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].

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.

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 $(x, y) = \frac{f(X,Y)}{Z}$, 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

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

(1)

Let $w = (u, v)$ 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)
This can be further simplified as,
\[ (u, v) = f' \frac{(\Delta X, \Delta Y)}{Z} - f'' \frac{\Delta Z}{Z^2} (X, Y) \]  
(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,
\[
\begin{align*}
(u_l, v_l) &= f' \frac{(\Delta X, \Delta Y)}{Z} - f'' \frac{\Delta Z}{Z^2} (X, Y) \\
(u_r, v_r) &= f' \frac{(\Delta X, \Delta Y)}{Z} - f'' \frac{\Delta Z}{Z^2} (X - B, Y)
\end{align*}
\]  
(3)

This can be further simplified as,
\[
\begin{align*}
&\begin{align*}
&u_r - u_l = f' B \frac{\Delta Z}{Z} \\
v_r - v_l = 0
\end{align*}
\end{align*}
\]  
(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 \( d = f' B \frac{\Delta Z}{Z} \). Take the time derivative, we have
\[
\frac{\Delta d}{\Delta t} = -f' B \frac{\Delta Z}{Z^2} \frac{\Delta t}{\Delta t}
\]  
(5)

Similarly, we set time step to be one unit, then
\[
\Delta t_{t+1} - \Delta t = -f' B \frac{\Delta Z}{Z^2}
\]  
(6)

With Eq. (4) and (6), we finally obtain,
\[
\begin{align*}
&\begin{align*}
&u_r - u_l = -(\Delta t_{t+1} - \Delta t) \\
v_r - v_l = 0
\end{align*}
\end{align*}
\]  
(7)

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 \( w_{1 \rightarrow 3} \) denote the optical flow from \( I_1 \) to \( I_3 \), \( w_{1 \rightarrow 2} \) 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 \( w_{1 \rightarrow 4} \) (green row in Fig. 1). In this case,
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 \( p_i \) is a pixel in \( I_1 \), \( p_r, p_{r+1}, p_{r+1}^t \) are its correspondence pixels in \( I_2, I_3 \) and \( I_4 \) respectively, then we have,

\[
\begin{align*}
    p_r &= p_i + w_{1 \rightarrow 2}(p_i) \\
    p_{r+1} &= p_r + w_{2 \rightarrow 3}(p_r) \\
    p_{r+1}^t &= p_r^t + w_{3 \rightarrow 4}(p_r^t)
\end{align*}
\]

(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,

\[
    w_{1 \rightarrow 4}(p_i^t) = (p_{r+1}^t - p_r^t) + (p_i^t - p_i) \\
    = w_{2 \rightarrow 4}(p_r^t) + w_{1 \rightarrow 2}(p_i).
\]

(9)

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

\[
    w_{1 \rightarrow 3}(p_i^t) = (p_{r+1}^t - p_r^t) + (p_i^t - p_i) \\
    = w_{3 \rightarrow 4}(p_r^t) + w_{1 \rightarrow 2}(p_i).
\]

(10)

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

\[
    w_{2 \rightarrow 4}(p_r^t) - w_{1 \rightarrow 3}(p_i^t) = w_{3 \rightarrow 4}(p_{r+1}^t) - w_{1 \rightarrow 2}(p_i^t).
\]

(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

\[
\begin{align*}
    u_{2 \rightarrow 4}(p_r^t) - u_{1 \rightarrow 3}(p_i^t) &= u_{3 \rightarrow 4}(p_{r+1}^t) - u_{1 \rightarrow 2}(p_i^t) \\
    v_{2 \rightarrow 4}(p_r^t) - v_{1 \rightarrow 3}(p_i^t) &= 0
\end{align*}
\]

(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:

\[
\begin{align*}
    u_{1 \rightarrow 4}(p_i^t) &= u_{2 \rightarrow 4}(p_r^t) + u_{1 \rightarrow 2}(p_i^t) \\
    v_{1 \rightarrow 4}(p_i^t) &= v_{2 \rightarrow 4}(p_r^t) \\
    u_{1 \rightarrow 4}(p_r^t) &= u_{3 \rightarrow 4}(p_{r+1}^t) + u_{1 \rightarrow 3}(p_i^t) \\
    v_{1 \rightarrow 4}(p_i^t) &= v_{1 \rightarrow 3}(p_i^t)
\end{align*}
\]

(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
Then we measure the difference between the warped target image and the reference image.

Stage 1: Predicting confident optical flow with geometric constraints. With the estimated optical flow map \( w_{i \rightarrow j} \), we warp the target image \( I_j \) toward the reference image \( I_i \). Then we measure the difference between the warped target image \( I_{j \rightarrow i} \) 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, but also for non-occluded pixels. Because the model just does not have the ability to predict them.

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. 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, so 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].

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. \( f_g \) and \( b_g \) denote results of foreground and background regions respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>Train</th>
<th>KITTI 2012</th>
<th>KITTI 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Stereo</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>EPE-all</td>
<td>EPE-noc</td>
</tr>
<tr>
<td>SpyNet [32]</td>
<td>✓</td>
<td>3.36</td>
<td>–</td>
</tr>
<tr>
<td>FlowFieldsCNN [1]</td>
<td>✗</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>DCFlow [45]</td>
<td>✓</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>FlowNet2 [15]</td>
<td>✓</td>
<td>(1.28)</td>
<td>–</td>
</tr>
<tr>
<td>LiteFlowNet [14]</td>
<td>✓</td>
<td>(1.05)</td>
<td>–</td>
</tr>
<tr>
<td>PWC-Net [39]</td>
<td>✓</td>
<td>(1.45)</td>
<td>–</td>
</tr>
<tr>
<td>MFF [34]</td>
<td>✓</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>SelFlow [26]</td>
<td>✓</td>
<td>(0.76)</td>
<td>–</td>
</tr>
<tr>
<td>BackToBasic [17]</td>
<td>✗</td>
<td>11.3</td>
<td>4.3</td>
</tr>
<tr>
<td>DSTFlow [35]</td>
<td>✗</td>
<td>10.43</td>
<td>3.29</td>
</tr>
<tr>
<td>UnFlow-CSS [30]</td>
<td>✗</td>
<td>3.29</td>
<td>1.26</td>
</tr>
<tr>
<td>OccAwareFlow [44]</td>
<td>✗</td>
<td>3.55</td>
<td>–</td>
</tr>
<tr>
<td>MultiFrameOccFlow-None [16]</td>
<td>✗</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>MultiFrameOccFlow-Soft [16]</td>
<td>✗</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>DSTFlow [25]</td>
<td>✓</td>
<td>2.35</td>
<td>1.02</td>
</tr>
<tr>
<td>SelFlow [26]</td>
<td>✓</td>
<td>1.69</td>
<td>0.91</td>
</tr>
<tr>
<td>LiteFlowNet [14]</td>
<td>✓</td>
<td>2.56</td>
<td>1.39</td>
</tr>
<tr>
<td>UnFlow-CSS [30]</td>
<td>✓</td>
<td>1.64</td>
<td>1.04</td>
</tr>
<tr>
<td>UnFlow-CSS [30]</td>
<td>✓</td>
<td>1.64</td>
<td>1.04</td>
</tr>
<tr>
<td>LiteFlowNet [14]</td>
<td>✓</td>
<td>2.56</td>
<td>1.39</td>
</tr>
</tbody>
</table>

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.
GT Disparity

(a) $I_1$ 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. * 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>KITTI 2012</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EPE-all</td>
<td>EPE-noc</td>
<td>EPE-occ</td>
<td>D1-all</td>
<td>D1-noc</td>
<td>D1-all (test)</td>
<td>EPE-all</td>
<td>EPE-noc</td>
</tr>
<tr>
<td>Joong et al. [18]</td>
<td>2.12</td>
<td>1.44</td>
<td>30.91</td>
<td>10.41</td>
<td>8.33</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Godard et al. [8] *</td>
<td>2.37</td>
<td>1.63</td>
<td>33.62</td>
<td>9.64</td>
<td>7.89</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Zhou et al. [51]</td>
<td>2.37</td>
<td>1.63</td>
<td>33.62</td>
<td>9.64</td>
<td>7.89</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>OASM-Net [23]</td>
<td>1.78</td>
<td>1.68</td>
<td>6.25</td>
<td>11.57</td>
<td>10.61</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Guo et al. [9] *</td>
<td>1.56</td>
<td>1.38</td>
<td>8.62</td>
<td>5.82</td>
<td>4.89</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>UnOS [43]</td>
<td>1.16</td>
<td>1.09</td>
<td>4.14</td>
<td>6.45</td>
<td>5.82</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Ours+Lp</td>
<td>1.73</td>
<td>1.13</td>
<td>27.03</td>
<td>7.88</td>
<td>5.87</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Ours+Lp+Ls+Ls</td>
<td>1.62</td>
<td>0.94</td>
<td>29.26</td>
<td>6.69</td>
<td>4.69</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Ours+Lp+Ls+Ls+Self-Supervision</td>
<td>1.01</td>
<td>0.93</td>
<td>4.52</td>
<td>5.14</td>
<td>4.59</td>
<td>5.11</td>
<td>5.11</td>
<td>5.11</td>
</tr>
</tbody>
</table>

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 consensus transform, which has shown to be robust for illumination change [30, 25, 26]. Denote $M_{i\rightarrow j}$ as the confident map from $I_i$ to $I_j$ is $M_{i\rightarrow j}$, then $L_p$ is defined as,

$$L_p = \sum_{i,j} \psi(I_i(p) - I_j(p)) \odot M_{i\rightarrow j}(p),$$

(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 $w_{1\rightarrow 4}$, $w_{2\rightarrow 4}$, $w_{1\rightarrow 2}$ and $w_{3\rightarrow 4}$ for an example, we first compute the confident map for quadrilateral constraint $M_q(p_i) = M_{1\rightarrow 2}(p_i) \odot M_{1\rightarrow 3}(p_i) \odot M_{1\rightarrow 4}(p_i)$. Then according to Eq. (12), we divide $L_q$ into two components on the $x$ direction $L_{qu}$ and $y$ direction $L_{qv}$ respectively:

$$L_{qu} = \sum_{p_i} \psi(u_{1\rightarrow 2}(p_i^l) + u_{2\rightarrow 4}(p_i^l) - u_{1\rightarrow 3}(p_i^l) -$$

$$u_{3\rightarrow 4}(p_i^l + 1)) \odot M_q(p_i^l) / \sum_{p_i^l} M_q(p_i^l),$$

(15)

$$L_{qv} = \sum_{p_i} \psi(v_{2\rightarrow 4}(p_i^l) - v_{1\rightarrow 3}(p_i^l)) \odot M_q(p_i^l) / \sum_{p_i^l} M_q(p_i^l),$$

(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
At test time, only the student model is needed, and we can compute the direction in the same way. Let $\lambda$ be 1, since there are 12 optical flow estimations for the 4 frames. During training, we use the raw multi-view extensions of KITTI 2012 [6] and KITTI 2015 [31] and exclude neighbor pixels. 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 achieve FL-all = 7.68% and Fl-noc = 4.02%, which is even better than state-of-the-art fully supervised learning methods.

5. Experiment

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 SelfFlow. 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
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, which further help in the 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


