Towards Equivariant Optical Flow Estimation with Deep Learning

Stefano Savian[1,2], Pietro Morerio[1], Alessio Del Bue[1], Andrea A. Janes[2], and Tammam Tillo[3]

{*stefano.savian, pietro.morerio, alessio.delbue*}*@iit.it, ajanes@unibz.it, tammam@iiitd.ac.in*
[1]Pattern Analysis & Computer Vision (PAVIS), Istituto Italiano di Tecnologia, Genova, Italy
[2]Free University of Bozen-Bolzano, Bolzano, Italy
[3]Indraprastha Institute of Information Technology Delhi (IIITD), Delhi, India

Abstract

Methods for Optical Flow (OF) estimation based on Deep Learning have considerably improved traditional approaches in challenging and realistic conditions. However, data-driven approaches can inherently be biased, leading to unexpected under-performance in real application scenarios. In this paper, we first observe that the OF estimation accuracy varies with motion direction, and name this phenomenon 'OF sign imbalance'. The sign imbalance cannot be assessed by means of the endpoint-error (EPE), the typical training and evaluation metric for Deep Optical Flow estimators. This paper tackles this issue by proposing a new metric to assess the sign imbalance, which is compared to the endpoint-error. We provide an extensive evaluation of the sign imbalance for the state-of-the-art optical flow estimators. Based on the evaluation, we propose two strategies to mitigate the phenomenon, i) by constraining the model estimations during inference, and, ii) by constraining the loss function during training. Testing and training code is available at: `www.github.com/stsavian/equivariant_of_estimation`.

1. Introduction

Optical flow estimation is an essential task for a wide range of real-world applications spanning across many areas of computer vision. Indeed, motion detection, object tracking, video compression [26], video interpolation [1], deblurring [38], structure-from-motion [36], simultaneous localization and mapping (SLAM) [39], surveillance [24], medical imaging [40], to mention a few, explicitly make use of optical flow. The OF is also used implicitly, to ensure that the models properly learn frame by frame relationships [41].

The most recent works on OF estimation are based on Deep Learning [35, 11, 33, 15]. These frameworks improved the OF estimation in typically challenging conditions, e.g. large displacements, non-rigid motion and illumination changes. However, the early works were still not performing on par with the more traditional variational and iterative methods [28]. More recent Deep Learning methods bridge this gap by embedding well-established traditional principles for OF estimation [35, 11], and, by identifying and mitigating different training bias of data-driven models [4].

This paper is motivated by the observation that there exists a common bias to all SOTA OF estimators, leading to substantially different estimates depending on motion direction. We name this phenomenon *sign imbalance bias*. The sign imbalance bias manifest itself when the same quantity of groundtruth motion is estimated differently depending on the direction of motion. Figure 1 shows the relevance of this problem. We experienced the sign imbalance bias on a video coding application domain, where we observed a significantly different signal-to-noise-ratio of the warped image with the OF, varying on motion direction. We believe the detection and mitigation of the sign imbalance to be very important for many different application scenarios, including the automotive domain, where a non-deterministic OF estimation could mean a safety hazard. The sign imbalance is detected by applying composable transformations (reflections) to the data and for this reason can be seen as a form of equivariance lack [18]. The term *equivariance* refers to the capability of the model to handle certain transformations of the data. With a little abuse of notation, assuming that the model can be represented by a function ϕ acting on an input x, a model is equivariant under a transformation T, if $\phi(T(\mathbf{x})) = T(\phi(\mathbf{x}))$, [18, 7, 17]. We also note that the commonly used training and testing metric, the endpoint-error (EPE), by construction cannot properly detect the sign imbalance. We will show that the sign imbalance is completely undetected, but accounts on average for about 50 %

We would like to thank Davide Talon and Matteo Bustreo for proofreading this paper.

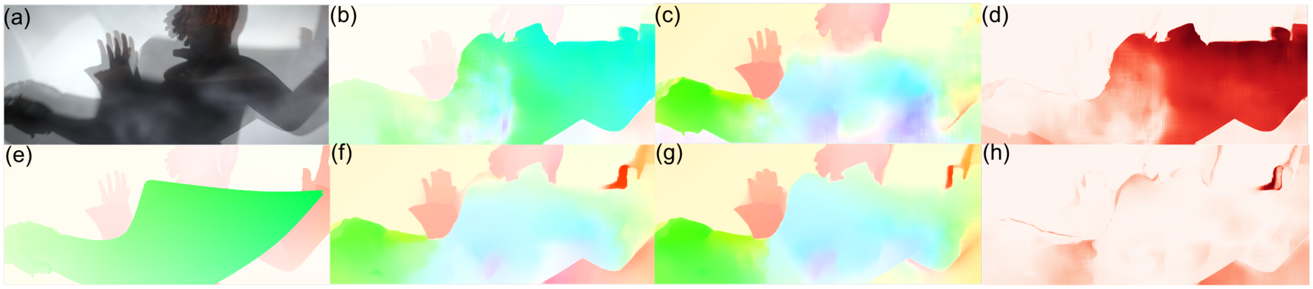


Figure 1: Sintel dataset [6] example. (a) Inputs F_n, F_{n+1} overlay, (e) groundtruth OF (Middlebury Visualization [3]). (b,f) RAFT [35] estimations from F_n, F_{n+1}, (c,g) RAFT estimations from the 180° rotated inputs $T_{180^\circ}(F_n), T_{180^\circ}(F_{n+1})$. The estimations are then rotated accordingly to allow for a comparison with the original (non rotated) estimations, Sec. 3.2. (b,c) are obtained by RAFT as originally trained by the authors (on FlyingThings3D [22]) (d) is the sign imbalance heatmap. (f,g) use our proposed loss function to limit the sign imbalance, and (h) corresponding sign imbalance heatmap. (f,g) are visually much closer than (b,c). This can be easily observed in (d,h).

of the EPE magnitude and can also reach magnitudes higher than the EPE, when trained on certain target data. For this reason this paper proposes an unsupervised methodology and metric to assess such phenomenon. This approach has been used to test the SOTA OF estimation models and their core components, showing that most of the SOTA OF estimators can considerably display sign imbalance.

Thus, this paper helps answering in detail the following questions: (RQ1) To which extent do the SOTA optical flow estimators quantitatively display sign imbalance? (RQ2) What are the main causing factors? (RQ3) Can such bias be mitigated, and how? We firstly show that the SOTA deep learning OF methods present a severe degree of sign imbalance. This bias cannot be solved by augmenting the training data with reflections, or by training using forward and backward OF groundtruth data. To solve this issue, we propose and extensively evaluate a novel loss function acting as a damping mechanism to mitigate the sign imbalance, and an ensemble technique to completely reduce the phenomenon during inference. The proposed loss function is based on a new metric and methodology designed to integrate the EPE limitations. Results show a considerable sign imbalance reduction (more than 50% on average, around 5 times when fine tuning on largely unbalanced data) when applying the loss function to the top performing OF estimator. The paper structure is as follow. Section 2 presents the related work, Sec. 3, proposes the evaluation and mitigation methodologies. Section 4, shows the experimental results. Sec. 5 discusses the results and concludes this paper. Testing and training scripts will be available online.

2. Related Works

This paper improves the OF estimation quality by assessing and reducing the sign imbalance, a special case of lack of equivariance of data-driven models. Thus, this section addresses the related works as follows: SOTA models for OF estimation, OF benchmarking, equivariance assessment and mitigation for Deep Learning models.

State-of-the-art OF estimators. Traditionally, variational optical flow methods estimate the OF by minimizing an energy functional with an additional regularization term [10]. Recently, a substantially different paradigm was introduced by FlowNet [8]. FlowNet is the first Deep Learning model trained end-to-end to estimate the OF. FlowNet could optionally embed a correlation layer (FlowNetC) to produce a cost volume between features of the input image pair. An important contribution of FlowNet is FlyingChairs, a large scale computer rendered dataset with OF groundtruth, composed of planar motion, used for training. Another important dataset is FlyingThings3D [22] which is composed of rigid objects moving of 3D motion. FlyingChairs and FlyingThings3D are very important datasets used by almost all learnt models for OF estimation. PWC-Net is a lightweight model improving FlowNet by introducing a pyramidal refinement of the optical flow, and the warping of the feature maps to build a cost volume. IRR [11] is a generally applicable fixed resolution iterative refinement which improves the estimates quality by iteratively improving the OF produced by FlowNet or PWC-Net [11]. RAFT [35] considerably improved the accuracy over the previous existing OF models. RAFT relies on a classically inspired iterative residual refinement of the estimated OF updated by a recurrent unit performing lookups on 4D correlation volumes. Different works use RAFT as backbone: The global Motion Aggregation module [14], Deep Equilibrium Networks [2], CRAFT [31]. Very recently, GMFlow [37] proposed a Transformer network feature enhancement, the correlation for global feature matching, and a self-attention layer for flow propagation. Their formulation allows to reduce the inference time, if compared to RAFT. Substantial steps forward have also been achieved by improving the training data, training schedule and data augmentation. Two notable works analyzing the characteristics of the training data are

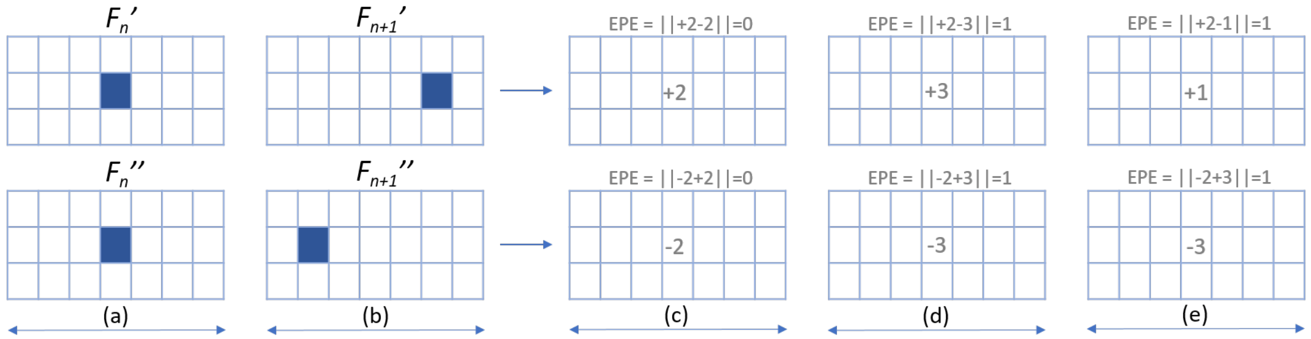


Figure 2: Two scenarios of motion of a pixel in two consecutive frames are shown in (a) and (b), producing a horizontal displacement of $+2$ and -2 respectively. Estimates of the horizontal displacement given by three hypothetical OF methods are shown in (c) - correct estimate, (d) - wrong but unbiased estimate, and (e) - wrong and biased estimate. As it can be noted, the endpoint errror (EPE) cannot account for the bias.

Mayer *et al.* [23] and Sun *et al.* [34]. Bar-Haim and Wolf, proposed ScopeFlow [4], showing considerable improvements by modifying the data sampling process by changing the data augmentation zoom and cropping size. Finally Sun *et al* show the importance of the training details on the models generalization, by retraining PWC-Net, IRR-PWC, and RAFT with the same schedule [32].

Unsupervised methods estimate the OF without the need of labelled groundtruth. One important SOTA unsupervised model is Uflow [15] which analyses the key components for unsupervised OF estimation, and embeds the best components into their proposed model. The currently top performing unsupervised method is SMURF [30], which uses RAFT as backbone and relies on multi frame inputs. Other SOTA unsupervised OF estimators are ARFlow [20] using data transformations as indirect-supervision, and DDFlow [21] using data distillation.

Optical flow benchmarking. The seminal work for OF benchmarking is the “Yosemite” synthetic sequence [5]. The most recent benchmarks are KITTI 2012 [9], KITTI 2015 [25], and HD1K [16], focusing on the automotive domain. Sintel [6] is a very challenging computer rendered dataset derived from the open source 3D animated short film *Sintel*. While being known benchmarks in the area of OF, none of the benchmarks described above have been designed to test network equivariance.

Equivariance in Deep Learning. A common strategy to improve equivariance is to train a neural network with data augmentation [18, 29, 19]. A different work by Lenc and Vedaldi [18], studies the equivariance and equivalence of Convolutional Neural Networks (CNN) feature representations. A more theoretical work by Kondor and Trivedi [17] proves that a convolutional structure is a necessary condition for assuring equivariance to the action of a compact group. The closest work to this paper is Jeong *et al* [13] which focuses on improving the OF consistency in presence of occlusions, and additionally proposes a transformation consistency loss. Instead, our work is rooted in the investigation of the sign imbalance and targets solutions to solve the bias. Finally, Savian *et al*, [27] originally observed the sign imbalance bias. Their work is largely limited by the usage of the EPE and do not propose solutions to the problem. Our work bridges this gap by providing a novel metric, a comprehensive evaluation of the sign imbalance and a number of strategies to mitigate it. This also give valuable insights on the limitations of learnt models for OF estimation and how to overcome them.

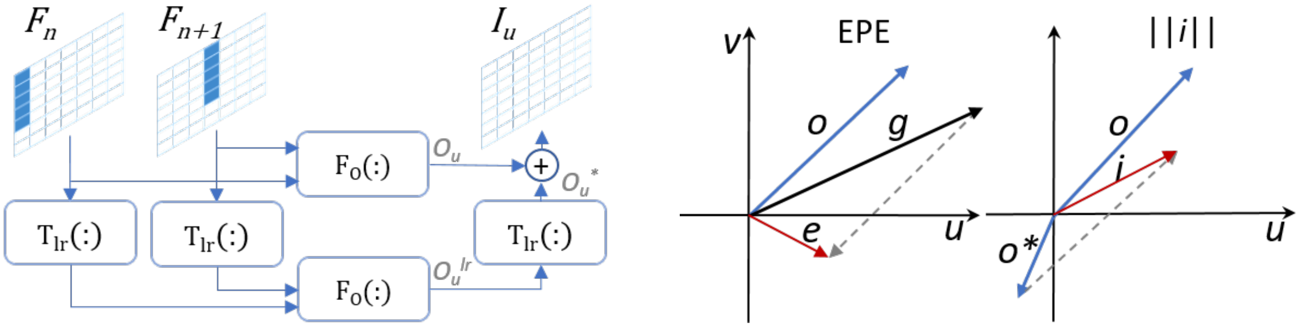
3. Methodology

We firstly show why the standard metric employed in OF learning, the EPE, is limited for detecting the sign imbalance bias – Sec. 3.1. Then, we present our proposed methodology and metric to assess the sign imbalance – Sec. 3.2, and our strategy to constrain the loss function to mitigate the sign imbalance – Sec. 3.3 and Sec. 3.4.

3.1. Why the EndPoint-Error (EPE) cannot detect the sign imbalance bias?

Suppose that $F_n^{'}$ and $F_n^{''}$ in Fig. 2 are two identical white 3×7 pixel frames with a dark pixel in the center. Let the single and double quotes superscript represent two scenarios of motion, namely, Scenario I and Scenario II, respectively.

In Scenario I the dark pixel moves to the right by 2 pixels, whereas, in Scenario II the dark pixel moves to the left by the same amount, as in Fig. 2 (b). This figure shows also in (c), (d), and (e) three hypothetical OF estimates. Fig. 2 (c) shows a correct estimation; (d) shows an inaccurate estimation, where in both Scenario I and II the motion was overestimated by the same amount; (e) shows an inaccurate estimation, where in Scenario I the amount of motion was overestimated, while in Scenario II it was underestimated by the same amount. An optical flow estimator showing a



(a)

(b)

Figure 3: (a) block diagram to produce the horizontal sign imbalance matrix I_u, (b) pixelwise cartesian representation of the EPE and of the Euclidean sign imbalance $||i||$.

behavior like the one shown in Fig. 2 (e), i.e. an estimator behaving differently given specific motion directions, is an *imbalanced estimator*. Differently, if the behavior of the optical flow estimator is like in Fig. 2 (c) or (d), the estimator is a *balanced estimator*. Assuming an OF estimate o and its groundtruth OF g for a single pixel, where the pedix u, v represent respectively the horizontal and vertical directions (throughout this paper), the EPE is formally defined as the Euclidean distance between the estimate and groundtruth; in other words, $EPE = \sqrt{(g_u - o_u)^2 + (g_v - o_v)^2}$. It is worth noticing that the EPE, would produce the same values for the dark pixel of the two scenarios represented in Fig. 2 in (d,e). Thus, such metric would not account for the two different cases and the related sign imbalance. Therefore, to evaluate imbalanced behavior of an OF estimator, a metric and a different methodology is necessary.

3.2. A metric for assessing the sign imbalance bias

Figure 2 is limited to a single pixel moving. Figure 3 (a) depicts our methodology to generalize this case to more pixels. In this figure, the dark object in frame F_n moves right of three pixels in F_{n+1}. Assuming only horizontal motion, scenario I is generated by estimating the OF starting from F_n, F_{n+1}, and scenario II is obtained by applying a horizontal reflection to the inputs for estimating the OF:

$$\begin{aligned} &F_n^{'} = F_n, \quad F_{n+1}^{'} = F_{n+1}; \\ &F_n^{''} = T_{lr}(F_n), \quad F_{n+1}^{''} = T_{lr}(F_{n+1}); \end{aligned} \tag{1}$$

where T_{lr} indicates a horizontal reflection of the data. The estimation $O = f_O(F_n, F_{n+1})$ defines a flow field $O = O_u, O_v$. Let $O^{lr} = f_O(T_{lr}(F_n), T_{lr}(F_{n+1}))$. Let O_u and O_u^{lr} be two matrices representing the estimated horizontal displacement of the pixels for Scenario I and Scenario II, respectively. The relationship between O_u, and O_u^{lr}, for an unbiased estimator, is: $O_u^{lr} = -T_{lr}(O_u)$, (vertical extension later). At this point Eq. (1) could be generalized to compute the sign imbalance for the estimated horizontal displacements, as:

$$I_u = O_u + T_{lr}(O_u^{lr}), \tag{2}$$

This matrix contains the imbalance of the estimated horizontal displacement for each pixel of F_n and F_{n+1}. The same reasoning can be applied to the vertical component, $F_n^{''} = T_{ud}(F_n)$, and $F_{n+1}^{''} = T_{ud}(F_{n+1})$, where T_{ud} indicates a vertical reflection of the data. Leading to $I_v = O_v + T_{ud}(O_v^{ud})$. In the remainder of this paper, let F_n, F_{n+1} be a generic frame pair. Thus, the previous equation and eq. (2) could be rewritten explicitly as:

$$I_u = f_{Ou}(F_n, F_{n+1}) + T_{lr}(f_{Ou}(T_{lr}(F_n), T_{lr}(F_{n+1}))) \tag{3}$$

$$I_v = f_{Ov}(F_n, F_{n+1}) + T_{ud}(f_{Ou}(T_{ud}(F_n), T_{ud}(F_{n+1}))) \tag{4}$$

Eq. (3),(4) show that it is possible to evaluate the extent of imbalance for the estimated displacement by using two consecutive frames, namely F_n, and F_{n+1}. A block diagram representing the previous equation to estimate $I_u(.)$, is shown in Fig. 3 (a); the extension to the vertical case is trivial. From a practical point of view, this process requires to perform two estimations of the optical flow. Thus, to evaluate the extent of the vertical and horizontal imbalance behavior of an optic flow estimator three estimates are needed.

We note that one 180° rotation equals to the composition of vertical and horizontal reflections, meaning that the sign imbalance matrix $I = I_u, I_v$, could be produced with two forward propagations, as in:

$$\begin{aligned} I = f_O(F_n, F_{n+1}) + T_{180^\circ}(f_O(T_{180^\circ}(F_n), T_{180^\circ}(F_{n+1}))) = \\ = O + T_{180^\circ}(O^{180^\circ}) = O + O^*. \end{aligned} \tag{5}$$

Equation (5) shows that this procedure produces a matrix I, composed by I_u, I_v. Every element $i \in I$ will have two entries i_u, i_v representing the sign imbalance. These values can be reduced to a layerwise mean $\overline{I_u}, \overline{I_v}$, (the overline indicates the arithmetic average). We average the pixelwise sign imbalance values using the Euclidean norm, this allows for a direct comparison of the sign imbalance over the groundtruth magnitude, or with the EPE. Figure 3 (b), shows how the error e is computed given the groundtruth vector $g \in G$, and the output vector $o \in O$ for a single pixel coordinate. The EPE is the L2-norm of the error e, i.e. $||e||$. Similarly, Fig. 3 (b) shows how the Euclidean imbalance $||i||$ is calculated, starting from the output $o \in O$, and the transformed output $o^* \in O^*$, $(O^* = O_u^*, O_v^*)$. The sign imbalance is equal to $i = o + o^*$, and the Euclidean sign imbalance is the L2 norm of i, i.e. $||i||$. Assuming $\mathcal{P}$ is the set of all pixels p considered, the average Euclidean sign imbalance can be evaluated with the arithmetic average and is noted as $\overline{||I||}$:

$$\overline{||I||} = \frac{1}{\mathcal{P}} \sum_{\forall p \in \mathcal{P}} ||i(p)||. \tag{6}$$

This metric has the advantage of directly allowing a comparison between $||i||$, and the EPE.

3.3. Sign imbalance mitigation - ensemble inference

A straightforward approach to limiting the sign imbalance bias is to average the estimates during inference [18]. We start with the following observation: if we average the vectors o and $-o^*$ at inference time, $o_m = \frac{o-o^*}{2}$, then the sign imbalance is zero by definition. O^* is here obtained following the block diagram in Fig. 3a, but using the T_{180° transform. Moreover, the averaged EPE for the error should be lower. This can be better explained with Fig. 4. Lets suppose that we have a groundtruth displacement g, and two estimates o_1, o_2, producing respectively the errors e_1, e_2. The estimates o_1, o_2 can be written as:

$$o_1 = g + e_1; \qquad o_2 = g + e_2. \tag{7}$$

If we take the average of o_1, o_2, we get:

$$o_m = \frac{o_1 + o_2}{2} = \frac{g + e_1 + g + e_2}{2} = g + \frac{e_1 + e_2}{2}. \tag{8}$$

Equation, (8) shows that if we perform different estimates using an optical flow model on the same, but transformed inputs, and average them, the estimation needs to present an lower error. Thus, assuming to any $||e_1||, ||e_2||$ then the averaged error is bounded by:

$$||\frac{e_1 + e_2}{2}|| \leq max(||e_1||, ||e_2||). \tag{9}$$

If $||e_1|| = ||e_2||$, then: $||\frac{e_1+e_2}{2}|| \leq ||e_1||$. Meaning that if the two errors present the same EPE, the averaged error must present a lower EPE, Fig. 4 b).

Thus, if this strategy is applied without any assumption on $||e_1||$, $||e_2||$, then the averaged error should be lower than the maximum EPE. Thus, by using this strategy we could jointly completely reduce the sign imbalance and improve the EPE, at the cost of a doubled inference time. The ensemble inference can also be used during training.

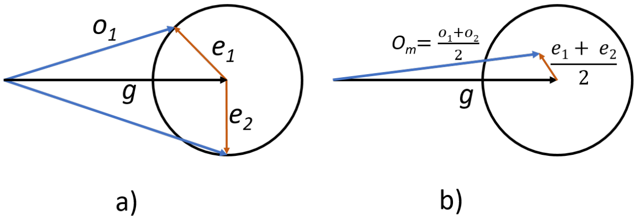


Figure 4: a) example of estimates averaged during inference. The groundtruth g is estimated by o_1 and o_2, producing the errors e_1, e_2, which in this case produce the same EPE, as the two estimates lay in the same circumference. b) by taking the average of the estimates o_1 and o_2, the error is also averaged, and it must be lower than e_1 and e_2.

3.4. Sign imbalance mitigation - training loss function

Using the ensemble inference as described in Sec. 3.3, results in a doubled inference time, and does not give insights on the origin of the sign imbalance bias. In this section we investigate how to integrate any loss function with an additional loss, exploiting the developed framework and metric, to constrain the models to produce balanced estimates. The auxiliary loss should act as a damping mechanism to mitigate unbalanced outputs: high values of sign imbalance should be penalized by an imbalance loss increase. However, the sign imbalance loss alone could lead to inaccurate balanced estimates. For this reason the EPE loss and sign imbalance loss should be weighed accordingly, $\mathcal{L} = \mathcal{L}_E + \beta\mathcal{L}_I$, where $\mathcal{L}$ is the total loss, $\mathcal{L}_E$ is the EPE loss, $\mathcal{L}_I$ is the sign imbalance loss, and β is a weighing hyperparameter. The sign imbalance loss $\mathcal{L}_I$ is calculated starting from I, obtained by taking the norm of eq. (5), as: $\mathcal{L}_I = |O + O^*|$. (the choice of L1-Norm or L2-Norm depends on the EPE training norm, Supplementary Sec. 3)

This process requires to obtain O^* as defined in eq. (5), using the methodology depicted in Fig. 3 (a) (extended to the T_{180° transformation). To this extent, we propose Multiple Forward Propagations (FWD suffix) of the input frames and the transformed frames, $(F_n, F_{n+1}, F_n^{180^\circ}, F_{n+1}^{180^\circ})$, to sequentially obtain O, O^*, for then computing the losses $\mathcal{L}_E, \mathcal{L}_I$ and finally populate the gradient by backpropagation.

Algorithm 1, summarizes the approach for every training iteration. F_n, F_{n+1} are forward propagated to obtain O. After that, $F_n^{180^\circ}, F_{n+1}^{180^\circ}$, are computed and forward propagated into the same network to obtain O^{180°, and then 180°rotated again to obtain O^*. The sign imbalance and EPE loss are then computed, and the gradients are popu-

Algorithm 1 Multiple forward propagations

```
1: procedure TRAINING-ITERATION($F_n, F_{n+1}, G$)
2:     $O$ = inference($F_n, F_{n+1}$)
3:     if FWDs $\implies$ detach gradients.
4:     Generate $F_n^{180^\circ}, F_{n+1}^{180^\circ}$
5:     $O^{180^\circ}$ = inference($F_n^{180^\circ}, F_{n+1}^{180^\circ}$)
6:     $O^*$ = $T_{180^\circ}(O^{180^\circ})$
7:     Compute $\mathcal{L}_E(O,G)$
8:     Compute $\mathcal{L}_I(O,O^*)$
9:     $\mathcal{L}$= $\mathcal{L}_E$+ $\beta\cdot$ $\mathcal{L}_I$
10:    if FWDg $\implies$ $\mathcal{L}$= $\mathcal{L}/2$
11:    Compute gradients
12:    Update learnable weights
```

lated by back-propagation. The double forward propagation does not require the network input size to be changed and can be applied to virtually any learnt OF estimator. However, the double forward propagation before the backward pass changes how the dynamic graph to calculate the gradient is computed. Stopping the gradient (FWDs), during the second inference, set its computational graph differentials to zero, (the symbol $\implies$ in line 3 and 10 means "then"). When letting the gradient flow, (FWDg), the gradients during the second forward propagation are computed and accumulated. This also requires to halve the total loss $\mathcal{L}$ (line 9 in Alg. 1) to maintain the same gradient magnitude (supplementary Sec. 3). From a practical perspective, FWDg strategies double the memory footprint during training, whereas FWDs strategies do not add memory overhead. Finally, the loss function strategy can be combined with the ensemble inference (Sec. 3.3) during training, to minimize the averaged EPE.

4. Experiments

The experiments in this section benchmark the sign imbalance for different OF estimators and their core components. Sec. 4.1 describes the testing datasets, the models, their training scheme and the terminology used in the experiments. Section 4.2 presents the sign imbalance evaluation results, while Sec. 4.3 shows how the sign imbalance can be mitigated with our approach applied to the top performing method for OF estimation.

Table 1: List of acronyms used.

Acronym	Data	Acronym	Meaning
C	FlyingChairs	DF	DDFlow
Co	FlyingChairsOcc	G	GMA
C2	FlyingChairs2	IP	IRR-PWC
C2f	FlyingChairs2 (FWD)	R	RAFT
T	FlyingThings	'	Retrained
Tf	FlyingThings (FWD)	o	original probability
S	Sintel		
K	KITTI	-M	50 % probability

4.1. Experimental setting

Testing datasets. The Sintel [6] *training* subset is used as our main testing benchmark. The models have been additionaly tested on Monkaa [22], and KITTI [25]. Results on Sintel and Monkaa are comparable (tested with RAFT - Supplementary Sec. 11 -). Detailed dataset statistics can be found in the supplementary material, Sec. 6.

Models and training datasets. The SOTA networks considered are: RAFT (R) [35], GMA (G) [14], and IRR-PWC (IP) [11], the current leading OF estimation methods based on a single resolution iterative refinement. We also evaluate DDFlow [21], (unsupervised method) based on PWC-Net. IRR and DDFlow both include bi-directional flow estimation to improve the estimates consistency. The acronyms used can be found in table 1.

Training schedule. All networks follow the common procedure of pretraining on FlyingChairs (C) [8], (or similarly FlyingChairs2 (C2) [12], or FlyingChairsOcc (Co) [11] which also provide backward OF); then train on FlyingThings3D (T) [22]. Eventually some models are then fine-tuned on the target dataset. We use the acronyms in table 1 and only report the last training dataset.

Experimental protocol. We perform ablation studies on RAFT to evaluate the effects of: training mirroring data augmentation, forward and backward OF groundtruth. Moreover, we test the networks with our ensemble inference strategy, and retrain RAFT using our developed sign imbalance loss function. Models marked with a prime symbol (') have been retrained by us.

4.2. Benchmarking the sign imbalance

Sign imbalance dependence on training data. For all models in table 2, $\overline{||I||}$ decreases with more training data, except when fine tuning on KITTI. However, the training data distribution has a limited impact on the sign imbalance (despite the extreme cases), as the training is performed on randomly cropped patches. Moreover, FlyingChairs and FlyingThings3D, have a somehow uniform motion distribution. Sintel is more unbalanced, and KITTI is largely unbalanced (supplementary Sec.6). For RAFT and IRR-PWC the sign imbalance is reduced of 40% its original value with the training on FlyingThings3D, this can be observed on IP(Co), R-M'(C), and IP(T), R-M'(T). The models R-M'(S) and IP(S) show that fine tuning on Sintel does not significantly affect the sign imbalance. On the other hand, fine tuning on KITTI leads to largely unbalanced models, mostly for IP(K).

Sign imbalance and groundtruth magnitude comparison. The sign imbalance $\overline{||I||}$ and the EPE may vary depending on the testing groundtruth average magnitude $\overline{||G||}$: larger displacements on the test set lead to larger errors. For this reason, we evaluate the ratio between sign imbalance and groundtruth motion $I_G = \overline{||I||}/\overline{||G||} \cdot 100 = [\%]$. Sin-

Table 2: Network performance summary, organized by label, model, training data "DATA", forward (FWD) or backward (BCK) training optical flow direction, probability of applying horizontal reflections $p(T_{lr})$, and vertical reflections $p(T_{ud})$. "EPE" is the EPE(O, G), "EPE$_{180}$ is EPE$(O^*, -G)$", $\overline{||I||}$ is the Euclidean sign imbalance. Evaluating the mirroring data augmentation for FlyingChairs, for FlyingThings3D. The "x" means "True", the label "-" means "False".

							Sintel *clean*			Sintel *final*		
label	model	DATA	FWD	BCK	$p(T_{lr})$	$p(T_{ud})$	EPE	EPE$_{180}$	$\overline{\lvert\lvert I\rvert\rvert}$	EPE	EPE$_{180}$	$\overline{\lvert\lvert I\rvert\rvert}$
DF(C)	DDFlow	C	x	x	0.5	0.5	3.85	3.87	1.66	4.93	4.93	2.06
IP(Co)	IRR-PWC	Co	x	x	0.5	0.5	2.34	2.36	1.3	3.96	4.05	2.09
Ro(C)	RAFT	C	x	-	0.5	0.1	2.15	2.23	1.27	4.44	4.45	2.35
G(T)	RAFT	T	x	x	0.5	0.1	**1.3**	**1.36**	**0.73**	2.73	**2.75**	**1.32**
IP(T)	IRR-PWC	T	x	x	0.5	0.5	1.87	1.86	0.92	3.46	3.46	1.35
Ro(T)	RAFT	T	x	x	0.5	0.1	1.47	1.44	0.84	**2.71**	2.8	1.4
IP(S)	IRR-PWC	S	x	-	0.5	0.5	1.91	1.87	0.98	2.5	2.43	1.32
Ro(S)	RAFT	S	x	-	0.5	0.1	0.74	1.01	0.7	1.19	1.7	1.17
IP(K)	IRR-PWC	K	x	-	0.5	0.5	7.41	7.09	7.07	8.07	7.72	6.05
RAFT for different mirroring data augmentation probabilities.												
R'(C)	RAFT	C	x	-	0	0	2.25	2.29	1.49	4.36	**4.31**	2.74
R-M'(C)	RAFT	C	x	-	0.5	0.5	**2.19**	2.25	**1.2**	4.39	4.37	**2.1**
Ro'(C)	RAFT	C	x	-	0.5	0.1	**2.19**	**2.17**	**1.22**	**4.24**	4.49	2.39
R'(T)	RAFT	T	x	x	0	0	1.54	1.52	1.13	2.8	**2.77**	1.46
Ro'(T)	RAFT	T	x	x	0.5	0.1	1.58	1.44	0.95	2.83	2.95	1.47
R-M'(T)	RAFT	T	x	x	0.5	0.5	**1.42**	**1.39**	**0.78**	**2.73**	2.84	**1.33**
RAFT trained on forward and backward optical flow.												
R-M'(C2)	RAFT	C2	x	x	0.5	0.5	2.15	**2.11**	**1.11**	**3.5**	**3.61**	**1.75**
R-M'(C2f)	RAFT	C2f	x	-	0.5	0.5	**2.14**	2.12	1.19	3.7	3.73	2.08
R-M'(Tf)	RAFT	Tf	x	-	0.5	0.5	1.45	**1.49**	**0.77**	2.75	2.75	**1.32**
R-M'(C2-T)	RAFT	C2-T	x	x	0.5	0.5	**1.43**	1.59	0.91	**2.71**	2.68	1.39
RAFT fine tuned models.												
R-M'(S)	RAFT	S	x	-	0.5	0.5	0.83	**0.79**	**0.55**	1.36	**1.33**	**0.73**
R'(S)	RAFT	S	x	-	0	0	**0.56**	1.24	1.05	**0.87**	2.31	2.05
R'(K)	RAFT	K	x	-	0	0	4.63	4.83	5.24	6.91	6.91	8.27
Ro'(K)	RAFT	K	x	-	0.5	0.1	4.43	4.19	4.27	6	**6.05**	6.23
R-M'(K)	RAFT	K	x	-	0.5	0.5	**3.95**	**3.95**	**4.03**	**5.6**	6.14	**5.93**

tel $\overline{||G||} \approx 13.5$ px. The mean ratio I_G for all networks trained on FlyingChairs, or FlyingChairs-FlyingThings3D, and tested on Sintel is between 10%, and 20%. When fine tuning on KITTI and testing on Sintel the imbalance is very large, I_G =52%.

Sign imbalance and EPE comparison. Overall, models with lower EPE exhibit lower $\overline{||I||}$. However, this is not always true. Among the models trained on FlyingChairs, DDFlow shows a higher EPE, if compared to IP(Co). However, IP(Co) and DF(C) show a similar $\overline{||I||}$. This should not depend on to the forward and backward optical flow consistency check, as both network perform forward and backward consistency check for occlusion reasoning. Thus, it might be due to the unsupervised loss. This can be better observed by evaluating the ratio $I_R = 100 \cdot \overline{||I||}/EPE$. On average, the imbalance is around 60% of the EPE, for all networks (trained on C or C-T). However, on Sintel, DDFlow display the lowest $I_R \approx 40\%$.

Mirroring data augmentation effect on sign imbalance. The yellow and green highlighted models in Tab. 2 help evaluate the training mirroring data augmentation effects. RAFT (Ro) has an uneven mirroring probability distribution. This is counterintuitive, but it can be explained by observing the EPE. Among the RAFT trainings on FlyingChairs, Ro, obtains the lowest EPE, however if we extend the evaluation to the EPE$_{180}$ and to the sign imbalance, Ro is not the best model. This might be due to the fact that Sintel motion distribution is slightly unbalanced (supplementary Sec.6), and could positively reward a moderate sign imbalance. In fact, the models fine tuned on Sintel obtain a lower EPE when the data augmentation is not applied, R'(S). Applying the mirroring data augmentation, R-M'(S), significantly decreases the sign imbalance but also increases the EPE in certain cases. When RAFT is trained on FlyingThings3D, the best retrained model, R-M'(T), uses mirroring.

Table 3: Sign imbalance mitigation. “Ens. Testing” refers to the ensemble inference used during testing, “Ens. training” refers to ensemble inference applied during training (Sec. 3.3) , “GRAD” indicates the gradient aggregation type (Sec. 3.4); β loss weighing hyperparameter. Results show: the ensemble testing completely solve the sign imbalance bias and reduces the EPE. Models retrained with the loss function reduce the sign imbalance. Lower values of β improve accuracy. Our retrained models also obtain the lowest EPE. “x” means “presence of.”, “-” means “absence of.”.

							Sintel *clean*			Sintel *final*		
label	model	data	β	GRAD	Ens. Training	Ens. Testing	EPE	EPE_{180}	$\overline{\lvert\lvert I\rvert\rvert}$	EPE	EPE_{180}	$\overline{\lvert\lvert I\rvert\rvert}$
DF(C)	DDFlow	C	-	-	-	x	3.78	3.78	0	4.82	4.82	0
IP(Co)	IRR-PWC	Co	-	-	-	x	2.27	2.27	0	3.87	3.87	0
IP(T)	IRR-PWC	T	-	-	-	x	1.81	1.81	0	3.4	3.4	0
R’(T)	RAFT	T	-	-	-	x	1.35	1.35	0	2.7	2.7	0
Ro’(T)	RAFT	T	-	-	-	x	1.45	1.45	0	2.8	2.8	0
R-M’(T)	RAFT	T	-	-	-	x	1.36	1.36	0	2.71	2.71	0
R-M’(T)	RAFT	T	0.3	FWDs	x	-	**1.28**	**1.28**	**0.54**	**2.6**	**2.64**	1.02
R-M’(T)	RAFT	T	0.6	FWDg	x	-	1.42	1.38	0.43	2.66	2.7	0.77
R-M’(T)	RAFT	T	1	FWDg	x	-	1.6	1.63	**0.37**	2.78	2.79	**0.54**
R-M’(T)	RAFT	T	0.3	FWDs	x	x	**1.25**	**1.25**	0	**2.57**	**2.57**	0
R-M’(T)	RAFT	T	0.6	FWDg	x	x	1.39	1.39	0	2.65	2.65	0

4.3. Mitigating the sign imbalance

In this section we evaluate our proposed methods to reduce the sign imbalance. Table 3 shows the effect of our ensemble inference and loss function strategies.

Ensemble inference. Table 3 results show that the ensemble inference strategy proposed in Sec. 3.3 completely reduce the imbalance, and always improves the EPE on all the testing datasets, if compared with the same models in table 2. This comes at the cost of a doubled inference time.

Sign imbalance loss. Overall, when training on FlyingChairs, very large values of β, considerably decrease $\overline{||I||}$, and considerably decrease the EPE on Sintel *final* (supplementary Sec. 10-11). When fine tuning on FlyingThings3D, reducing the sign imbalance is more difficult. For this training dataset; large values of β do not significantly reduce the EPE on Sintel. Values of $\beta \leq 0.6$ reduce the imbalance and the EPE slightly; $\beta = 1$ provide a strong sign imbalance mitigation, but can show a slight EPE penalty Sintel *clean*. Instead, when fine tuning on strongly unbalanced datasets, such as KITTI, the loss function can dramatically reduce the sign imbalance (Supplementary Sec. 11). The FWDs and the FWDg strategies perform similarly. FWDg shows a higher imbalance reduction for the same values of β, but shows an overall slightly higher EPE on Sintel *clean*. However, there is a complex interplay between our loss strategies and the mirroring data augmentation, on a high level, FWDs strategies obtain a lower EPE without mirroring augmentation (Supplementary Sec.10-11).

Best models. Applying the ensemble method during training together with our loss function leads to the best performance. Table 3 shows that the models showing the lowest EPE and sign imbalance: is R-M’(T) trained with $\beta = 0.3$ and tested with the ensemble inference. The second best model in terms of EPE is **R-M’(T)** trained with $\beta = 0.3$, **without using the ensemble inference during testing**. (Please note that the previous two models listed seem to outperform the very recent deep equilibrium networks [2], on Sintel training). Increasing β to 0.6 still leads to an EPE lower than the baseline and leads roughly to a 60% sign imbalance decrease. Increasing β to 1.0 increase the imbalance mitigation, but starts to increase the EPE on Sintel *clean*.

5. Discussion and conclusion

In this paper we provide a methodology and metric to measure and mitigate the sign imbalance, a special case of lack of equivariance, for optical flow estimators. Supported by the experimental evidence we answer the three questions raised in the introduction (RQ1): *To which extent do the SOTA OF estimators quantitatively display sign imbalance?* We tested the top performing optical flow estimators based on challenging leaderboards and measured the amount of sign imbalance. We found that almost all the tested models show sign imbalance to a considerable extent, and that more accurate models lead to a lower sign imbalance. (RQ2) *What are the main causing factors?* We analyzed different components in the Deep Learning pipeline: the training and testing data, the architecture, the data augmentation. We tested different models relying on considerably different architectures. Mirroring the training data, or using forward and backward optical flow provides only a marginal sign imbalance mitigation: the EPE training metric cannot account for the sign imbalance by design, leading to unbalanced estimates. (RQ3) *Can such bias be mitigated, and how?* The sign imbalance can be completely mitigated at the cost of a doubled inference time, or partially mitigated, without any increase in inference time. Reducing the sign imbalance also reduces the EPE.

References

[1] Argaw, D.M., Kim, J., Rameau, F., Cho, J.W., Kweon, I.S.: Optical flow estimation from a single motion-blurred image. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 891–900 (2021)

[2] Bai, S., Geng, Z., Savani, Y., Kolter, J.Z.: Deep equilibrium optical flow estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 620–630 (June 2022)

[3] Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. International Journal of Computer Vision **92**(1), 1–31 (2011)

[4] Bar-Haim, A., Wolf, L.: Scopeflow: Dynamic scene scoping for optical flow. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7995–8004 (2020)

[5] Barron, J.L., Fleet, D.J., Beauchemin, S.S.: Performance of optical flow techniques. International journal of computer vision **12**(1), 43–77 (1994)

[6] Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: European Conf. on Comp. Vision (ECCV). Part IV, LNCS 7577, Springer-Verlag (Oct 2012)

[7] Dieleman, S., Fauw, J.D., Kavukcuoglu, K.: Exploiting Cyclic Symmetry in Convolutional Neural Networks. In: Proc. of The 33rd Int. Conf. on Machine Learning. Proc. of Machine Learning Research, vol. 48. PMLR, New York, New York, USA (20–22 Jun 2016)

[8] 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: Proceedings of the IEEE International Conference on Computer Vision. pp. 2758–2766 (2015)

[9] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE Conf. on Comp. Vision and Pattern Recognition (2012)

[10] Horn, B.K., Schunck, B.G.: Determining optical flow. Artificial intelligence **17**(1-3), 185–203 (1981)

[11] Hur, J., Roth, S.: Iterative residual refinement for joint optical flow and occlusion estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5754–5763 (2019)

[12] Ilg, E., Saikia, T., Keuper, M., Brox, T.: 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)

[13] Jeong, J., Lin, J.M., Porikli, F., Kwak, N.: Imposing consistency for optical flow estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3181–3191 (2022)

[14] Jiang, S., Campbell, D., Lu, Y., Li, H., Hartley, R.: Learning to estimate hidden motions with global motion aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9772–9781 (2021)

[15] Jonschkowski, R., Stone, A., Barron, J.T., Gordon, A., Konolige, K., Angelova, A.: What matters in unsupervised optical flow. In: European Conference on Computer Vision. pp. 557–572. Springer (2020)

[16] Kondermann, D., Nair, R., Honauer, K., Krispin, K., Andrulis, J., Brock, A., Gussefeld, B., Rahimimoghaddam, M., Hofmann, S., Brenner, C., et al.: The hci benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 19–28 (2016)

[17] Kondor, R., Trivedi, S.: On the generalization of equivariance and convolution in neural networks to the action of compact groups. In: International Conference on Machine Learning. pp. 2747–2755. PMLR (2018)

[18] Lenc, K., Vedaldi, A.: Understanding Image Representations by Measuring Their Equivariance and Equivalence. Int. J. of Comp. Vision **127**(5) (May 2019)

[19] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42** (2017)

[20] Liu, L., Zhang, J., He, R., Liu, Y., Wang, Y., Tai, Y., Luo, D., Wang, C., Li, J., Huang, F.: Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)

[21] Liu, P., King, I., Lyu, M.R., Xu, J.: Ddflow: Learning optical flow with unlabeled data distillation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8770–8777 (2019)

[22] Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: 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

[23] Mayer, N., Ilg, E., Fischer, P., Hazirbas, C., Cremers, D., Dosovitskiy, A., Brox, T.: What makes good synthetic training data for learning disparity and optical flow estimation? International Journal of Computer Vision pp. 1–19 (2018)

[24] Mendes, P.A.S., Paulo Coimbra, A.: Movement detection and moving object distinction based on optical flow for a surveillance system. In: Ao, S.I., Gelman, L., Kim, H.K. (eds.) Transactions on Engineering Technologies. pp. 143–158. Springer Singapore, Singapore (2021)

[25] Menze, M., Heipke, C., Geiger, A.: Joint 3d estimation of vehicles and scene flow. In: ISPRS Workshop on Image Sequence Analysis (ISA) (2015)

[26] Rippel, O., Nair, S., Lew, C., Branson, S., Anderson, A.G., Bourdev, L.: Learned video compression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3454–3463 (2019)

[27] Savian, S., Elahi, M., Tillo, T.: Benchmarking the imbalanced behavior of deep learning based optical flow estimators. In: 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). pp. 151–158. IEEE (2019)

[28] Savian, S., Elahi, M., Tillo, T.: Optical flow estimation with deep learning, a survey on recent advances. In: Deep biometrics, pp. 257–287. Springer (2020)

[29] Simard, P.Y., Steinkraus, D., Platt, J.C.: Best practices for convolutional neural networks applied to visual document analysis. In: 7th Int. Conf. on Document Analysis and Recognition, 2003. Proceedings. (2003)

[30] Stone, A., Maurer, D., Ayvaci, A., Angelova, A., Jonschkowski, R.: Smurf: Self-teaching multi-frame unsupervised raft with full-image warping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3887–3896 (2021)

[31] Sui, X., Li, S., Geng, X., Wu, Y., Xu, X., Liu, Y., Goh, R., Zhu, H.: Craft: Cross-attentional flow transformer for robust optical flow. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17602–17611 (2022)

[32] Sun, D., Herrmann, C., Reda, F., Rubinstein, M., Fleet, D., Freeman, W.T.: What makes raft better than pwc-net? arXiv preprint arXiv:2203.10712 (2022)

[33] Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8934–8943 (2018)

[34] Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Models matter, so does training: An empirical study of CNNs for optical flow estimation. IEEE transactions on pattern analysis and machine intelligence **42**(6), 1408–1423 (2019)

[35] Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European Conference on Computer Vision. pp. 402–419. Springer (2020)

[36] Wang, J., Zhong, Y., Dai, Y., Birchfield, S., Zhang, K., Smolyanskiy, N., Li, H.: Deep two-view structure-from-motion revisited (2021)

[37] Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Tao, D.: Gmflow: Learning optical flow via global matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8121–8130 (2022)

[38] Yuan, Y., Su, W., Ma, D.: Efficient dynamic scene deblurring using spatially variant deconvolution network with optical flow guided training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)

[39] Zhang, T., Zhang, H., Li, Y., Nakamura, Y., Zhang, L.: Flowfusion: Dynamic dense rgb-d slam based on optical flow. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 7322–7328 (2020)

[40] Zhao, C., Li, D., Feng, C., Li, S.: Of-umrn: Uncertainty-guided multitask regression network aided by optical flow for fully automated comprehensive analysis of carotid artery. Medical Image Analysis **70**, 101982 (2021)

[41] Zhou, H., Ummenhofer, B., Brox, T.: Deeptam: Deep tracking and mapping with convolutional neural networks. International Journal of Computer Vision **128**(3), 756–769 (2020)