (2007) 1, 3, 4, 6, 8
7. Barron, J.L., Fleet, D.J., Beauchemin, S.S.: Performance of optical flow techniques. IJCV **12**(1), 43–77 (1994) 4
8. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. In: ECCV (2016) 5
9. Bhandari, K., Zong, Z., Yan, Y.: Revisiting optical flow estimation in 360 videos. In: ICPR (2021) 2, 3, 4, 12
10. Blender: Https://www.blender.org/ 5
11. Boomsma, W., Frellsen, J.: Spherical convolutions and their application in molecular modelling. In: NeurIPS (2017) 3
12. Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. IJPRAI **7**(04), 669–688 (1993) 5
13. Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. IJPRAI **7**(04), 669–688 (1993) 5
14. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: ECCV (2004) 1
15. Brox, T., Malik, J.: Large displacement optical flow: descriptor matching in variational motion estimation. TPAMI **33**(3), 500–513 (2010) 1
16. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: ECCV (2012) 1, 3, 4, 8, 11
17. Chen, Q., Koltun, V.: Full flow: Optical flow estimation by global optimization over regular grids. In: CVPR (2016) 1
18. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR (2021) 5, 9, 10, 11
19. Cohen, T.S., Geiger, M., Koehler, J., Welling, M.: Spherical cnns. arXiv (2018) 3, 12
20. Coors, B., Condurache, A.P., Geiger, A.: Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In: ECCV (2018) 3, 12
21. Demonceaux, C., Kachi-Akkouche, D.: Optical flow estimation in omnidirectional images using wavelet approach. In: CVPRW (2003) 4
22. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: ICCV (2015) 3, 4

23. Eder, M., Shvets, M., Lim, J., Frahm, J.M.: Tangent images for mitigating spherical distortion. In: CVPR (2020) 2, 3
24. Esteves, C., Allen-Blanchette, C., Makadia, A., Daniilidis, K.: Learning so (3) equivariant representations with spherical cnns. In: ECCV (2018) 3, 12
25. Feng, B.Y., Yao, W., Liu, Z., Varshney, A.: Deep depth estimation on 360° images with a double quaternion loss. In: 3DV (2020) 3
26. Fernandez-Labrador, C., Facil, J.M., Perez-Yus, A., Demonceaux, C., Civera, J., Guerrero, J.J.: Corners for layout: End-to-end layout recovery from 360 images. RA-L **5**(2), 1255–1262 (2020) 5
27. Field, D.J.: Relations between the statistics of natural images and the response properties of cortical cells. Josa a **4**(12), 2379–2394 (1987) 4, 9
28. Garg, R., Roussos, A., Agapito, L.: A variational approach to video registration with subspace constraints. IJCV **104**(3), 286–314 (2013) 1
29. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. IJRR **32**(11), 1231–1237 (2013) 1, 3, 4
30. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR (2012) 1, 4
31. Geyer, C., Daniilidis, K.: A unifying theory for central panoramic systems and practical implications. In: ECCV (2000) 10
32. Goralczyk, A.: Nishita sky demo (2020), creative Commons CC0 (Public Domain) - Blender Studio - cloud.blender.org 5
33. Horn, B.K., Schunck, B.G.: Determining optical flow. AI **17**(1-3), 185–203 (1981) 1
34. Horn, B., Schunck, B.: Techniques and applications of image understanding (1981) 1
35. Hui, T.W., Tang, X., Loy, C.C.: Liteflownet: A lightweight convolutional neural network for optical flow estimation. In: CVPR (2018) 1, 3, 4
36. Hui, T.W., Tang, X., Loy, C.C.: A lightweight optical flow cnn —revisiting data fidelity and regularization. TPAMI **43**(8), 2555–2569 (2021) 5
37. Hulle, S.V.: Bcon19 (2019), 2019 Blender Conference - cloud.blender.org 5
38. Hur, J., Roth, S.: Mirrorflow: Exploiting symmetries in joint optical flow and occlusion estimation. In: ICCV (2017) 3
39. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: CVPR (2017) 1, 3, 4
40. Jason, J.Y., Harley, A.W., Derpanis, K.G.: Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In: ECCV (2016) 4
41. Jiang, S., Campbell, D., Lu, Y., Li, H., Hartley, R.: Learning to estimate hidden motions with global motion aggregation. arXiv (2021) 1
42. Liu, C., Freeman, W.T., Adelson, E.H., Weiss, Y.: Human-assisted motion annotation. In: CVPR (2008) 4
43. Liu, P., Lyu, M., King, I., Xu, J.: Selflow: Self-supervised learning of optical flow. In: CVPR (2019) 3
44. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: IJCAI. vol. 2 (1981) 1
45. McCane, B., Novins, K., Crannitch, D., Galvin, B.: On benchmarking optical flow. CVIU **84**(1) (2001) 4
46. Meister, S., Hur, J., Roth, S.: Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In: AAAI (2018) 3

47. Meister, S., Jähne, B., Kondermann, D.: Outdoor stereo camera system for the generation of real-world benchmark data sets. Optical Engineering **51**(2), 021107 (2012) 4
48. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: CVPR (2015) 1, 4
49. Menze, M., Heipke, C., Geiger, A.: Discrete optimization for optical flow. In: GCPR (2015) 1
50. Otte, M., Nagel, H.H.: Optical flow estimation: advances and comparisons. In: ECCV (1994) 4
51. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: CVPR (2017) 4
52. Seidel, R., Apitzsch, A., Hirtz, G.: Omniflow: Human omnidirectional optical flow. In: CVPR (2021) 2, 4, 8, 9
53. Shakernia, O., Vidal, R., Sastry, S.: Omnidirectional egomotion estimation from back-projection flow. In: CVPRW (2003) 4
54. Simoncelli, E.P., Olshausen, B.A.: Natural image statistics and neural representation. Annual review of neuroscience **24**(1), 1193–1216 (2001) 9
55. Sketchfab: Https://sketchfab.com/ 5
56. Steinbrücker, F., Pock, T., Cremers, D.: Large displacement optical flow computation withoutwarping. In: ICCV (2009) 1
57. Su, Y.C., Grauman, K.: Learning spherical convolution for fast features from 360°imagery. In: NeurIPS (2017) 3, 12
58. Su, Y.C., Grauman, K.: Kernel transformer networks for compact spherical convolution. In: CVPR (2019) 2, 3, 4, 12, 13
59. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Models matter, so does training: An empirical study of cnns for optical flow estimation. TPAMI **42**(6), 1408–1423 (2019) 1
60. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: CVPR (2014) 5
61. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV (2020) 1, 3, 4, 10, 11, 12, 13
62. Teney, D., Hebert, M.: Learning to extract motion from videos in convolutional neural networks. In: ACCV (2016) 4
63. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Deep end2end voxel2voxel prediction. In: CVPRW (2016) 4
64. Turbosquid: Https://www.turbosquid.com 5
65. Wang, R., Geraghty, D., Matzen, K., Szeliski, R., Frahm, J.M.: Vplnet: Deep single view normal estimation with vanishing points and lines. In: CVPR (2020) 3
66. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: Deepflow: Large displacement optical flow with deep matching. In: ICCV (2013) 1
67. Woliński, M.: City - 3d model, sketchfab.com 5
68. Wulff, J., Black, M.J.: Efficient sparse-to-dense optical flow estimation using a learned basis and layers. In: CVPR (2015) 4
69. Zhang, Z., Xu, Y., Yu, J., Gao, S.: Saliency detection in 360 videos. In: ECCV (2018) 12
70. Zhao, S., Sheng, Y., Dong, Y., Chang, E.I., Xu, Y., et al.: Maskflownet: Asymmetric feature matching with learnable occlusion mask. In: CVPR (2020) 1, 3
71. Zioulis, N., Karakottas, A., Zarpalas, D., Daras, P.: Omnidepth: Dense depth estimation for indoors spherical panoramas. In: ECCV (2018) 3