Flow2Stereo: Effective Self-Supervised Learning of Optical Flow and Stereo Matching

Pengpeng Liu[†*] Irwin King[†] Michael Lyu[†] Jia Xu[§]
[†] The Chinese University of Hong Kong [§] Huya AI

Abstract

In this paper, we propose a unified method to jointly learn optical flow and stereo matching. Our first intuition is stereo matching can be modeled as a special case of optical flow, and we can leverage 3D geometry behind stereoscopic videos to guide the learning of these two forms of correspondences. We then enroll this knowledge into the state-of-the-art self-supervised learning framework, and train one single network to estimate both flow and stereo. Second, we unveil the bottlenecks in prior self-supervised learning approaches, and propose to create a new set of challenging proxy tasks to boost performance. These two insights yield a single model that achieves the highest accuracy among all existing unsupervised flow and stereo methods on KITTI 2012 and 2015 benchmarks. More remarkably, our self-supervised method even outperforms several state-of-the-art fully supervised methods, including PWC-Net and FlowNet2 on KITTI 2012.

1. Introduction

Estimating optical flow and stereo matching are two fundamental computer vision tasks with a wide range of applications [6, 31]. Despite impressive progress in the past decades, accurate flow and stereo estimation remain a long-standing challenge. Traditional stereo matching estimation approaches often employ different pipelines compared with prior flow estimation methods [13, 2, 36, 19, 40, 37, 12, 11, 7]. These methods merely share common modules, and they are computationally expensive.

Recent CNN-based methods directly estimate optical flow [4, 15, 32, 39, 14] or stereo matching [20, 3] from two raw images, achieving high accuracy with real-time speed. However, these fully supervised methods require a large amount of labeled data to obtain state-of-the-art performance. Moreover, CNNs for flow estimation are drastically different from those for stereo estimation in terms of network architecture and training data [4, 28].

*Work mainly done during an internship at Huya AI.

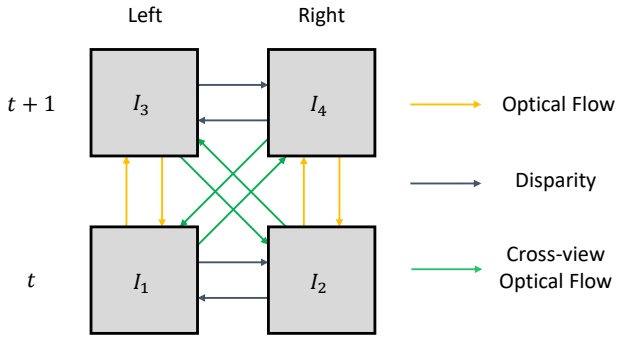


Figure 1. Illustration of 12 cross-view correspondnce maps among 4 stereoscopic frames. We leverage all these geometric consistency constraints, and train one single network to estimate both flow and stereo.

Is it possible to train one single network to estimate both flow and stereo using only one set of data, even unlabeled? In this paper, we show conventional self-supervised methods can learn to estimate these two forms of dense correspondences with one single model, when fully utilizing stereoscopic videos with inherent geometric constraints.

Fig. 2 shows the geometric relationship between stereo disparity and optical flow. We consider stereo matching as a special case of optical flow, and compute 12 cross-view correspondence maps between images captured at different time and different view (Fig. 1. This enables us to train one single network with a set of photometric and geometric consistency constraints. Besides, after digging into conventional two-stage self-supervised learning framework [25, 26], we show that creating challenging proxy tasks is the key for performance improvement. Based on this observation, we propose to employ additional challenging conditions to further boost the performance.

These two insights yield a method outperforming all existing unsupervised flow learning methods by a large margin, with Fl-noc = 4.02% on KITTI 2012 and Fl-all = 11.10% on KITTI 2015. Remarkably, our self-supervised method even outperforms several state-of-the-art fully supervised methods, including PWC-Net [39], FlowNet2 [15], and MFF [34] on KITTI 2012. More impor-

tantly, when we directly estimate stereo matching with our optical flow model, it also achieves state-of-the-art unsupervised stereo matching performance. This further demonstrates the strong generalization capability of our approach.

2. Related Work

Optical flow and stereo matching have been widely studied in the past decades [13, 2, 36, 38, 45, 48, 49, 29, 21, 11]. Here, we briefly review recent deep learning based methods.

Supervised Flow Methods. FlowNet [4] is the first end-to-end optical learning method, which takes two raw images as input and output a dense flow map. The followup FlowNet 2.0 [15] stacks several basic FlowNet models and refines the flow iteratively, which significantly improves accuracy. SpyNet [32], PWC-Net [39] and LiteFlowNet [14] propose to warp CNN features instead of image at different scales and introduce cost volume construction, achieving state-of-the-art performance with a compact model size. However, these supervised methods rely on pre-training on synthetic datasets due to lacking of real-world ground truth optical flow. The very recent SelFlow [26] employs self-supervised pre-training with real-world unlabeled data before fine-tuning, reducing the reliance of synthetic datasets. In this paper, we propose an unsupervised method, and achieve comparable performance with supervised learning methods without using any labeled data.

Unsupervised & Self-Supervised Flow Methods. Labeling optical flow for real-world images is a challenging task, and recent studies turn to formulate optical flow estimation as an unsupervised learning problem based on the brightness constancy and spatial smoothness assumption [17, 35]. [30, 44, 16] propose to detect occlusion and exclude occluded pixels when computing photometric loss. Despite promising progress, they still lack the ability to learn optical flow of occluded pixels.

Our work is most similar to DDFlow [25] and SelFlow [26], which employ a two-stage self-supervision strategy to cope with optical flow of occluded pixels. In this paper, we extend the scope to utilize geometric constraints in stereoscopic videos and jointly learn optical flow and stereo disparity. This turns out to be very effective, as our method significantly improves the quality of flow prediction in the first stage. Recent works also propose to jointly learn flow and depth from monocular videos [52, 53, 47, 33, 24] or jointly learn flow and disparity from stereoscopic videos [22, 43]. Unlike these methods, we make full use of geometric constraints between optical flow and stereo matching in a self-supervised learning manner, and achieve much better performance.

Unsupervised & Self-Supervised Stereo Methods. Our method is also related to a large body of unsupervised stereo learning methods, including image synthesis and warping with depth estimation [5], left-right consistency [8, 51, 9], employing additional semantic information *et al*. [46], cooperative learning [23], self-adaptive fine-tuning *et al*. [41, 50, 42]. Different from all these methods that design a specific network for stereo estimation, we train one single unified network to estimate both flow and stereo.

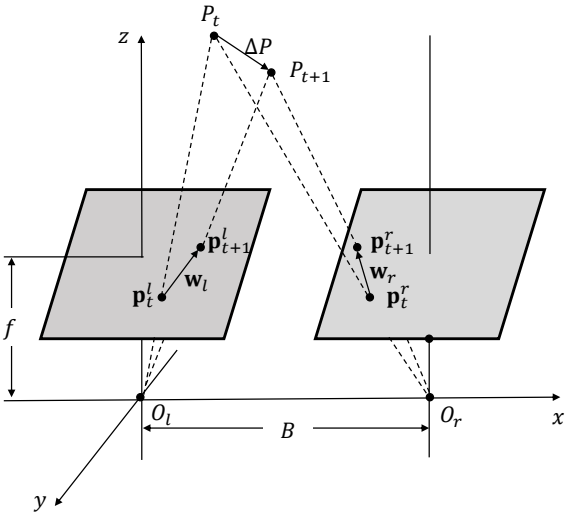


Figure 2. 3D geometric constraints between optical flow ($\mathbf{w}_l$ and $\mathbf{w}_r$) and stereo disparity from time t to $t+1$ in the 3D projection view.

3. Geometric Relationship of Flow and Stereo

In this section, we review the geometric relationship between optical flow and stereo disparity from both the 3D projection view [10] and the motion view.

3.1. Geometric Relationship in 3D Projection

Fig. 2 illustrates the geometric relationship between stereo disparity and optical flow from a 3D projection view. O_l and O_r are rectified left and right camera centers, B is the baseline distance between two camera centers.

Suppose $P(X, Y, Z)$ is a 3D point at time t, and it moves to $P + \Delta P$ at time $t+1$, resulting in the displacement as $\Delta P = (\Delta X, \Delta Y, \Delta Z)$. Denote f as the focal length, $p = (x, y)$ as the projection point of P on the image plane, then $(x, y) = \frac{f}{s}\frac{(X,Y)}{Z}$, where s is the scale factor that converts the world space to the pixel space, *i.e.*, how many meters per pixel. For simplicity, let $f' = \frac{f}{s}$, we have $(x, y) = f'\frac{(X,Y)}{Z}$. Take the time derivative, we obtain

$$\frac{(\Delta x, \Delta y)}{\Delta t} = f'\frac{1}{Z}\frac{(\Delta X, \Delta Y)}{\Delta t} - f'\frac{(X, Y)}{Z^2}\frac{\Delta Z}{\Delta t} \quad (1)$$

Let $\mathbf{w} = (u, v)$ be the optical flow vector (u denotes motion in the x direction and v denotes motion in the y direction) and the time step is one unit (from t to $t+1$), then Eq. (1)

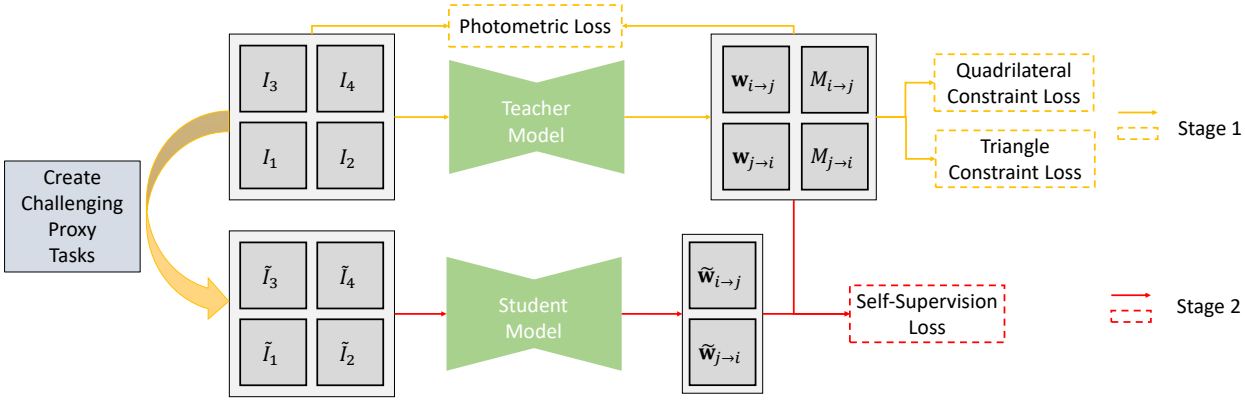


Figure 3. Our self-supervised learning framework contains two stages: In stage 1, we add geometric constraints between optical flow and stereo disparity to improve the quality of confident predictions; In stage 2, we create challenging proxy tasks to guide the student model for effective self-supervised learning.

becomes,

$$(u, v) = f' \frac{(\Delta X, \Delta Y)}{Z} - f' \frac{\Delta Z}{Z^2}(X, Y) \tag{2}$$

For calibrated stereo cameras, we let P in the coordinate system of O_l. Then $P_l = P = (X, Y, Z)$ in the coordinate system of O_l and $P_r = (X - B, Y, Z)$ in the coordinate system of O_r. With Eq. (2), we obtain,

$$\begin{cases} (u_l, v_l) = f' \frac{(\Delta X, \Delta Y)}{Z} - f' \frac{\Delta Z}{Z^2}(X, Y) \\ (u_r, v_r) = f' \frac{(\Delta X, \Delta Y)}{Z} - f' \frac{\Delta Z}{Z^2}(X - B, Y) \end{cases}, \tag{3}$$

This can be further simplified as,

$$\begin{cases} u_r - u_l = f' B \frac{\Delta Z}{Z^2} \\ v_r - v_l = 0 \end{cases}, \tag{4}$$

Suppose d is the stereo disparity ($d \geq 0$), according to the depth Z and the distance between two camera centers B, we have $d = f' \frac{B}{Z}$. Take the time derivative, we have

$$\frac{\Delta d}{\Delta t} = -f' \frac{B}{Z^2} \frac{\Delta Z}{\Delta t} \tag{5}$$

Similarly, we set time step to be one unit, then

$$d_{t+1} - d_t = -f' B \frac{\Delta Z}{Z^2} \tag{6}$$

With Eq. (4) and (6), we finally obtain,

$$\begin{cases} u_r - u_l = (-d_{t+1}) - (-d_t) \\ v_r - v_l = 0 \end{cases}. \tag{7}$$

Eq. (7) demonstrates the 3D geometric relationship between optical flow and stereo disparity, *i.e.*, the difference between optical flow from left and right view is equal to the difference between disparity from time t to $t + 1$. Note that Eq. (7) also works when cameras move, including rotating and translating from t to $t + 1$. Eq. (7) assumes the focal length is fixed, which is common for stereo cameras.

Next, we review the geometric relationship between flow and stereo in the motion view.

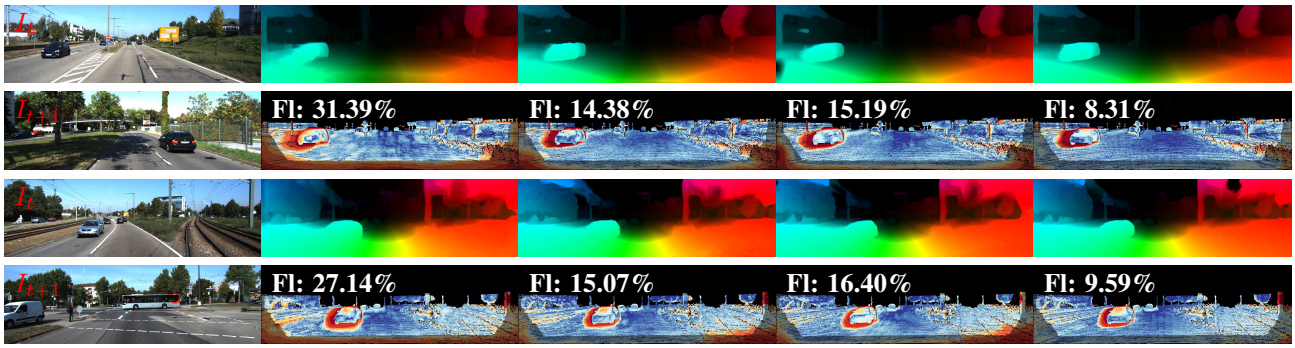
3.2. Geometric Relationship in Motion

Optical flow estimation and stereo matching can be viewed as one single problem: correspondence matching. Optical flow describes the pixel motion between two adjacent frames recorded at different time, while stereo disparity represents the pixel displacement between two stereo images recorded at the same time. According to stereo geometry, the correspondence pixel shall lie on the epipolar line between stereo images. However, optical flow does not have such a constraint, this is because both camera and object can move at different times.

To this end, we consider stereo matching as a special case of optical flow. That is, the displacement between stereo images can be seen as a one dimensional movement. For rectified stereo image pairs, the epipolar line is horizontal and stereo matching becomes finding the correspondence pixel along the horizontal direction x.

In the following, we consider stereo disparity as a form of motion between stereo image pairs. For simplicity, let I_1, I_3 denote the left-view images at time t and $t + 1$, I_2, I_4 denote the right-view images at time t and $t+1$ respectively. Then we let $\mathbf{w}_{1\to3}$ denote the optical flow from I_1 to I_3, $\mathbf{w}_{1\to2}$ denote the stereo disparity from I_1 to I_2. For stereo disparity, we only keep the horizontal direction of optical flow. For optical flow and disparity of other directions, we denote them in the same way.

Apart from optical flow in the left and right view, disparity at time t and $t + 1$, we also compute the cross-view optical flow between images captured at different time and different view, such as $\mathbf{w}_{1\to4}$ (green row in Fig. 1). In this case,



(a) Input Images (b) MFO-Flow [16] (c) DDFlow [25] (d) SelFlow [26] (e) Ours

Figure 4. Qualitative evaluation on KITTI 2015 optical flow benchmark. For each case, the top row is optical flow and the bottom row is error map. Our model achieves much better results both quantitatively and qualitatively (*e.g.*, shaded boundary regions). Lower Fl is better.

we compute the correspondence between every two images, resulting in 12 optical flow maps as shown in Fig. 1. We employ the same model to compute optical flow between every two images.

Suppose $\mathbf{p}_t^l$ is a pixel in I_1, $\mathbf{p}_t^r, \mathbf{p}_{t+1}^l, \mathbf{p}_{t+1}^r$ are its correspondence pixels in I_2, I_3 and I_4 respectively, then we have,

$$\begin{cases} \mathbf{p}_t^r = \mathbf{p}_t^l + \mathbf{w}_{1\to2}(\mathbf{p}_t^l) \\ \mathbf{p}_{t+1}^l = \mathbf{p}_t^l + \mathbf{w}_{1\to3}(\mathbf{p}_t^l) \\ \mathbf{p}_{t+1}^r = \mathbf{p}_t^l + \mathbf{w}_{1\to4}(\mathbf{p}_t^l) \end{cases}. \tag{8}$$

A pixel directly moves from I_1 to I_4 shall be identical to the movement from I_1 to I_2 and from I_2 to I_4. That is,

$$\begin{aligned} \mathbf{w}_{1\to4}(\mathbf{p}_t^l) &= (\mathbf{p}_{t+1}^r - \mathbf{p}_t^r) + (\mathbf{p}_t^r - \mathbf{p}_t^l) \\ &= \mathbf{w}_{2\to4}(\mathbf{p}_t^r) + \mathbf{w}_{1\to2}(\mathbf{p}_t^l). \end{aligned} \tag{9}$$

Similarly, if the pixel moves from I_1 to I_3 and from I_3 to I_4, then

$$\begin{aligned} \mathbf{w}_{1\to4}(\mathbf{p}_t^l) &= (\mathbf{p}_{t+1}^r - \mathbf{p}_{t+1}^l) + (\mathbf{p}_{t+1}^l - \mathbf{p}_t^l) \\ &= \mathbf{w}_{3\to4}(\mathbf{p}_{t+1}^l) + \mathbf{w}_{1\to3}(\mathbf{p}_t^l). \end{aligned} \tag{10}$$

From Eq. (9) and (10), we obtain,

$$\mathbf{w}_{2\to4}(\mathbf{p}_t^r) - \mathbf{w}_{1\to3}(\mathbf{p}_t^l) = \mathbf{w}_{3\to4}(\mathbf{p}_{t+1}^l) - \mathbf{w}_{1\to2}(\mathbf{p}_t^l). \tag{11}$$

For stereo matching, the correspondence pixel shall lie on the epipolar lines. Here, we only consider rectified stereo cases, where epipolar lines are horizontal. Then, Eq.(11) becomes

$$\begin{cases} u_{2\to4}(\mathbf{p}_t^r) - u_{1\to3}(\mathbf{p}_t^l) = u_{3\to4}(\mathbf{p}_{t+1}^l) - u_{1\to2}(\mathbf{p}_t^l) \\ v_{2\to4}(\mathbf{p}_t^r) - v_{1\to3}(\mathbf{p}_t^l) = 0 \end{cases}. \tag{12}$$

Note Eq. (12) is exactly the same as Eq. (7).

In addition, since epipolar lines are horizontal, we can re-write Eq. (9) and (10) as follows:

$$\begin{cases} u_{1\to4}(\mathbf{p}_t^l) = u_{2\to4}(\mathbf{p}_t^r) + u_{1\to2}(\mathbf{p}_t^l) \\ v_{1\to4}(\mathbf{p}_t^l) = v_{2\to4}(\mathbf{p}_t^r) \\ u_{1\to4}(\mathbf{p}_t^l) = u_{3\to4}(\mathbf{p}_{t+1}^l) + u_{1\to3}(\mathbf{p}_t^l) \\ v_{1\to4}(\mathbf{p}_t^l) = v_{1\to3}(\mathbf{p}_t^l) \end{cases}. \tag{13}$$

This leads to the two forms of geometric constraints we used in our training loss functions: quadrilateral constraint (12) and triangle constraint (13).

4. Method

In this section, we first dig into the bottlenecks of the state-of-the-art two-stage self-supervised learning framework [25, 26]. Then we describe an enhanced proxy learning approach, which can improve its performance greatly in both two stages.

4.1. Two-Stage Self-Supervised Learning Scheme

Both DDFlow [25] and SelFlow [26] employ a two-stage learning approaches to learning optical flow in a self-supervised manner. In the first stage, they train a teacher model to estimate optical flow for `non-occluded` pixels. In the second stage, they first pre-process the input images, *e.g.*, cropping and inject superpixel noises to create hand-crafted `occlusions`, then the predictions of teacher model for those `non-occluded` pixels are regarded as ground truth to guide a student model to learn optical flow of hand-crafted `occluded` pixels.

The general pipeline is reasonable, but the definition of `occlusion` is in a heuristic manner. At the first stage, forward-backward consistency is employed to detect whether the pixel is `occluded`. However, this brings in errors because many pixels are still `non-occluded` even they violate this principle, and vice versa. Instead, it would be more proper to call those pixels reliable or confident if they pass the forward-backward consistency check. From

Table 1. Quantitative evaluation of optical flow estimation on KITTI. Bold fonts highlight the best results among supervised and unsupervised methods. Parentheses mean that training and testing are performed on the same dataset. `fg` and `bg` denote results of foreground and background regions respectively.

	Method	Train	KITTI 2012						KITTI 2015				
			train		test				train		test		
		Stereo	EPE-all	EPE-noc	EPE-all	EPE-noc	Fl-all	Fl-noc	EPE-all	EPE-noc	Fl-all	Fl-fg	Fl-bg
Supervised	SpyNet [32]	✗	3.36	–	4.1	2.0	20.97%	12.31%	–	–	35.07%	43.62%	33.36%
	FlowFieldsCNN [1]	✗	–	–	3.0	1.2	13.01%	4.89%	–	–	18.68%	20.42%	18.33%
	DCFlow [45]	✗	–	–	–	–	–	–	–	–	14.86%	23.70%	13.10%
	FlowNet2 [15]	✗	(1.28)	–	1.8	1.0	8.80%	4.82%	(2.3)	–	10.41%	8.75%	10.75%
	UnFlow-CSS [30]	✗	(1.14)	(0.66)	1.7	0.9	8.42%	4.28%	(1.86)	–	11.11%	15.93%	10.15%
	LiteFlowNet [14]	✗	(1.05)	–	1.6	**0.8**	7.27%	**3.27%**	(1.62)	–	9.38%	7.99%	9.66%
	PWC-Net [39]	✗	(1.45)	–	1.7	0.9	8.10%	4.22%	(2.16)	–	9.60%	9.31%	9.66%
	MFF [34]	✗	–	–	1.7	0.9	7.87%	4.19%	–	–	**7.17%**	**7.25%**	**7.15%**
	SelFlow [26]	✗	**(0.76)**	–	**1.5**	0.9	**6.19%**	3.32%	**(1.18)**	–	8.42%	7.61%	12.48%
Unsupervised	BackToBasic [17]	✗	11.3	4.3	9.9	4.6	43.15%	34.85%	–	–	–	–	–
	DSTFlow [35]	✗	10.43	3.29	12.4	4.0	–	–	16.79	6.96	39%	–	–
	UnFlow-CSS [30]	✗	3.29	1.26	–	–	–	–	8.10	–	23.30%	–	–
	OccAwareFlow [44]	✗	3.55	–	4.2	–	–	–	8.88	–	31.2%	–	–
	MultiFrameOccFlow-None [16]	✗	–	–	–	–	–	–	6.65	3.24	–	–	–
	MultiFrameOccFlow-Soft [16]	✗	–	–	–	–	–	–	6.59	3.22	22.94%	–	–
	DDFlow [25]	✗	2.35	1.02	3.0	1.1	8.86%	4.57%	5.72	2.73	14.29%	20.40%	13.08%
	SelFlow [26]	✗	1.69	0.91	2.2	1.0	7.68%	4.31%	4.84	2.40	14.19%	21.74%	12.68%
	Lai *et al.* [22]	✓	2.56	1.39	–	–	–	–	7.134	4.306	–	–	–
	UnOS [43]	✓	1.64	1.04	1.8	–	–	–	5.58	–	18.00%	–	–
	Our+L_p+L_q+L_t	✓	4.91	0.84	–	–	–	–	7.88	2.24	–	–	–
	Ours+L_p+L_q+L_t+Self-Supervision	✓	**1.45**	**0.82**	**1.7**	**0.9**	**7.63%**	**4.02%**	**3.54**	**2.12**	**11.10%**	**16.67%**	**9.99%**

this point of view, creating hand-crafted `occlusions` can be regard as creating more challenging conditions, under which the prediction would be less confident. Then in the second stage, the key point is to let confident predictions to supervise those less confident predictions.

During the self-supervised learning stage, the student model is able to handle more challenging conditions. As a result, its performance improves not only for those occluded pixels, but also for non-occluded pixels. Because when creating challenging conditions, both occluded regions and non-occluded regions become more challenging. The reason why optical flow for occluded pixels improves more than non-occluded regions is that, during the first stage, photometric loss does not hold for occluded pixels, the model just does not have the ability to predict them. In the second stage, the model has the ability to learn optical flow of occluded pixels for the first time, therefore its performance improves a lot. To lift the upper bound of confident predictions, we propose to utilize stereoscopic videos to reveal their geometric nature.

4.2. Proxy Learning Scheme

Following [25, 26], our proxy learning scheme contains two stages and our network structure is built upon PWC-Net [39].

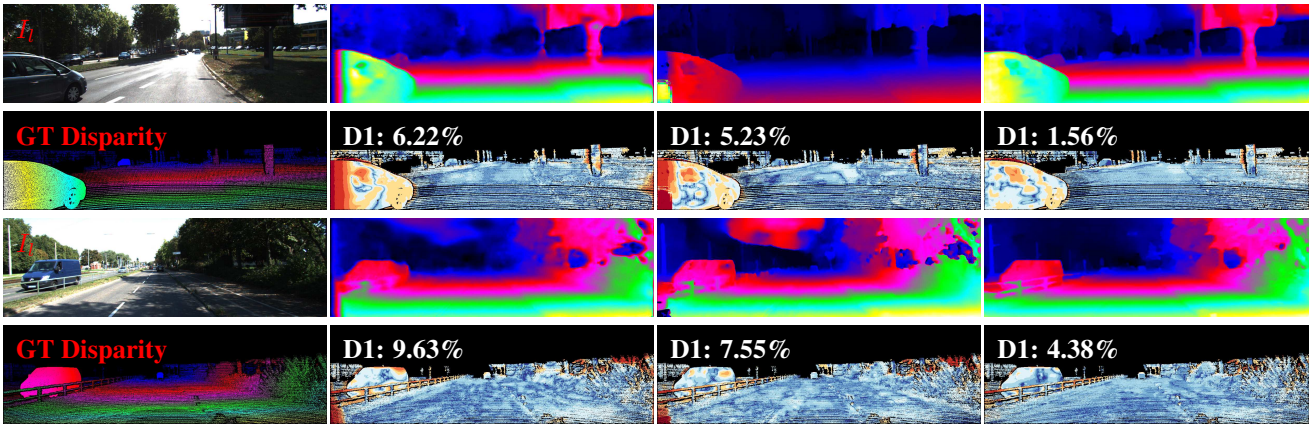
Stage 1: Predicting confident optical flow with geometric constraints. With the estimated optical flow map $\mathbf{w}_{i\to j}$, we warp the target image I_j toward the reference image I_i. Then we measure the difference between the warped target image $I^w_{j\to i}$ and the reference image I_i with a photometric loss. Similar to [30, 25, 26], we employ forward-backward consistency check to compute a confident map, where value 1 indicates the prediction is confident, 0 indicates the prediction is non-confident.

Apart from photometric loss, we also apply geometric constraints to our teacher model, including the triangle constraint and quadrilateral constraint. Note that geometric constraints are only applied to those confident pixels. This turns out to highly effective and greatly improves the accuracy of those confident predictions.

Stage 2: Self-supervised learning from teacher model to student model. As discussed earlier, the key point of self-supervision is to create challenging input-output pairs. In our framework, we create challenging conditions by random cropping input image pairs, injecting random noise into the second image, random scale (down-sample) the input image pairs, to make correspondence learning more difficult. These hard input-output pairs push the network to capture more information, resulting in a large performance gain in practice.

Different from [25, 26], we do not distinguish between occluded and non-occluded pixels anymore in the self-supervision stage. As forward-backward consistency check cannot perfectly determine whether a pixel is occluded, there may be many erroneous judgments. In this case, the confident prediction from teacher model will provide guidance for both occluded or non-occluded pixels no matter forward-backward check is employed or not.

Next, we describe our training losses for each stage.



(a) I_l and GT Disparity (b) Godard *et al*. [8] (c) SeqStereo [46] (d) Ours

Figure 5. Qualitative evaluation with other unsupervised stereo matching methods on KITTI 2015 training dataset. For each case, the top row is stereo disparity and the bottom row is error map. Our models estimate more accurate disparity maps (*e.g*., image boundary regions and moving-object boundary regions). Lower D1 is better.

Table 2. Quantitative evaluation of stereo disparity on KITTI training datasets (apart from the last columns). Our single model achieves the highest accuracy among all unsupervised stereo learning methods. $\star$ denotes that we use their pre-trained model to compute the numbers, while other numbers are from their paper. Note that Guo *et al.* [9] pre-train stereo model on synthetic Scene Flow dataset with ground truth disparity before fine-tuning on KITTI dataset.

Method	KITTI 2012						KITTI 2015					
	EPE-all	EPE-noc	EPE-occ	D1-all	D1-noc	D1-all (test)	EPE-all	EPE-noc	EPE-occ	D1-all	D1-noc	D1-all (test)
Joung *et al.* [18]	–	–	–	–	–	13.88%	–	–	–	13.92%	–	–
Godard *et al.* [8] *	2.12	1.44	30.91	10.41%	8.33%	–	1.96	1.53	24.66	10.86%	9.22%	–
Zhou *et al.* [51]	–	–	–	–	–	–	–	–	–	9.41%	8.35%	–
OASM-Net [23]	–	–	–	8.79%	6.69%	8.60%	–	–	–	–	–	8.98%
SeqStereo *et al.* [46] *	2.37	1.63	33.62	9.64%	7.89%	–	1.84	1.46	26.07	8.79%	7.7%	–
Liu *et al.* [24] *	1.78	1.68	6.25	11.57%	10.61%	–	1.52	1.48	4.23	9.57%	9.10%	–
Guo *et al.* [9] *	1.16	1.09	4.14	6.45%	5.82%	–	1.71	1.67	4.06	7.06%	6.75%	–
UnOS [43]	–	–	–	–	–	5.93%	–	–	–	**5.94%**	–	6.67%
Ours+L_p	1.73	1.13	27.03	7.88%	5.87%	–	1.79	1.40	25.24	9.83%	7.74%	–
Ours+L_p+L_q+L_t	1.62	0.94	29.26	6.69%	4.69%	–	1.67	**1.31**	19.55	8.62%	7.15%	–
Ours+L_p+L_q+L_t+Self-Supervision	**1.01**	**0.93**	**4.52**	**5.14%**	**4.59%**	**5.11%**	**1.34**	**1.31**	**2.56**	6.13%	**5.93%**	**6.61%**

4.3. Loss Functions

For stage 1, our loss function mainly contains three parts: photometric loss L_p, triangle constraint loss L_t and quadrilateral constraint loss L_q. For stage 2, we only apply self-supervision loss L_s.

Photometric loss. Photometric loss is based on the brightness consistency assumption, which only works for non-occluded pixels. During our experiments, we employ census transform, which has shown to be robust for illumination change [30, 25, 26]. Denote $M_{i\to j}$ as the confident map from I_i to I_j is $M_{i\to j}$, then L_p is defined as,

$$L_p = \sum_{i,j} \frac{\sum_{\mathbf{p}} \psi(I_i(\mathbf{p}) - I^w_{j\to i}(\mathbf{p})) \odot M_{i\to j}(\mathbf{p})}{\sum_{\mathbf{p}} M_{i\to j}(\mathbf{p})}, \quad (14)$$

where $\psi(x) = (|x| + \epsilon)^q$. During our experiments, we set $\epsilon = 0.01$ and $q = 0.4$.

Quadrilateral constraint loss. Quadrilateral constraint describes the geometric relationship between optical flow and stereo disparity. Here, we only employ L_q to those confident pixels. Take $\mathbf{w}_{1\to4}$, $\mathbf{w}_{2\to4}$, $\mathbf{w}_{1\to2}$ and $\mathbf{w}_{3\to4}$ for an example, we first compute the confident map for quadrilateral constraint $M_q(\mathbf{p}) = M_{1\to2}(\mathbf{p}) \odot M_{1\to3}(\mathbf{p}) \odot M_{1\to4}(\mathbf{p})$. Then according to Eq. (12), we divide L_q into two components on the x direction L_{qu} and y direction L_{qv} respectively:

$$L_{qu} = \sum_{\mathbf{p}^l_t} \psi(u_{1\to2}(\mathbf{p}^l_t) + u_{2\to4}(\mathbf{p}^r_t) - u_{1\to3}(\mathbf{p}^l_t) - u_{3\to4}(\mathbf{p}^l_{t+1})) \odot M_q(\mathbf{p}^l_t) / \sum_{\mathbf{p}^l_t} M_q(\mathbf{p}^l_t) \quad (15)$$

$$L_{qv} = \sum_{\mathbf{p}^l_t} \psi(v_{2\to4}(\mathbf{p}^r_t) - v_{1\to3}(\mathbf{p}^l_t)) \odot M_q(\mathbf{p}^l_t) / \sum_{\mathbf{p}^l_t} M_q(\mathbf{p}^l_t) \quad (16)$$

where $L_q = L_{qu} + L_{qv}$. Quadrilateral constraint loss at other directions are computed in the same way.

Triangle constraint loss. Triangle constraint describes the relationship between optical flow, stereo disparity and

cross-view optical flow. Similar to quadrilateral constraint loss, we only employ L_t to confident pixels. Take $\mathbf{w}_{1\to3}$, $\mathbf{w}_{2\to4}$, $\mathbf{w}_{1\to2}$ as an example, we first compute the confident map for triangle constraint $M_t(\mathbf{p}) = M_{1\to2}(\mathbf{p}) \odot M_{1\to4}(\mathbf{p})$, then according to Eq. (9), L_t is defined as follows,

$$L_{tu} = \frac{\sum_{\mathbf{p}_t^l} \psi(u_{1\to4}(\mathbf{p}_t^l) - u_{2\to4}(\mathbf{p}_t^r) - u_{1\to2}(\mathbf{p}_t^l)) \odot M_t(\mathbf{p})}{\sum_{\mathbf{p}_t^l} M_t(\mathbf{p}_t^l)} \tag{17}$$

$$L_{tv} = \sum_{\mathbf{p}_t^l} \psi(v_{1\to4}(\mathbf{p}_t^l) - v_{2\to4}(\mathbf{p}_t^r)) \odot M_t(\mathbf{p}) \sum_{\mathbf{p}_t^l} M_t(\mathbf{p}_t^l), \tag{18}$$

where $L_t = L_{tu} + L_{tv}$. Triangle constraint losses at other directions are computed in the same way.

The final loss function for teacher model is $L = L_p + \lambda_1 L_q + \lambda_2 L_t$, where we set $\lambda_1 = 0.1$ and λ_2 = 0.2 during experiments.

Self-Supervision loss. During the first stage, we train our teacher model to compute proxy optical flow $\mathbf{w}$ and confident map M, then we define our self-supervision loss as,

$$L_s = \sum_{i,j} \frac{\sum_{\mathbf{p}} \psi(\mathbf{w}_{i\to j}(\mathbf{p}) - \widetilde{\mathbf{w}}_{i\to j}(\mathbf{p})) \odot M_{i\to j}(\mathbf{p})}{\sum_{\mathbf{p}} M_{i\to j}(\mathbf{p})}. \tag{19}$$

At test time, only the student model is needed, and we can use it to estimate both optical flow and stereo disparity.

5. Experiments

We evaluate our method on the challenging KITTI 2012 and KITTI 2015 datasets and compare our method with state-of-the-art unsupervised and supervised optical flow learning methods. Besides, since our method is able to predict stereo disparity, we also compare its stereo matching performance with related methods.

5.1. Experimental Setting

During training, we use the raw multi-view extensions of KITTI 2012 [6] and KITTI 2015 [31] and exclude neighboring frames (frame 9-12) as [35, 44, 25, 26]. For evaluation, we use the training sets of KITTI 2012 and KITTI 2015 with ground truth optical flow and disparity. We also submit our results to optical flow and stereo matching benchmarks for comparison with current state-of-the-art methods.

We implement our algorithm using TensorFlow with Adam optimizer. For teacher model, we set batch size to be 1, since there are 12 optical flow estimations for the 4 images. For student model, batch size is 4. We adopt similar data augmentation strategy as [4]. During training, we random crop [320, 896] as input, while during testing, we resize images to resolution [384, 1280]. We employ a two-stage training procedure as [25, 26]. The key difference is that during the first stage, we add geometric constraints which enable our model to predict more accurate reliable predictions. Besides, during the second stage, we do not distinguish between occluded and non-occluded pixels, and set all our confident predictions as ground truth. For each experiment, we set initial learning rate to be 1e-4 and decay it by half every 50k iterations.

For evaluation metrics, we use the standard EPE (average end-point error) and Fl (percentage or erroneous pixels). A pixel is considered as correctly estimated if end-point error is <3 pixel or <5%. For stereo matching, there is another metric $D1$, which shares the same definition as Fl.

5.2. Main Results

Our method achieves the best unsupervised results for all evaluation metrics on both KITTI 2012 and KITTI 2015 datasets. More notably, our unsupervised results are even comparable with state-of-the-art supervised learning methods. Our approach bridges the performance gap between supervised learning and unsupervised learning methods for optical flow estimation.

Optical Flow. As shown in Tab. 1, our method outperforms all unsupervised learning method for all metrics on both KITTI 2012 and KITTI 2015 datasets. Specially, on KITTI 2012 dataset, we achieve EPE-all = 1.45 pixels, which achieves 14.2% relative improvement than previous best SelFLow [26]. For testing set, we achieve EPE = 1.7 pixels, resulting in 22.7% improvement. More notably, we achieve FL-all = 7.68% and Fl-noc = 4.02%, which is even better than state-of-the-art fully supervised learning methods including PWC-Net [39], MFF [34], and is highly competitive with LiteFlowNet [14] and SelFlow [26].

On KITTI 2015, the improvement is also impressive. For the training set, we achieve EPE-all = 3.54 pixels, resulting in 26.9% relative improvement than previous best method SelFlow. On the testing benchmark, we achieve Fl-all = 11.10%, which is not only better than previous best unsupervised learning methods by a large margin (21.8% relative improvement), but also competitive with state-of-the-art supervised learning methods. To the best of our knowledge, this is the first time that an unsupervised method achieves comparable performance compared with state-of-the-art fully supervised learning methods. Qualitative comparisons with other methods on KITTI 2015 optical flow benchmark are shown in Fig. 4.

Stereo Matching. We directly apply our optical flow model to stereo matching (only keeping the horizontal direction of flow), it achieves state-of-the-art unsupervised stereo matching performance as shown in Tab. 2. Specially, we reduce EPE-all from 1.61 pixels to 1.01 pixels on KITTI 2012 training dataset and from 1.71 pixels to 1.34 pixels on KITTI 2015 dataset.

Compared with previous state-of-the-art method UnOS [43], we reduce Fl-all from 5.93% to 5.11% on

Table 3. Ablation study on KITTI training datasets. For self-supervision, `v1` means employing self-supervision of [25, 26], `v2` means not distinguishing between occluded and non-occluded pixels, `v3` means adding more challenging conditions (our final model), and `v4` means adding geometric constraints in the self-supervision stage (slightly degrade the performance).

L_p	L_q	L_t	Self-Supervision				KITTI 2012					KITTI 2015				
			v1	v2	v3	v4	EPE-all	EPE-noc	EPE-occ	Fl-all	Fl-noc	EPE-all	EPE-noc	EPE-occ	Fl-all	Fl-noc
✓	✗	✗	✗	✗	✗	✗	4.41	1.06	26.54	14.18%	5.13%	8.20	2.85	42.01	19.50%	9.97%
✓	✓	✗	✗	✗	✗	✗	5.15	0.84	33.74	13.53%	3.42%	8.24	2.33	45.46	18.31%	8.15%
✓	✗	✓	✗	✗	✗	✗	4.98	0.86	32.33	12.64%	3.54%	7.99	2.34	43.50	17.89%	8.14%
✓	✓	✓	✗	✗	✗	✗	4.91	0.84	31.81	12.57%	3.47%	7.88	2.24	43.92	17.68%	7.97%
✓	✗	✗	✓	✗	✗	✗	1.92	0.95	7.86	6.56%	3.82%	5.85	2.96	24.17	13.26%	9.06%
✓	✗	✗	✗	✓	✗	✗	1.89	0.93	7.76	6.44%	3.76%	5.48	2.78	22.05	12.62%	8.53%
✓	✗	✗	✗	✗	✓	✗	1.62	0.89	6.21	5.62%	3.38%	4.12	2.36	15.04	10.93%	8.31%
✓	✓	✓	✗	✗	✓	✗	**1.45**	**0.82**	**5.52**	**5.29%**	**3.27%**	**3.54**	**2.12**	**12.58**	**10.04%**	**7.57%**
✓	✓	✓	✗	✗	✗	✓	1.56	0.86	6.20	5.83%	3.41%	3.66	2.16	13.18	10.44%	7.80%

KITTI 2012 testing dataset and from 6.67% to 6.61% on KITTI 2015 testing dataset. This is a surprisingly impressive result, since our optical flow model performs even better than other models specially designed for stereo matching. It also demonstrates the generalization capability of our optical flow model toward stereo matching. Qualitative comparisons with other unsupervised stereo matching approaches are shown in Fig. 5.

5.3. Ablation Study

We conduct a thorough analysis for different components of our proposed method.

Quadrilateral and Triangle Constraints. We add both constraints during our training in the first stage, aiming to improve the accuracy of confident pixel, since only these confident pixels are used for self-supervised training in the second stage. confident pixels are usually non-occluded in the first stage, because we optimize our model with photometric loss, which only holds for non-occluded pixels. Therefore, we are concerned about the performance over those non-occluded pixels (not for all pixels). As shown in the first 4 rows of Tab. 3, both constraints significantly improve the performance over those non-occluded pixels, and the combination of them produces the best results, while the `EPE-occ` may degrade. This is because we are concerned about the performance over those non-occluded pixels, since only confident pixels are used for self-supervised training. Specially, EPE-noc decreases from 1.06 pixels to 0.84 pixels on KITTI 2012 and from 2.85 pixels to 2.24 pixels on KITTI 2015. It is because that we achieve more accurate confident flow predictions, we are able to achieve much better results in the second self-supervision stage. We also achieve big improvement for stereo matching performance over non-occluded pixels as in Tab. 2.

Self-Supervision. We employ four types of self-supervision (check comparison of row 5, 6, 7, 8 in Tab. 3). For row 5 and row 6 (`v1` and `v2`), we show that it does not make much difference to distinguish occluded or non-occluded pixels denoted by forward-backward consistency check. Because forward-backward consistency predicts confident or non-confident flow predictions, but not occluded or non-occluded pixels. Therefore, the self-supervision will be employed to both occluded and non-occluded pixels whenever forward-backward check is employed. Comparing row 6 and row 7 (`v2` and `v3`), we show that after adding additional challenging conditions, flow estimation performance is improved greatly. Currently, we are not able to successfully apply geometric constraints in the self-supervision stage. As shown in row 7 and row 8 (`v2` and `v3`), geometric constraints will slightly degrade the performance. This is mainly because there is a correspondence ambiguity within occluded pixels, and it is challenging for our geometric consistency to hold for all occluded pixels.

6. Conclusion

We have presented a method to jointly learning optical flow and stereo matching with one single model. We show that geometric constraints improve the quality of those confident predictions, which further help in the self-supervision stage to achieve much better performance. Besides, after digging into the self-supervised learning approaches, we show that creating challenging conditions is the key to improve the performance. Our approach has achieved the best unsupervised optical flow performance on KITTI 2012 and KITTI 2015, and our unsupervised performance is comparable with state-of-the-art supervised learning methods. More notably, our unified model also achieves state-of-the-art unsupervised stereo matching performance, demonstrating the generalization capability of our model.

7. Acknowledgment.

This work was partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (RGC C5026-18GF and No. CUHK 14210717 of the General Research Fund).

References

[1] Christian Bailer, Kiran Varanasi, and Didier Stricker. Cnn-based patch matching for optical flow with thresholded hinge embedding loss. In *CVPR*, 2017.

[2] Thomas Brox and Jitendra Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *TPAMI*, 2011.

[3] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, 2018.

[4] 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 *ICCV*, 2015.

[5] Ravi Garg, BG Vijay Kumar, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016.

[6] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[7] Andreas Geiger, Martin Roser, and Raquel Urtasun. Efficient large-scale stereo matching. In *ACCV*, 2010.

[8] Clement Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

[9] Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. Learning monocular depth by distilling cross-domain stereo networks. In *ECCV*, 2018.

[10] Sang Hyun Han, Yan Sheng, and Hong Jeong. Geometric relationship between stereo disparity and optical flow and an efficient recursive algorithm. *Journal of Pattern Recognition Research*.

[11] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *TPAMI*, 2008.

[12] Heiko Hirschmuller and Daniel Scharstein. Evaluation of cost functions for stereo matching. In *CVPR*, 2007.

[13] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial Intelligence*, 1981.

[14] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *CVPR*, 2018.

[15] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017.

[16] Joel Janai, Fatma Güney, Anurag Ranjan, Michael J. Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *ECCV*, 2018.

[17] J Yu Jason, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *ECCV*, 2016.

[18] Sunghun Joung, Seungryong Kim, Kihong Park, and Kwanghoon Sohn. Unsupervised stereo matching using confidential correspondence consistency. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[19] Takeo Kanade and Masatoshi Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. In *ICRA*, 1991.

[20] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, 2017.

[21] Andreas Klaus, Mario Sormann, and Konrad Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *International Conference on Pattern Recognition*, 2006.

[22] Hsueh-Ying Lai, Yi-Hsuan Tsai, and Wei-Chen Chiu. Bridging stereo matching and optical flow via spatiotemporal correspondence. In *CVPR*, 2019.

[23] Ang Li and Zejian Yuan. Occlusion aware stereo matching via cooperative unsupervised learning. In *ACCV*, 2018.

[24] Liang Liu, Guangyao Zhai, Wenlong Ye, and Yong Liu. Unsupervised learning of scene flow estimation fusing with local rigidity. In *IJCAI*, 2019.

[25] Pengpeng Liu, Irwin King, Michael R. Lyu, and Jia Xu. Ddflow: Learning optical flow with unlabeled data distillation. In *AAAI*, 2019.

[26] Pengpeng Liu, Irwin King, Michael R. Lyu, and Jia Xu. Selflow: Self-supervised learning optical flow. In *CVPR*, 2019.

[27] D. Maurer and A. Bruhn. Proflow: Learning to predict optical flow. In *BMVC*, 2018.

[28] 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 *CVPR*, 2016.

[29] Xing Mei, Xun Sun, Weiming Dong, Haitao Wang, and Xiaopeng Zhang. Segment-tree based cost aggregation for stereo matching. In *CVPR*, 2013.

[30] Simon Meister, Junhwa Hur, and Stefan Roth. UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In *AAAI*, 2018.

[31] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015.

[32] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *CVPR*, 2017.

[33] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *CVPR*, 2019.

[34] Zhile Ren, Orazio Gallo, Deqing Sun, Ming-Hsuan Yang, Erik B. Sudderth, and Jan Kautz. A fusion approach for multi-frame optical flow estimation. In *IEEE Winter Conference on Applications of Computer Vision*, 2019.

[35] Zhe Ren, Junchi Yan, Bingbing Ni, Bin Liu, Xiaokang Yang, and Hongyuan Zha. Unsupervised deep learning for optical flow estimation. In *AAAI*, 2017.

[36] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *CVPR*, 2015.

[37] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 2002.

[38] Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *CVPR*, 2010.

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

[40] Jian Sun, Nan-Ning Zheng, and Heung-Yeung Shum. Stereo matching using belief propagation. *TPAMI*, 2003.

[41] Alessio Tonioni, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Unsupervised adaptation for deep stereo. In *ICCV*, 2017.

[42] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *CVPR*, 2019.

[43] Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, and Wei Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *CVPR*, 2019.

[44] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, and Wei Xu. Occlusion aware unsupervised learning of optical flow. In *CVPR*, 2018.

[45] Jia Xu, René Ranftl, and Vladlen Koltun. Accurate Optical Flow via Direct Cost Volume Processing. In *CVPR*, 2017.

[46] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In *ECCV*, 2018.

[47] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018.

[48] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *ECCV*, 1994.

[49] Ke Zhang, Jiangbo Lu, and Gauthier Lafruit. Cross-based local stereo matching using orthogonal integral images. *IEEE Transactions on Circuits and Systems for Video Technology*, 2009.

[50] Yiran Zhong, Hongdong Li, and Yuchao Dai. Open-world stereo video matching with deep rnn. In *ECCV*, 2018.

[51] Chao Zhou, Hong Zhang, Xiaoyong Shen, and Jiaya Jia. Unsupervised learning of stereo matching. In *ICCV*, 2017.

[52] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.

[53] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *ECCV*, 2018.