Imposing Consistency for Optical Flow Estimation

Jisoo Jeong[1] Jamie Menjay Lin[2,†] Fatih Porikli[1] Nojun Kwak[3,‡]
[1]Qualcomm AI Research* [2]Google Research [3]Seoul National University
{jisojeon,fporikli}@qti.qualcomm.com jmlin@google.com nojunk@snu.ac.kr

Abstract

Imposing consistency through proxy tasks has been shown to enhance data-driven learning and enable self-supervision in various tasks. This paper introduces novel and effective consistency strategies for optical flow estimation, a problem where labels from real-world data are very challenging to derive. More specifically, we propose occlusion consistency and zero forcing in the forms of self-supervised learning and transformation consistency in the form of semi-supervised learning. We apply these consistency techniques in a way that the network model learns to describe pixel-level motions better while requiring no additional annotations. We demonstrate that our consistency strategies applied to a strong baseline network model using the original datasets and labels provide further improvements, attaining the state-of-the-art results on the KITTI-2015 scene flow benchmark in the non-stereo category. Our method achieves the best foreground accuracy (4.33% in Fl-all) over both the stereo and non-stereo categories, even though using only monocular image inputs.

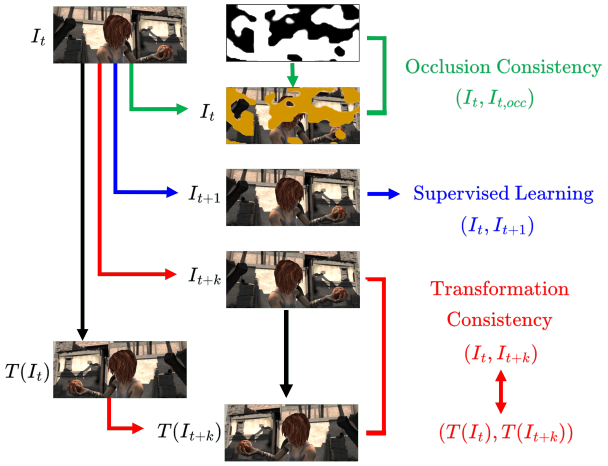

Figure 1. During training, we enforce occlusion consistency with self-supervision by applying random occlusion patterns and imposing the network to detect the regions under occlusion between consecutive images (I_t, I_{t+1}). We also employ transformation consistency (equivariance to geometric transformations) in a semi-supervised manner for an image pair (I_t, I_{t+k}) and the transformed pair $(T(I_t), T(I_{t+k}))$ with $k \geq 1$.

1. Introduction

Optical flow characterizes dense displacements between corresponding pixels across images, e.g. between two consecutive frames in a video [9, 19, 40, 43]. It is widely employed in video analysis applications including video compression [32, 46], action recognition [6, 29], video denoising [3, 8], and object tracking [25, 52], to point out a few.

As important as its, optical flow estimation comes with significant challenges. Occlusions due to camera and object motions present one inherent difficulty, where a part of the scene is visible in one but not in the other image of the pair. Several methods addressed this problem by explicitly estimating regions to be excluded [34, 51], by applying self-supervision [31], or by incorporating contextual information [43]. These methods, however, had limited reception since they rely on multiple forward-backward iterations for predicting occlusion areas [34, 41] or fail for larger occlusions.

Obtaining precise annotations for optical flow is another challenge that directly impacts the learning performance. Since pixel-level motion annotation requires specialized and costly data acquisition systems, and in many cases, such annotations do not support high precision and spatial resolution, optical flow datasets are limited in number, variety, and degree of realism [9, 19]. The need for large-scale real-world datasets, therefore, becomes a bottleneck.

To mitigate the annotation issues, unsupervised learning [20, 24, 34, 45] and semi-supervised learning [27, 47] methods have been proposed in the past. Unsupervised learning schemes, however, typically result in degraded performance, lagging behind fully supervised learning

* Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
† This work was done while at Qualcomm AI Research.
‡ Nojun Kwak was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (2021R1A2C3006659).

counterparts [24, 30, 45]. In comparison, semi-supervised learning [27] may offer potential performance gains with data augmentation along with generative adversarial networks [14].

In this paper, we introduce two consistency strategies for optical flow estimation to address these challenges as depicted in Fig. 1. First, we propose occlusion consistency that generates a random occlusion mask, which is used to create additional image pairs, and constrains the network to predict the mask and a zero-forced flow field in a self-supervised manner. Unlike other approaches, our occlusion consistency allows generating occlusion ground truth without forward-backward iterations. Although this intuitive strategy is simple, it enables the network not to confuse occlusion patterns as motion indicators without losing its representative capacity for the unoccluded image regions. It also helps the network to derive more informative features for the partially occluded regions within local receptive fields of the kernels without requiring additional labeling.

We also incorporate a transformation-based consistency regularization that has been shown useful in semi-supervised image classification and object detection tasks [21, 22, 28, 36, 42]. This strategy helps the model impose *equivariance* through such consistency regularization. We apply whole-image geometric transformations including flippings, translations, and rotations. Then we restore the transformation before evaluating the overall transformation consistency losses. While our transformation consistency is derived with two passes of forward flow estimation, the cycle consistency [44] is computed with one pass of forward and the other pass of backward flow estimation. To the best of our knowledge, this is the first attempt to impose equivariance through consistency regularization for optical flow estimation. Note that our approach is different from conventional data augmentation schemes, which expand training samples without imposition of sophisticated consistency losses during training.

Our proposed self- and semi-supervised consistency learning strategies not only complement the previous state-of-the-art RAFT [43] baseline, but enable significant improvement in the model accuracy performance as evidenced in our experiment results. Our proposed method achieves the new state-of-the-art accuracies and has ranked at the top of the KITTI-2015 scene flow non-stereo leaderboard (Ours: 4.33%, 6.01%, 3.99% vs. RAFT: 5.10%, 6.87%, 4.74% in Fl-all, Fl-fg, and Fl-bg, respectively). Our training with consistency strategies can potentially be adapted to other dense prediction tasks.

In summary, our main contributions are as follows:

- We propose a novel occlusion consistency strategy, which facilitates learning occlusion-robust representations efficiently in a self-supervised manner.
- We incorporate transformation consistency equivariance enabling learning from a more diverse set of image pairs without additional labeling.
- Applying these two consistency strategies jointly in training and integrating an occlusion estimation channel in the architecture, our model generates superior results over its baseline achieving state-of-the-art performance in the KITTI-2015 scene flow non-stereo monocular dataset.

2. Related Work

Optical Flow: Classic solutions have been studied for decades [4, 15], and recent advancements have been made with deep learning methods [9,19,37,39,43,51]. RAFT [43] demonstrates notable improvement by extracting per-pixel features from the corresponding image pair (I_t, I_{t+1}), building multi-scale 4-dimensional correlation volumes for all pixel pairs, and iteratively adjust the flow estimates through a refinement module with gated recurrent units (GRUs) [7] with repeated lookups in the correlation volume. The loss is computed between the ground truth optical flow $f(I_t, I_{t+1})$ and the predicted optical flow $\tilde{f}^i(I_t, I_{t+1})$ in each iteration i with ℓ_1 norm

$$\mathcal{L}_{RAFT} = \sum_{i=1}^{N} \gamma^{N-i} \left\| f(I_t, I_{t+1}) - \tilde{f}^i(I_t, I_{t+1}) \right\|_1, \quad (1)$$

where N is the number of GRU iterations and γ is a decay factor ($\gamma < 1$). The final predicted flow is then $\tilde{f}(I_t, I_{t+1}) = \tilde{f}^N(I_t, I_{t+1})$, the prediction after all iterations.

Methods for Occlusion Handling: UnFlow [34] identifies occlusions with the forward-backward constraint assumption [41] and excludes the occlusion area during training. For the forward-backward constraint, a bidirectional optical flow is required, and the errors could accumulate and propagate, partially due to the discretization of continuous values in the estimates. Self-supervised learning has also been introduced in recent works for optical flow estimation. SelFlow [31], as an example, performs flow estimation for non-occluded regions and uses these predictions to estimate flows in occluded regions. However, it requires four optical flow inferences (forward/backward$\times$occlusion/non-occlusion pairs) and significantly increases computational and memory costs to obtain occlusion maps and non-occlusion/occlusion flows. Maskflownet [51] proposes a learnable occlusion mask, which is applied to the next image frame I_{t+1} when calculating the correlation between the features of I_t and I_{t+1}. Recent studies [18, 24] also propose predicting the occlusion mask with an additional channel, and we adopt this approach.

Another solution is to integrate contextual information. Recently, RAFT [43] presented a context sub-network to

(a) minor occlusion

(b) major occlusion

Figure 2. Illustration of occlusion problems: (a) In a case of minor occlusion, incorrect optical flow estimations for the occlusion area can be corrected using larger spatial context (red to green). (b) However, in a case of major occlusion, the occlusion area can degenerate the accurately estimated optical flow of the smaller visible region (green to red)

incorporate neighborhood pixels' information. By assuming the pixels in an object or segment to have a similar flow, it refines the estimated flow fields in occlusion areas. However, as shown in Fig. 5 (RAFT results), the matched parts can be incorrectly updated in case of severe occlusions. We analyze contextual information in more detail in the following subsection.

In contrast to previous algorithms, our method generates occlusion itself and enforces the network to predict the occlusion areas without multiple inferences.[1]

Contextual Information: Using context to regularize estimations within an image segment can improve optical flow as [43] intended with the context sub-network. However, such a regularization needs to be imposed while keeping the degree of occlusion in mind. Figure 2 shows an example. In the case of minor occlusions, most pixels (in green) in a contextual segment (car) are likely to be estimated correctly. Here, the context sub-network may provide adequate support over the refinement iterations. On the other hand, in the case of major occlusions, the dominating portion of the occlusion region (in red) can be biased towards incorrect context, creating possibly significant deterioration in the correspondence estimation. RAFT estimation in Figure 5 gives a real example of this problem occurring under a major occlusion. To tackle this problem, we propose the occlusion consistency strategy, as described in Section 3.1.

Self-Supervised Learning: By defining pretext tasks for unlabeled data and then using them to pretrain models, self-supervision allows making the best use of the unlabeled data and enhancing the performance of the downstream tasks [13, 50]. In [50], the image is rotated by a random angle, and this angle is predicted. With this auxiliary task of rotation estimation, the network makes room for performance improvement in the original task. However, the use of this auxiliary task is reported to underperform in supervised settings while it performs better in semi-supervised and self-supervised settings [13, 50].

Semi-Supervised Learning: Data augmentation with consistency regularization has been popular in semi-supervised learning [28, 36, 42] where a set of predefined transformations are applied to the original labeled data and the outputs of the perturbed inputs are enforced to agree with the outputs of the original data [28]. The loss is defined as the mismatch between the outputs for the original and perturbed inputs. It is shown that consistency regularization improves robustness by smoothing the underlying data manifold [36]. The consistency regularization loss and the supervised loss is often aggregated. Similar ideas are also applied localization problems, and demonstrated better performance [21, 22]. In our work, we extend this promising concept to optical flow estimation.

There have also been studies on semi-supervised optical flow estimation to reduce dependency on the labeled data. In [27], an adversarial learning setup is used where the discriminator learns whether an optical flow is real (by comparison with the ground truth) or generated with a model. In the process of minimizing the discriminator loss, the generator with unlabeled data pairs is trained. In [47], clean images are generated from foggy images, and foggy images are generated from clean images. A model is trained with interchangeable samples among clean and foggy images. These algorithms require additional networks to translate images into flow estimates. In our proposal, we do not require any separate network as a part of our training framework as we derive equivariance-based consistency losses simply by comparing the original pairs with the transformation pairs.

3. Consistency for Optical Flow

Here, we summarize the notations used in this paper. We denote the ground truth optical flow as $f(I_t, I_{t+k})$ and the predicted optical flow as $\tilde{f}(I_t, I_{t+k})$ between two images I_t and I_{t+k} that are k apart in time. Image size is $w \times h$. An occluded version of the original image I_t and its corresponding occlusion mask are denoted as $I_{t,occ}$ and O_t, respectively. We denote the predicted occlusion mask as $\tilde{O}_t$. We also use $T(\cdot)$ and $R(\cdot)$ to denote the operations of transformation and transformation restoration, respectively.

The consistency strategies we describe below are applied in a self- and semi-supervised manner, which requires no additional ground truths.

[1] Note that our contribution is not simply adding a channel but proposing a new scheme that generates and trains occlusion without occlusion prediction.

3.1. Occlusion Consistency

In this subsection, we discuss two techniques in our occlusion consistency strategy: zero forcing and mask match loss.

Zero Forcing: In order to apply meaningful occlusions to images, we define an occlusion mask $O_t \in \mathbb{R}^{w \times h}$. We adopt the cow-mask [10, 11] to create sufficiently random yet locally connected occlusion patterns as an occlusion could occur in any size, any shape, and at any position in an image while exhibiting locally explainable structures. Occlusions are mainly perpendicular to motion direction (depth discontinuities) for moving objects (camera motion) around object boundaries (scene depth discontinuities), thus occlusion regions are often connected. Using self-supervised learning with random occlusion masks enables our network to respond and learn such complex occlusion structures in the scene.

In a self-supervised manner, we apply the occlusion mask to a single image by multiplying pixel-wise the occlusion mask with the image, which allows us to obtain a new image pair $(I_t,\ I_{t,occ})$ without requiring any ground truth. Each entry of the occlusion mask O_t takes a binary value; $O_t(p) = 1$ indicating a non-occluded pixel p and $O_t(p) = 0$ corresponds to a masked pixel. We impose the flow to be zero, i.e, $\tilde{f}(I_t, I_{t,occ}) = 0$, as there is no motion but only occlusion. This allow us to compute the zero-forcing loss as

$$\mathcal{L}_{ZF} = \sum_{i=1}^{N} \gamma^{N-i} \left\| \tilde{f}^i(I_t, I_{t,occ}) \right\|_1 . \tag{2}$$

As an enhancement to the occlusion consistency, we further introduce a special case in which $O_t = 1$ (no occlusion), meaning two images in the newly formed pair are identical, i.e., the pair to be (I_t, I_t), which results in the new zero-forcing loss

$$\mathcal{L}_{ZF^*} = \sum_{i=1}^{N} \gamma^{N-i} \left\| \tilde{f}^i(I_t, I_t) \right\|_1 . \tag{3}$$

Mask Match Loss: Since we can generate occlusion masks automatically, our intuition is that we can also estimate them in our network and reinforce another consistency by matching the generated O_t and estimated $\tilde{O}_t$ masks. To achieve this, we introduce one additional channel in the output of our network to estimate the occlusion status of pixels. This also facilitates better feature correspondences for correlation volumes as the network can directly access an internal occlusion mask in its layers. Furthermore, occlusion mask estimation can be refined over iterations and along with supervision. Therefore, we employ the zero-forcing loss together with an occlusion mask match loss simultaneously and iteratively in our occlusion consistency strategy.

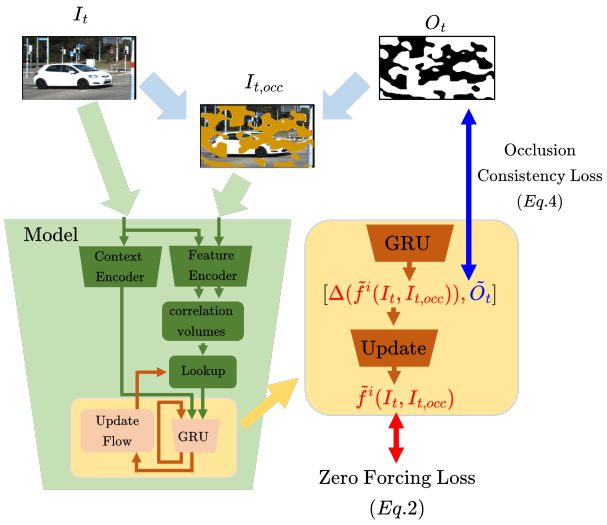

Figure 3. **Occlusion consistency:** A random mask is applied to the original image I_t to construct $I_{t,occ}$. Then, the optical flow, as well as the occlusion mask, are estimated for the image pair $(I_t, I_{t,occ})$. In this case, the target ground truth is $f(I_t, I_{t,occ}) = 0$.

We define the mask match loss as

$$\mathcal{L}_{MM} = \sum_{i=1}^{N} \gamma^{N-i} \left(-\frac{1}{wh} \sum_{p} O_t(p) log(\tilde{O}_t^i(p)) \right) \tag{4}$$

Here, we use the cross entropy, γ and N are the same parameters as defined in (1).

3.2. Transformation Consistency

Transformation consistency strategy leverages two methods; consistency regularization and frame-hopping with semi-supervised learning.

We apply spatial transformation consistency to the input image pair, creating cases for enforcing equivariance between the estimated optical flow for the original pair and the estimated optical flow for the transformed pair, in addition to the supervised loss of optical flow (See Fig. 4). In addition, as an enhancement to this transformation consistency methods, we extend the temporal gap from $k = 1$ to $k \geq 1$ to include pairs where the images depict larger motions. Existing datasets typically provide ground truth flow fields $f(I_t, I_{t+1})$ only between consecutive image frames I_t and I_{t+1}, while the image sampling rates may vary[2] significantly from one dataset to another. Allowing pairs with larger frame gaps enables more versatile characterization of underlying object and camera motion with different speeds.

Consistency Regularization: Optical flow estimations should equivariantly change when the input images in the pair undergo the same spatial (geometric) transformations

[2]For example, the frame rate of the Sintel [5] dataset is 24 frames-per-second, while that of the KITTI [12] is 10 frames-per-second.

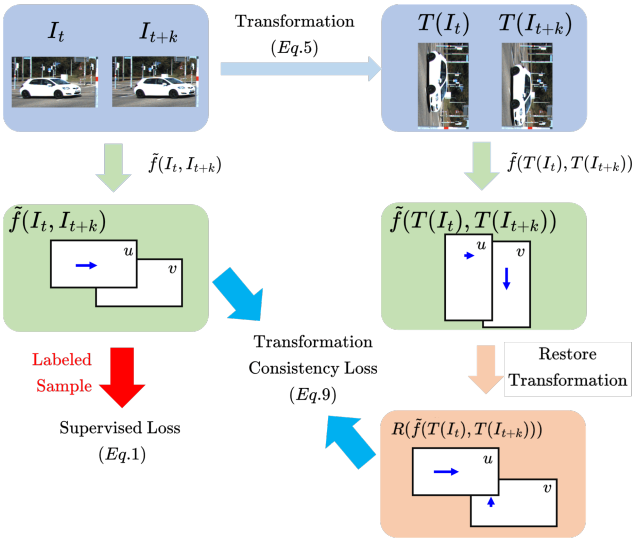

Figure 4. **Transformation consistency**. $T(I_t)$ and $T(I_{t+k})$ are generated with image-wise transformations (random rotation as illustrated) for the image pair (I_t, I_{t+k}). Optical flows $\tilde{f}(I_t, I_{t+k})$ and $\tilde{f}(T(I_t), T(I_{t+k}))$ are computed by the same model for the image pair and its transformed image pair. Then, the estimated flow for the transformed pair are remapped by applying the transformation restoration operation. In case we have labeled data, a supervised loss is calculated between $\tilde{f}$ and the ground truth f.

that are bijective. We take advantage of this property and impose an intuitive consistency regularization for the image pairs during the training process. More specifically, we apply 2D image transformations, including flips and random rotations that we observed to be effective choices, to the input images and corresponding estimated optical flows.

Figure 4 shows an example for the transformation consistency regularization. We transform both images I_t and I_{t+k} in the pair

$$I_t, I_{t+k} \overset{T}{\mapsto} T(I_t), T(I_{t+k}) \tag{5}$$

and compute the optical flow for the original and transformed pairs using our model. Our assumption is that after applying transformation restoration, the estimated optical flows should be equivalent

$$\tilde{f}(I_t, I_{t+k}) = R\left(\tilde{f}(T(I_t), T(I_{t+k}))\right). \tag{6}$$

Using this, we compute the transformation consistency loss $\mathcal{L}_{tr}$ between $\tilde{f}$ and $R(\tilde{f})$ as follow

$$\mathcal{L}_{tr} = \left\| \tilde{f}(I_t, I_{t+k}) - R\left(\tilde{f}(T(I_t), T(I_{t+k}))\right) \right\|_2^2. \tag{7}$$

During the initial phase of training, a larger transformation inconsistency $\mathcal{L}_{tr}$ is more likely to occur, thus the training may diverge. To alleviate this issue, we introduce an *identifier mask* α ($\alpha \in \mathbb{R}^{w \times h}$) as follows

$$\alpha^i = \begin{cases} 1, & \text{if } \mathcal{L}_{tr}^i < \epsilon \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Here, ϵ is a small positive constant, which is then used in the final loss function to prevent the network from diverging

$$\mathcal{L}_{TR} = \sum_{i=1}^{N} \gamma^{N-i} \cdot \mathbb{E}_{\mathbb{I}\{\alpha^i=1\}}(\mathcal{L}_{tr}^i). \tag{9}$$

where $\mathbb{I}\{\alpha^i = 1\}$ indicates that the expectation is fulfilled only for the ones in mask. For iterative flow refinement, $\mathcal{L}_{tr}^i$ is calculated in the i-th iteration as in (7) and γ and N are the same parameters as (1).

Frame Hopping: We also utilize *frame hopping*, a technique inspired by ScopeFlow [2]. Our intuition is that larger displacements in the datasets [5, 12] exist mostly near edges of images; thus, training with samples containing larger displacements can benefit model performance. Frame hopping (for image pairs (I_t, I_{t+k}) with $k > 1$) provides not only more training samples but also samples with larger displacements to enhance learning.

3.3. Aggregated Loss

Our total loss consists of the conventional supervised loss ($\mathcal{L}_{base}$), the zero-forcing loss ($\mathcal{L}_{ZF}$), the mask match loss ($\mathcal{L}_{MM}$), and the transformation consistency loss ($\mathcal{L}_{TR}$) as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{base} + \mathcal{L}_{ZF} + \lambda_1 \mathcal{L}_{MM} + \lambda_2 \mathcal{L}_{TR}. \tag{10}$$

The supervised loss (L_{base} in (1)) for labeled data and the unsupervised loss (L_{ZF} in (2)), (L_{MM} in (4)), and (L_{TR} in (9)) for unlabeled data are combined by using a balance parameter λ_1 and λ_2 to derive the final loss[3].

4. Experiments

Datasets & Implementation Details: In our experiments, we have utilized the FlyingChairs (C) [9], FlyingThings3D (T) [33], Sintel (S) [5], KITTI (K) [12, 35], and HD1K(H) [26] datasets, which are the most popular benchmarks in the optical flow estimation problem. More details on our experimental analysis are provided in the supplementary material.

All experiments have been conducted under the same setting with the official code of RAFT[4]. We followed the same batch sizes, optimizer, number of GRU iterations, and so on. As the number of image pairs increased in our method, we increased the number of iterations proportionally. Similar to RAFT, we pretrained our model in sequence

[3]Zero Forcing loss is computed with the same balance with supervised learning.

[4]https://github.com/princeton-vl/RAFT

Table 1. Optical Flow results for Sintel and KITTI. We trained the model with the Flyingchairs (C) and Flyingthings (T) datasets and tested the model on the training dataset of the Sintel (S) and KITTI (T). For Sintel and KITTI tests, we finetuned the model with a pre-trained model (C+T) with the Sintel, KITTI, and HD1K (H) training dataset. (Smaller numbers are better. The numbers in gray have little meaning because they are measured on the training data. † is trained including test images without label as unlabeled data, and ‡ is trained on KITTI-2012 and KITTI-2015 datasets. * is the results of warm-start, and § is the results of undisclosed method.)

Method	Training dataset	Sintel (train-EPE)		KITTI (train)		Sintel (test-EPE)		KITTI (test)
		(Clean)	(Final)	(Fl-epe)	(Fl-all)	(Clean)	(Final)	(Fl-all)
HD3 [49]	C+T	3.84	8.77	13.17	24.0	-	-	-
FlowNet2 [19]		2.02	3.54	10.08	30.0	3.96	6.02	-
PWC-Net [39]		2.55	3.93	10.35	33.7	-	-	-
LightFlowNet [16]		2.48	4.04	10.39	28.5	-	-	-
LightFlowNet2 [17]		2.24	3.78	8.97	25.9	-	-	-
VCN [48]		2.21	3.68	8.36	25.1	-	-	-
MaskFlowNet [51]		2.25	3.61	-	23.1	-	-	-
RAFT-small [43]		2.21	3.35	7.51	26.9	-	-	-
Ours (RAFT-small + OCTC)		**1.95**	**3.13**	**6.53**	**22.1**	-	-	-
RAFT [43]		1.43	2.71	5.04	17.4	-	-	-
Ours (RAFT + OCTC)		**1.31**	**2.67**	**4.72**	**16.3**	-	-	-
SelFlow [31]	C+T+S+K	1.68	1.77	-	1.18	3.74	4.26	8.42
ScopeFlow [2]		-	-	-	-	3.59	4.10	6.82
LiteFlowNet2 [49]	C+T+S+K+H	1.30	1.62	1.47	4.8	3.48	4.69	7.62
PWC-Net+ [40]		1.71	2.34	1.50	5.3	3.45	4.60	7.72
VCN [48]		1.66	2.24	1.16	4.1	2.81	4.40	6.30
MaskFlowNet [51]		-	-	-	-	2.52	4.17	6.10
RAFT [43]		0.76	1.22	0.63	1.5	1.94/1.61*	3.18/2.86*	5.10
CRAFT [1]	Undisclosed	-	-	-	-	1.45§	2.42§	4.79
RAFT-A [38]	A+T+S+K+H	-	-	-	-	2.01/ – *	3.14/ – *	4.78
GMA [23]	C+T+S+K+H	0.62	1.06	0.57	1.2	– /**1.39***	– /**2.47***	5.15
Ours (RAFT + OCTC)		0.73	1.23	0.67	1.7	1.82/ – *	3.09/ – *	4.72
Ours† (RAFT + OCTC)		0.74	1.24	0.71	2.0	1.58/ – *	**2.95**/ – *	–
Ours‡ (RAFT + OCTC)		-	-	0.78	2.3	**1.55**/1.41*	2.98/2.57*	**4.33**

with FlyingChairs and FlyingThings3D. Since Flyingchair samples do not have more than two consecutive images, only self-supervised learning was applied. The parameters are set to $(\lambda_1, \lambda_2) = (0.1, 0.01)$ in (10), $\epsilon = 5^2$ in (8), and k is set to 2. [5] For a wide variety of random patterns in occlusion consistency learning, we applied cowmask[6] with the same parameters used in [11]. All samples applied in our experiments are from the original datasets without additional data.

Experimental Results: Table 1 shows the performances of the proposed method and some very recent optical flow estimation algorithms. The model trained with C+T, RAFT reported the state-of-the-art performance previously. Nevertheless, we improved its performance even further when we applied our learning scheme OCTC (Occlusion Consistency and Transformation Consistency). In addition, our method outperformed others on the KITTI benchmark that contains real images. Our method achieved 0.26 and 0.22 EPE improvements in Sintel-clean and Sintel-final, respectively, in relation to RAFT-small. For the KITTI dataset, EPE decreased by an impressive 0.98, and Fl-all decreased by 4.8%. Using the RAFT-large model, our performance in predicting the optical flow still attained additional improvements; 0.12 and 0.04 smaller EPE for Sintel-clean and Sintel-final, and 0.32 EPE decrease and 1.1% Fl-all decrease for the KITTI dataset.

The bottom half of Table 1 presents the performance on the test datasets of Sintel and KITTI. The models are trained with the training datasets of Sintel and KITTI. For the model trained on the Sintel dataset, the test EPE decreased by 0.12 and 0.09 for clean and final, respectively, compared to RAFT. For the model trained on the KITTI-2015 dataset, the Fl-all score our model improves down to 4.72%. Furthermore, we trained our model with test images without labels treating them as unlabeled data. In Sintel, the test EPEs are 1.58 and 2.95 in the clean and final versions, respectively. Like MaskFlowNet, when we finetune on KITTI-2012 and KITTI-2015 together, our model shows further performance improvement with an Fl-all score of 4.33%, which achieves the new state of the art on the KITTI-2015 dataset. The proposed method has a gain of about 0.77% over the conventional RAFT model. And, when we applied

[5] We performed a grid search in $\{3^2, 5^2, 7^2, \infty\}$ for ϵ value in Eq.8 and over the values in $\{1.0, 0.1, 0.01, 0.001\}$ for each λ in Eq.10. The best hyperparameters found were $[\epsilon = 5^2, (\lambda_1, \lambda_2) = (0.1, 0.01)]$. More details and results of these experiments are provided in Supplementary File.

[6] https://github.com/google-research/google-research/tree/master/milking_cowmask

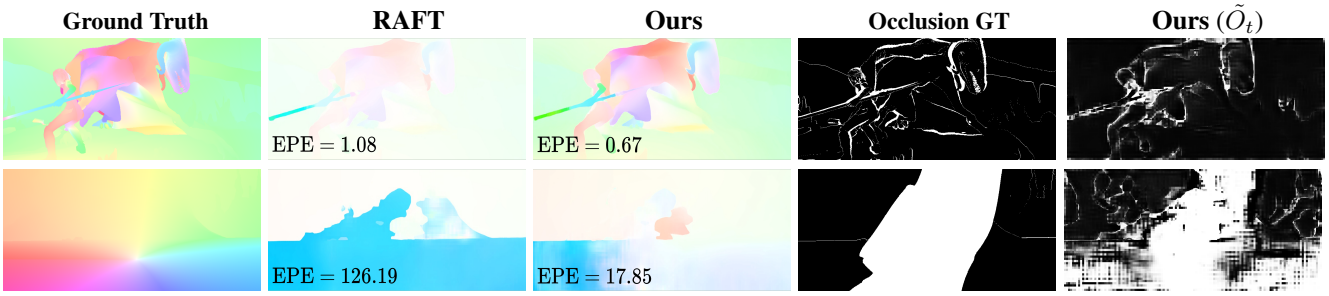

Figure 5. Qualitative results for the Sintel training set using RAFT and our RAFT+OCTC (Occlusion Consistency and Transformation Consistency) models (trained with C+T). The first row shows that our RAFT+OCTC, which adopts frame hopping in transformation consistency, works better for large displacements than RAFT. The second row shows that our RAFT+OCTC can predict occlusion area, and it helps our model prevent incorrect predictions.

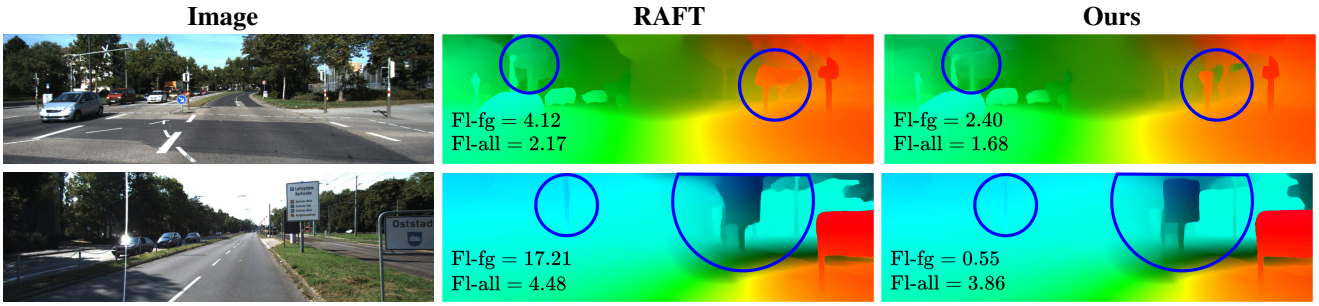

Figure 6. Qualitative results for the KITTI test set using RAFT and our RAFT+OCTC (Occlusion Consistency and Transformation Consistency) models (trained with C+T+S+K+H).

our method with warm-start, it also shows the performance improvement.

In comparison to other algorithms, our method brings robust improvements for both the Sintel and KITTI datasets. RAFT-A [38] shows performance improvement in the KITTI dataset, but its performance degrades in the Sintel dataset. GMA [23] reports state-of-the-art performance in the Sintel dataset, but its performance is not consistent; it is worse than the baseline RAFT in the KITTI dataset.

Qualitative Results: Figure 5 provides qualitative comparisons on the Sintel training dataset, where the scenarios of long-range movement and of large-area occlusion are shown in the top and bottom rows, respectively. In both scenarios, our model demonstrates improved accuracy than the RAFT baseline, indicating the effects of our consistency imposing strategies. Specifically, in the top row, our model trained with frame hopping enables improved handling with longer-range motions. In the bottom row, our RAFT+OCTC demonstrates improved robustness with large-area occlusions (see Supplementary file for more examples).

Figure 6 provides qualitative comparisons on the KITTI test dataset, where our algorithm also demonstrates improved consistency in the prediction outputs.

Table 2. Ablation study for Occlusion Consistency (OC). We trained our models with the Flyingchairs (C) and Flyingthings (T) datasets and tested on the training dataset of the Sintel (S) and KITTI (T). L_{ZF} and L_{MM} are zero-forcing loss in (2) and mask match loss in (4), respectively.

Method (small)	Additional Loss	Sintel (train-EPE)		KITTI-15 (train)	
		Clean	Final	Fl-epe	Fl-all
RAFT (baseline)	-	2.21	3.35	7.51	26.9
RAFT + OC	$\mathcal{L}_{ZF^*}(I_t, I_t)$	2.23	3.59	8.27	25.8
	$\mathcal{L}_{ZF}(I_t, I_{t,occ})$	2.17	3.35	7.22	24.2
	L_{MM}	2.11	3.31	7.14	24.3
	$\mathcal{L}_{ZF}(I_t, I_{t,occ}) + L_{MM}$	**2.05**	**3.18**	**7.07**	**23.5**

5. Discussion

Occlusion Consistency Terms: As shown in Table 2, when we initially used (I_t, I_t) for zero forcing (i.e., identical samples as a special case without occlusions), we observed a performance degradation possibly due to overfitting. As we applied occlusions in one of the samples $(I_t, I_{t,occ})$, we started to observe accuracy gains. We noticed that the combination of L_{MM} and zero forcing produced remarkable performance improvements, possibly a result of *mutual learning* in GRU with the simultaneous flow and occlusion predictions in the availability of context information.

Table 3. Ablation study for Transformation Consistency (TC). H and R are horizontal flips and random rotations (Other notations are the same as Table 2)

Method (small)	k	Transformation	Sintel (train-EPE)		KITTI-15 (train)	
			Clean	Final	Fl-epe	Fl-all
RAFT (baseline)	-	-	2.21	3.35	7.51	26.9
RAFT + TC	1,2	H	2.06	3.19	**6.41**	22.6
		R	**2.05**	**3.15**	6.50	**22.5**
RAFT + TC	1,2	R	**2.05**	3.15	**6.50**	**22.5**
	1,2,3		**2.05**	**3.14**	6.69	22.6

Table 4. Combination of Transformation Consistency with Occlusion Consistency (Other notations are the same as Table 2)

Method (small)	Sintel (train-EPE)		KITTI-15 (train)	
	Clean	Final	Fl-epe	Fl-all
RAFT (baseline)	2.21	3.35	7.51	26.9
RAFT + OC	2.05	3.18	7.07	23.5
RAFT + TC	2.05	3.15	**6.50**	22.5
RAFT + OC + TC	**1.95**	**3.13**	6.53	**22.1**

Transformation Consistency: We use horizontal flips and random rotations in our transformation consistency strategy, and we evaluate the performance in each type of these transformations[7]. As shown in Table 3, the two types of transformations show comparable accuracy gains, although rotation works better empirically in Sintel. Such interesting observations could be attributed to the characteristics of data samples. For example, KITTI image samples are typically dominated by downwards pixel movements in the driving scenes while being quite balanced between rightwards and leftwards movements. This could suggest a strategy to whether apply symmetrical generalization in vertical and horizontal directions. In our supplemental materials, we provide some distribution curves on several datasets.

We also experiment with a range of k values. Within certain k ranges, both Sintel and KITTI samples produce noticeable improvements. It is interesting, however, that Sintel and KITTI empirically demonstrate somewhat different upper bounds for their most suitable k ranges, which could be, again, attributed to the data sample characteristics in flow distributions in vertical and horizontal directions. Systematic analysis may provide more insights into ways of accuracy improvements.

Combining Consistency Strategies: In Table 4, both consistency strategies show performance improvements over the baseline model (RAFT-small). And, applying both methods shows better performance. Our conjecture is that the impact of each strategy is enhanced, and generalizability is improved with joint learning.

[7]Some of the transformation methods could potentially improve the performance. Note that rotations (90°, 180°, and 270°) and horizontal flips guarantee one-to-one correspondences

Table 5. Comparisons against the RAFT baseline in accuracy, model size, and inference time on KITTI after 24 GRU iterations.

Model	KITTI		# of Parameters	Inference Time
	Fl-epe	Fl-all		
RAFT (small)	7.51	26.9	990,162	99.03 ms
RAFT + OCTC (small)	6.53	22.1	997,043	101.53 ms
RAFT	5.04	17.4	5,257,365	140.18 ms
RAFT + OCTC	4.72	16.3	5,263,803	143.21 ms

Transformation Restoration: We considered inverting not only the displacement quantities but also the signs and axes when restoring coordinates back from transformation. For example, in restoring the 90° rotation, we computed the inverse of the pixel location and changed the signs and flow vector axes.

Model Size and Speed: We measure the average inference times with KITTI dataset using Nvidia V100DX-8C GPU. Our models significantly outperform the baseline RAFT at only minimal model overhead as detailed in Table 5. To support transformation consistency, there is no model size increase. Occlusion consistency entails minor model size increases by only 0.12% and 0.69% on large and small models, respectively, for mask derivation, which also has a minimal impact on inference time. Besides, during training, our model computes the baseline and transformation outputs sequentially without needing extra memory.

Limitations: Our algorithm could be further improved to work for very large areas of occlusions. Besides, we currently use only self-supervised learning in our occlusion training with sample pairs created from individual images (I_t, $I_{t,occ}$). Furthermore, we speculate that it could be challenging to predict accurate optical flows in certain low-frequency regions, where boundaries may be hidden due to occlusion. This problem could be investigated using an occlusion generating network with labeled data.

Another area of further research for improvement could be an analysis on the frame rate. Beyond our methods of consistency, zero forcing, and frame hopping, aspects such as temporal consistency could be investigated.

6. Conclusion

In this paper, we have introduced novel and effective consistency learning strategies, promoting occlusion consistency and transformation consistency, for optical flow estimation. We further introduce enhancements, zero forcing as a special case of occlusion consistency and frame hopping as a generalization to transformation consistency, to our overall consistency learning framework. Applying these methods jointly, we demonstrate empirical outperformance over the baselines. Specifically, our method sets the new state-of-the-art performance and has ranked top in the KITTI-2015 scene flow non-stereo leaderboards. We intend to adapt our framework to wider tasks in our future study.

References

[1] Anonymous. Cross-attentional flow transformer. http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=flow, 2021. 6

[2] Aviram Bar-Haim and Lior Wolf. Scopeflow: Dynamic scene scoping for optical flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7998–8007, 2020. 5, 6

[3] Kireeti Bodduna and Joachim Weickert. Removing multiframe gaussian noise by combining patch-based filters with optical flow. *Journal of Electronic Imaging*, 30(3):033031, 2021. 1

[4] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *European conference on computer vision*, pages 25–36. Springer, 2004. 2

[5] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pages 611–625. Springer, 2012. 4, 5

[6] Zixi Cai, Helmut Neher, Kanav Vats, David A Clausi, and John Zelek. Temporal hockey action recognition via pose and optical flows. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[7] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 2

[8] Valéry Dewil, Jérémy Anger, Axel Davy, Thibaud Ehret, Gabriele Facciolo, and Pablo Arias. Self-supervised training for blind multi-frame video denoising. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2724–2734, 2021. 1

[9] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 1, 2, 5

[10] Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. *arXiv preprint arXiv:1906.01916*, 2019. 4

[11] Geoff French, Avital Oliver, and Tim Salimans. Milking cowmask for semi-supervised image classification. *arXiv preprint arXiv:2003.12022*, 2020. 4, 6

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 4, 5

[13] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 3

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[15] Berthold KP Horn and Brian G Schunck. Determining optical flow. In *Techniques and Applications of Image Understanding*, volume 281, pages 319–331. International Society for Optics and Photonics, 1981. 2

[16] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8981–8989, 2018. 6

[17] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. A lightweight optical flow cnn-revisiting data fidelity and regularization. *arXiv preprint arXiv:1903.07414*, 2019. 6

[18] Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5754–5763, 2019. 2

[19] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 1, 2, 6

[20] Joel Janai, Fatma Guney, Anurag Ranjan, Michael Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 690–706, 2018. 1

[21] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *Advances in Neural Information Processing Systems*, pages 10758–10767, 2019. 2, 3

[22] Jisoo Jeong, Vikas Verma, Minsung Hyun, Juho Kannala, and Nojun Kwak. Interpolation-based semi-supervised learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11602–11611, 2021. 2, 3

[23] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. *arXiv preprint arXiv:2104.02409*, 2021. 6, 7

[24] Rico Jonschkowski, Austin Stone, Jonathan T Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova. What matters in unsupervised optical flow. *arXiv preprint arXiv:2006.04902*, 2020. 1, 2

[25] Kiran Kale, Sushant Pawar, and Pravin Dhulekar. Moving object tracking using optical flow and motion vector estimation. In *2015 4th international conference on reliability, infocom technologies and optimization (ICRITO)(trends and future directions)*, pages 1–6. IEEE, 2015. 1

[26] Daniel Kondermann, Rahul Nair, Katrin Honauer, Karsten Krispin, Jonas Andrulis, Alexander Brock, Burkhard Gussefeld, Mohsen Rahimimoghaddam, Sabine Hofmann, Claus Brenner, 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 Vi-*

sion and Pattern Recognition Workshops, pages 19–28, 2016. 5

[27] Wei-Sheng Lai, Jia-Bin Huang, and Ming-Hsuan Yang. Semi-supervised learning for optical flow with generative adversarial networks. In *Advances in neural information processing systems*, pages 354–364, 2017. 1, 2, 3

[28] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 2, 3

[29] Myunggi Lee, Seungeui Lee, Sungjoon Son, Gyutae Park, and Nojun Kwak. Motion feature network: Fixed motion filter for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 387–403, 2018. 1

[30] Pengpeng Liu, Irwin King, Michael R Lyu, and Jia Xu. Ddflow: Learning optical flow with unlabeled data distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8770–8777, 2019. 2

[31] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selflow: Self-supervised learning of optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4571–4580, 2019. 1, 2, 6

[32] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. Dvc: An end-to-end deep video compression framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11006–11015, 2019. 1

[33] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 5

[34] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. *arXiv preprint arXiv:1711.07837*, 2017. 1, 2

[35] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015. 5

[36] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in neural information processing systems*, pages 3235–3246, 2018. 2, 3

[37] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4161–4170, 2017. 2

[38] Deqing Sun, Daniel Vlasic, Charles Herrmann, Varun Jampani, Michael Krainin, Huiwen Chang, Ramin Zabih, William T Freeman, and Ce Liu. Autoflow: Learning a better training set for optical flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10093–10102, 2021. 6, 7

[39] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 2, 6

[40] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. 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. 1, 6

[41] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *European conference on computer vision*, pages 438–451. Springer, 2010. 1, 2

[42] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017. 2, 3

[43] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pages 402–419. Springer, 2020. 1, 2, 3, 6

[44] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019. 2

[45] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. Occlusion aware unsupervised learning of optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4884–4893, 2018. 1, 2

[46] Chao-Yuan Wu, Nayan Singhal, and Philipp Krahenbuhl. Video compression through image interpolation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 416–431, 2018. 1

[47] Wending Yan, Aashish Sharma, and Robby T Tan. Optical flow in dense foggy scenes using semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13259–13268, 2020. 1, 3

[48] Gengshan Yang and Deva Ramanan. Volumetric correspondence networks for optical flow. In *Advances in neural information processing systems*, pages 794–805, 2019. 6

[49] Zhichao Yin, Trevor Darrell, and Fisher Yu. Hierarchical discrete distribution decomposition for match density estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6044–6053, 2019. 6

[50] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1476–1485, 2019. 3

[51] Shengyu Zhao, Yilun Sheng, Yue Dong, Eric I Chang, Yan Xu, et al. Maskflownet: Asymmetric feature matching with learnable occlusion mask. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6278–6287, 2020. 1, 2, 6

[52] Huizhong Zhou, Benjamin Ummenhofer, and Thomas Brox. Deeptam: Deep tracking and mapping. In *Proceedings of*

the European conference on computer vision (ECCV), pages
822–838, 2018. 1