Bridging Stereo Matching and Optical Flow via Spatiotemporal Correspondence

Hsueh-Ying Lai[1]   Yi-Hsuan Tsai[2]   Wei-Chen Chiu[1]

[1]National Chiao Tung University, Taiwan   [2]NEC Laboratories America

Abstract

Stereo matching and flow estimation are two essential tasks for scene understanding, spatially in 3D and temporally in motion. Existing approaches have been focused on the unsupervised setting due to the limited resource to obtain the large-scale ground truth data. To construct a self-learnable objective, co-related tasks are often linked together to form a joint framework. However, the prior work usually utilizes independent networks for each task, thus not allowing to learn shared feature representations across models. In this paper, we propose a single and principled network to jointly learn spatiotemporal correspondence for stereo matching and flow estimation, with a newly designed geometric connection as the unsupervised signal for temporally adjacent stereo pairs. We show that our method performs favorably against several state-of-the-art baselines for both unsupervised depth and flow estimation on the KITTI benchmark dataset.
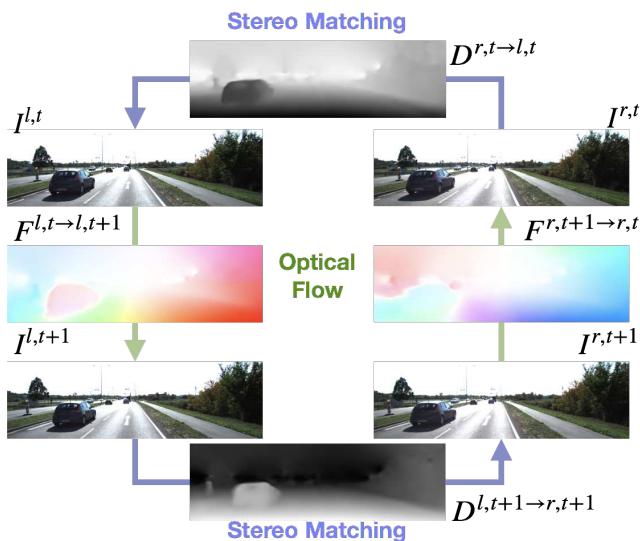
Figure 1. Using temporally adjacent stereo pairs as input, our model can estimate the correspondence maps of each pair via geometric connections, thus bridging stereo matching and optical flow through multiple reconstruction, forming a cycle.

1. Introduction

Reconstructing 3D motion from the real-world visual data has long been a fundamental problem in computer vision and is substantial for numerous applications such as robotics, virtual/augmented reality, and autonomous driving. Among the tasks of understanding 3D motion, two of the most commonly studied scenarios are optical flow estimation and stereo matching for depth estimation. Generally, the motion in 3D after projection into the image plane of a camera stands for the optical flow between two consecutive frames in a video, while the 3D structure captured by two horizontally displaced cameras builds the stereo rig as the binocular vision system of human eyes. Thus, the estimation of optical flow and stereo matching, which discover the pixel displacement across temporally adjacent frames and stereo pairs, provide crucial access to the 3D information.

Recently, deep learning-based approaches have shown tremendous improvement for both optical flow estimation and stereo matching in the supervised learning setting [2, 13, 14, 3, 11, 6, 10, 19]. However, these methods usually rely on large-scale datasets with ground truths, but such

annotating efforts are significantly expensive, especially in forms of pixel-wise displacement for optical flow and stereo matching. For eliminating limitation of datasets and potential issues such as poor model generalization across various scenes, several approaches are proposed recently to explore the unsupervised learning frameworks [31, 17, 9].

In the unsupervised learning setting, a common practice is to relate different tasks (e.g., optical flow, depth estimation, or camera pose estimation) and utilize photometric consistency to measure the pixel correspondences across frames [20, 27, 28, 31, 33]. Nevertheless, existing approaches utilize separate networks for each task, and thus the feature representations are not effectively shared across tasks. In this paper, we argue that there should exist a principled model, which is capable of learning joint representations for tasks that are highly co-related.

Although the properties of pixel correspondence used in stereo matching and optical flow estimation are slightly different, as the former considers the horizontal offset while the later has movement in both horizontal and vertical di-

rections, the common goal is obviously shared (i.e., finding pixel correspondences). By taking advantages of such a correlation, we propose to design a single network for simultaneously estimating optical flow and stereo matching, and show that these two tasks are beneficial to each other via learning shared feature representations. Moreover, we construct an unsupervised learning framework with modelling the geometric connections between both tasks based on temporally adjacent stereo pairs (as shown in Figure 1), in which this type of data is easily accessible as the popularity of stereo video cameras. We design a warping function that considers the consistency across adjacent video frames, and sequentially feed the training data both from flow and stereo pairs to meet the designed geometric constraints.

Extensive experiments are conducted on both KITTI2012 [8] and KITTI2015 [18] benchmark dataset to evaluate the effectiveness of the proposed method and show favorable performance against several state-of-the-art algorithms. In addition, we sequentially demonstrate the mutual benefit of jointly learning both tasks of optical flow estimation and stereo matching, successfully showing the improvement via utilizing the proposed geometric connections built upon stereo video data. The main contributions of the paper are summarized as the follows:

- We propose a single and principled network for joint estimating optical flow and stereo matching to account for their shared representations, in which the common goal is to find pixel correspondence across images.

- We introduce geometric constraints during the joint learning process, which provides an effective signal for modeling the consistency (i.e., spatiotemporal correspondence) across two tasks and is then utilized as an objective for unsupervised training.

- We develop an efficient training scheme for the joint optimization on two tasks within a single framework and show that both tasks benefit each other.

2. Related Works

We organize and discuss related approaches, including stereo matching, depth estimation, optical flow estimation, and the joint framework of them.

Unsupervised Learning of Depth Estimation. Stereo matching for depth estimation has been a classical computer vision problem for decades. Prior to the recent renaissance of deep learning, many approaches are proposed to tackle this problem based on diverse strategies, such as hand-crafting feature descriptors for matching local regions across frames, or formulating stereo matching upon a graphical model and resolving it by complicated energy minimization. With large annotated datasets are available

(e.g., KITTI [8]) in recent years, better matching functions to measure the similarity between image patches are learnt by deep neural networks [15, 30, 1] which obtain significant boost in performance. Simultaneously, estimating depth directly from monocular images based on deep models in the supervised learning manner is also widely explored [2, 14]. However, the requirement for training data with ground truths is expensive to meet, and thus the unsupervised learning scheme [31, 9, 32, 16] is popularly adopted. Here we review several of them as follows.

Godard *et al.* [9] learn to estimate disparity maps which are used to warp between images in a stereo pair for optimizing objectives of left-right consistency. Instead of exploring the pixel correspondence within stereo pairs, given a video sequence, [32] jointly estimates both the monocular depth of each frame as well as the camera motion such that consecutive frames can be reconstructed between each others, and are used for evaluating photometric consistency as loss functions. In [16], the authors combine the concept of monocular depth estimation and stereo matching, where binocular views in a stereo pair are first synthesized by using the depth map estimated from the monocular image. Then the stereo matching network is applied to produce the final depth estimation. Typically, these methods attempt to regress depth map solely from monocular inevitably depends on the quality of training data and hardly generalize to unseen scenes. In contrast, the models for stereo matching concentrate on learning to match pixels between images and thus have better generalizability, in which we aim to address the same stereo matching task in this paper. In the work of Zhou *et al.* [31], the authors propose to learn stereo matching via iterative left-right consistency check. Godard *et al.* [9] also extend their monocular depth estimation framework to perform stereo matching and obtain better performance with respect to its monocular version.

Unsupervised Learning of Optical Flow. The research works addressing optical flow estimation follow the same evolution as the ones for depth estimation, starting from conventional methods [4, 5], advancing to deep learning models based on supervised setting [3, 11], and then exploring unsupervised learning approaches [29, 21, 17]. When unsupervised learning of optical flow are first introduced in FlowNet-Simple [29] and DSTFlow [21], they utilize the similar objectives of photometric consistency across frames and local smoothness in the estimated flow map. However, these works do not take the severe occlusion issue into consideration when there are objects with large movement. In order to resolve the artifacts resulted from the warping operation, [25, 17, 12] handle regions of occlusion by analyzing the inconsistency between forward and backward flow maps. [17] further replaces the typical L1 loss with the ternary census transform for measuring photometric con-

sistency, providing more reliable constancy assumption in realistic situations. Moreover, [12] advance the optical flow estimation and occlusion handling by explicitly reasoning over multiple consecutive frames within a time window.

Joint Learning Framework of Depth and Optical Flow. Recently, numerous works have been proposed to jointly learn both depth and optical flow estimation models via employing geometric relations between flow, depth, and camera poses. In [32], based on the assumption of rigid scenes, the pixel correspondence between temporally adjacent frames caused by camera movement is derived from the estimates on both monocular depth and camera poses, and thus it becomes the key to define objectives for joint training. GeoNet [28] follows the similar idea as [32] but particularly introduces non-rigid motion localizer to handle moving objects in the optical flow map. Yang *et al.* [27] explicitly disentangle the dynamic objects from static background in a video based on a motion network, and carefully model it together with depth, flow, and camera pose estimation by using geometric constraints. The occlusion mask as well as 3D motion maps for dynamic and static regions can thus be obtained. DF-net [33] especially leverages the geometric consistency between the estimated flow from optical flow model and the synthetic 2D optical flow obtained from estimates of depth and camera motion, in which it shows benefits for simultaneously training monocular depth prediction and optical flow estimation networks. Along the same track of unsupervised learning but unlike the aforementioned research works where separate networks are learned for each task, our proposed method tackles stereo matching and optical flow estimation within a single and principled network, and relates them through geometric connections built upon temporally adjacent stereo pairs.

3. Proposed Method

In this section, we first describe the overall structure of how we construct the geometric relations among stereo videos. Second, we introduce each component of the proposed method, including unsupervised loss functions shared across stereo matching and flow estimation, a newly proposed 2-Warp loss that measures the consistency between two tasks, and occlusion handling for flow estimation.

3.1. Overall Structure

As motivated previously that both optical flow estimation and stereo matching aim to find pixel correspondences across images, our goal is to learn a single and principled network for these two tasks in an unsupervised learning manner with exploiting their geometric relations stemmed from stereo videos. Figure 3 illustrates the framework of the proposed method, which will be detailed in the following subsections.
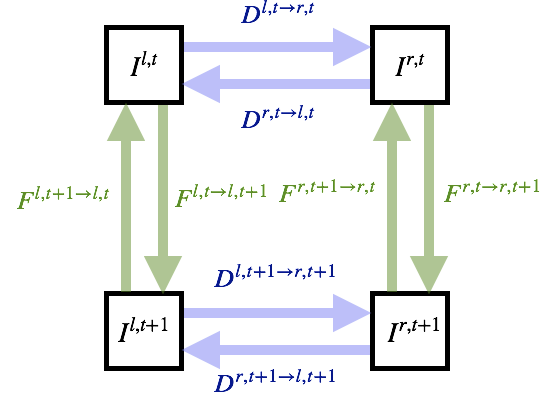

Figure 2. The relation of bridging stereo pairs and consequent frames. We can estimate the correspondence maps of any directions based on the input pairs and their reconstruction direction.

The network P in our method is based on the model used in *Monodepth* [9], which is now extended from its original usage of monocular depth estimation to take two input frames and output both horizontal and vertical offsets for pixel correspondences across input frames. Assuming two temporally adjacent stereo pairs are given as $\{I^{l,t}, I^{r,t}, I^{l,t+1}, I^{r,t+1}\}$ where the superscripts l, r denote left and right frames in a stereo pair respectively, and $t, t+1$ indicate their temporal indexes. Our network P is able to perform stereo matching to obtain the forward pixel correspondence $D^{l,t\to r,t}$ from $I^{l,t}$ to $I^{r,t}$ as well as the backward one $D^{r,t\to l,t}$ from $I^{r,t}$ to $I^{l,t}$:

$$
\begin{aligned}
D^{l,t\to r,t} &= P(I^{l,t}, I^{r,t}) \\
D^{r,t\to l,t} &= P(I^{r,t}, I^{l,t})
\end{aligned}
\tag{1}
$$

Likewise, for another stereo pair at time $t+1$, we obtain:

$$
\begin{aligned}
D^{l,t+1\to r,t+1} &= P(I^{l,t+1}, I^{r,t+1}) \\
D^{r,t+1\to l,t+1} &= P(I^{r,t+1}, I^{l,t+1})
\end{aligned}
\tag{2}
$$

The forward/backward optical flow maps on the left and right views can also be estimated using our network:

$$
\begin{aligned}
F^{l,t\to l,t+1} &= P(I^{l,t}, I^{l,t+1}) \\
F^{l,t+1\to l,t} &= P(I^{l,t+1}, I^{l,t}) \\
F^{r,t\to r,t+1} &= P(I^{r,t}, I^{r,t+1}) \\
F^{r,t+1\to r,t} &= P(I^{r,t+1}, I^{r,t})
\end{aligned}
\tag{3}
$$

The overall relations are shown in Figure 2. With these pixel correspondences, we aim to reconstruct a frame given its counterpart of a stereo pair or its temporal adjacency, based on a warping function W. For instance, frame $I^{r,t}$ can be reconstructed as:

$$
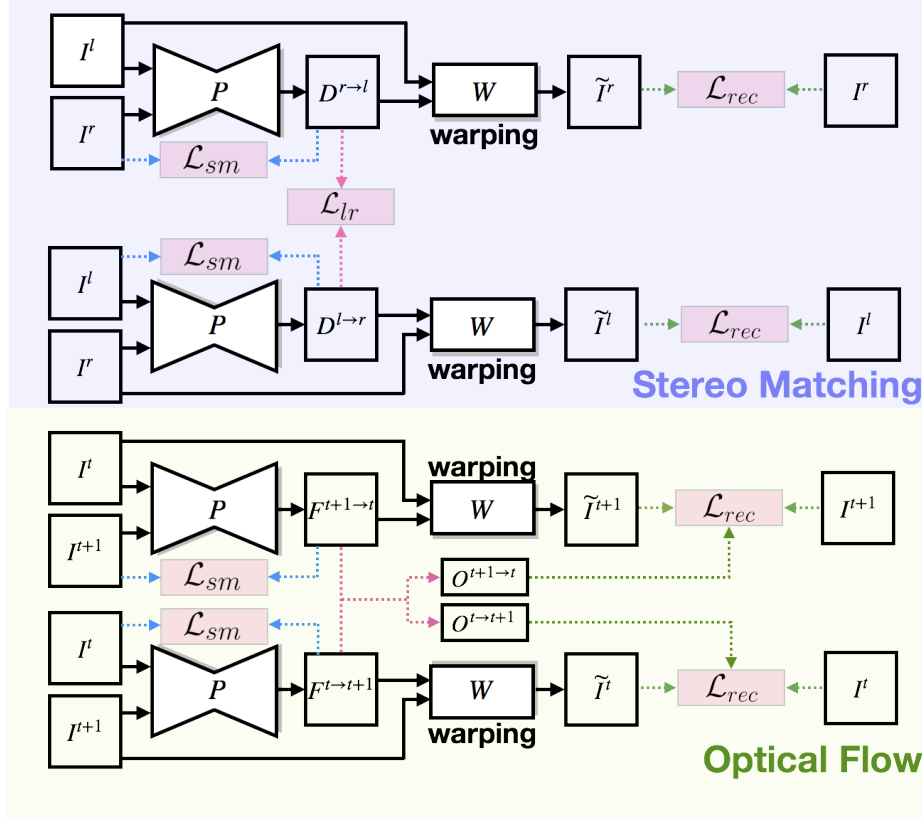\tilde{I}^{r,t} = W(I^{l,t}, D^{r,t\to l,t}),
\tag{4}
$$

Figure 3. Overall structure of our method. Our framework consists of a single model P that estimates dense correspondence maps based on the order of two input images for both stereo matching and optical flow. Each pair can be fed into P but in a different image order (e.g., (I_l, I_r) and (I_r, I_l)), and thus two reconstruction loss $\mathcal{L}_{rec}$ are able to be optimized based on two warping functions W obtained from each pair. Between these two tasks, two difference are: (1) we apply left-right consistency $\mathcal{L}_{lr}$ to stabilize the stereo matching part only; (2) occlusion map derived from the correspondence maps of two opposite directions is adopted on the reconstruction loss for solving the largely occluded area for optical flow only.

from its corresponding left view $I^{l,t}$ and the backward stereo matches $D^{r,t \to l,t}$. Similarly, $I^{l,t}$ can be reconstructed as:

$$\tilde{I}^{l,t} = W(I^{l,t+1}, F^{l,t \to l,t+1}), \qquad (5)$$

from its next frame $I^{l,t+1}$ via the flow $F^{l,t \to l,t+1}$. For simplicity, we skip listing here for other combinations across frames, which should be easily derivable.

3.2. Occlusion Estimation for Optical Flow

Before introducing the designed unsupervised loss functions in our framework, we describe first how we tackle the common occlusion issue for flow estimation. During training, there would be some occluded regions only visible at frame t but having no corresponding pixels at frame $t + 1$, as the camera or objects may have large movement. This causes the inconsistent warping process in appearance between the reconstructed image and the target one.

In order to deal with the occlusion issue, we utilize the forward-backward consistency check [23, 25, 33] to local-

ize the potentially occluded regions. More precisely, applying warping operation on a backward map by its corresponding forward map, e.g., $W(F^{l,t+1 \to l,t}, F^{l,t \to l,t+1})$, ideally could reconstruct the forward map with a negative sign in non-occluded regions. To this end, we follow the technique as used in [17] and by taking the pair of $\{I^{l,t}, I^{l,t+1}\}$ as an example, pixels are considered as occluded while the criterion below is violated:

$$|F^{l,t \to l,t+1} + W(F^{l,t+1 \to l,t}, F^{l,t \to l,t+1})|^2$$
$$< \alpha_1(|F^{l,t \to l,t+1}|^2 + |W(F^{l,t+1 \to l,t}, F^{l,t \to l,t+1})|^2) + \alpha_2, \qquad (6)$$

where the hyper-parameters α_1 and α_2 are set to 0.01 and 0.5 respectively. An occlusion map O is then obtained by setting 0 to those occluded regions and 1, otherwise.

3.3. Unsupervised Loss Functions

One key factor to make the proposed unsupervised method work is to design plausible loss functions that can exploit various connection across video frames. In the following, we sequentially introduce the utilized loss
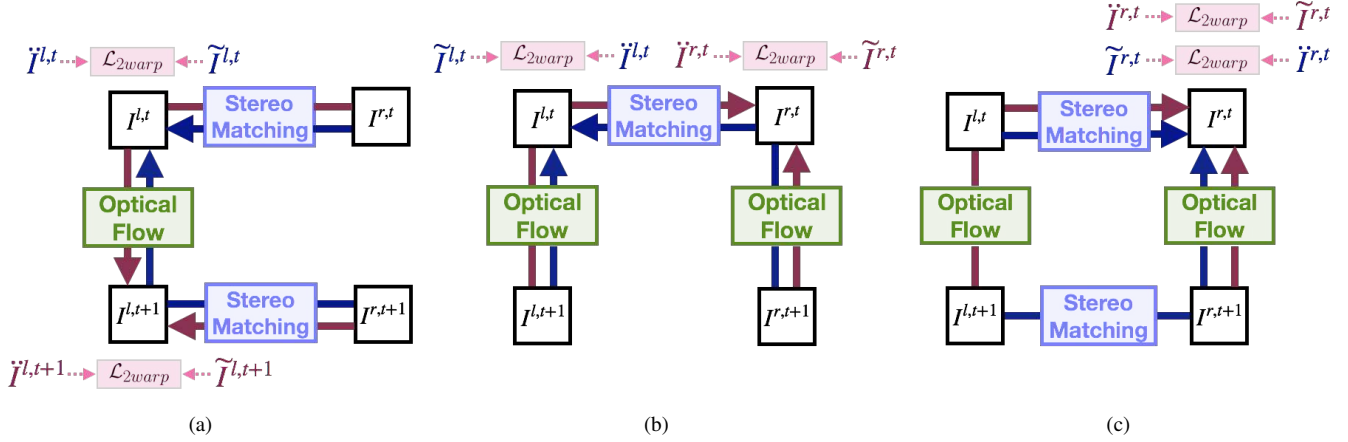
Figure 4. Our proposed 2-warp modules. The arrows indicate the warping direction and a 2-warp reconstruction loss is performed when the arrows with the same color meet, forcing reconstructed images via the 2-warp operations to be consistent. Here, we introduce three types of 2-warp functions and will discuss them in the section of experiments.

functions, including self-supervised reconstruction loss, smoothness loss, left-right consistency loss, and 2-Warp consistency loss to model the relation across stereo videos. Here, we use the pair of $\{I^{l,t}, I^{l,t+1}\}$ as an example for explanation and all the loss functions apply to both stereo/flow pairs, unless stated specifically. The overall structure of the proposed framework and loss functions is presented in Figure 3.

Reconstruction Loss. The reconstruction loss $\mathcal{L}_{rec}$ is similar to the one used in Monodepth [9] but with occlusion-aware constraints. The loss is the weighted sum of SSIM-based loss and L1 loss which compares $I^{l,t}$ and its reconstruction $\tilde{I}^{l,t}$:

$$\mathcal{L}_{rec}^{l,t \to l,t+1} = \frac{1}{\sum_{i,j} O}[\sum_{i,j}(\alpha \frac{1 - SSIM(I_{ij}^{l,t}, \tilde{I}_{ij}^{l,t})}{2} \quad (7)$$
$$+ (1-\alpha)|I_{ij}^{l,t} - \tilde{I}_{ij}^{l,t}|) \cdot O],$$

where O is the occlusion map derived from Section 3.2, subscript i, j indicates pixel coordinates, and α denotes the weights between SSIM and L1 loss. Since our occlusion maps are only used in the image pairs for flow estimation, all the elements in the occlusion map would be equal to 1 when $\mathcal{L}_{rec}$ is applied on image pairs for stereo matching.

Smoothness Loss. For the smoothness loss $\mathcal{L}_{sm}$, we adopt the formulation introduced in [25] which encourages the correspondence maps to be locally smooth but also maintains edges that should be aligned with the structure of images:

$$\mathcal{L}_{sm}^{l,t \to l,t+1} = \frac{1}{N} \sum_{i,j} \sum_{d \in (x,y)} |\partial_d^2 F^{l,t \to l,t+1}| e^{-\beta|\partial_d I_{ij}^{l,t}|} \quad (8)$$

where β denotes the edge-weighted hyperparameter. Here we adopt the second-order and the first-order derivatives on the correspondence map and the image, respectively.

Left-right Correspondence Consistency Loss. To improve the accuracy of correspondence map estimation as well as balance the performance of left-right estimation, we not only check the consistency of left-right reconstruction but also check left-right correspondence. Similar to the occlusion detection, our left-right consistency loss $\mathcal{L}_{lr}$ is derived from reconstructing the correspondence map pair by warping each other and compute the absolute L1 difference loss. Following [26], this consistency term is only adopted on stereo pairs:

$$\mathcal{L}_{lr}^{l,t \to r,t} = \frac{1}{N} \sum_{i,j} |D^{l,t \to r,t} + W(D^{r,t \to l,t}, D^{l,t \to r,t})| \quad (9)$$

2-Warp Consistency Loss. To reinforce the structure of stereo matching and optical flow estimation, we introduce a new 2-Warp consistency loss. That is, we warp an image twice through both the optical flow and stereo sides. Figure 4 presents three types of the possible 2-warp operations that we investigate. We will introduce the details of the first one as follows, while the others can be derived similarly.

Following previous works of depth estimation, we do not apply occlusion maps on stereo pairs, so that we could easily derive the 2-warp occlusion maps from estimated flow maps. To reconstruct $I^{r,t}$ from $I^{l,t+1}$ via $I^{l,t}$, the occlusion map and the 2-warp reconstructed image are written as:

$$O^{r,t \to l,t+1} = W(O^{l,t \to l,t+1}, D^{r,t+1 \to l,t+1}). \quad (10)$$

$$\ddot{I}^{r,t} = W(W(I^{l,t+1}, F^{l,t+1 \to l,t+1}), F^{r,t \to l,t}). \quad (11)$$

The occlusion regions between the stereo pairs at time t is the area where objects occlude the background at time

1894

$t+1$, so the occlusion area can be mapped by $D^{l,t+1\to r,t+1}$. Therefore, warping $O^{l,t\to l,t+1}$ by $D^{r,t+1\to l,t+1}$ as our 2-warp occlusion map is valid. Similar to (7), we could apply the occlusion-aware reconstruction loss between reconstructed $\ddot{I}^{r,t}$ from $I^{l,t+1}$ via $I^{l,t}$ and the reconstructed $\tilde{I}^{r,t}$ directly from $I^{r,t+1}$, as illustrated in Figure 4(a).

$$\mathcal{L}_{2warp}^{r,t\to l,t+1} = \frac{1}{\sum_{i,j} O^{r,t\to l,t+1}}[\sum_{i,j}(\alpha\frac{1-SSIM(\ddot{I}_{ij}^{r,t}, \tilde{I}_{ij}^{r,t})}{2}$$
$$+(1-\alpha)|\ddot{I}_{ij}^{r,t}-\tilde{I}_{ij}^{r,t}|)\cdot O^{r,t\to l,t+1}].$$
$$(12)$$

Total Loss. The total loss of the proposed framework is:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_{sm}\mathcal{L}_{sm} + \lambda_{lr}\mathcal{L}_{lr} + \lambda_{2warp}\mathcal{L}_{2warp} \quad (13)$$

We note that, all these terms except 2-Warp consistency, have its mirrored counterpart at each scale for the multi-scale estimation as in Monodepth [9].

4. Experimental Results

We evaluate the proposed method for both depth estimation and flow estimation on the KITTI dataset [8]. We show that our framework is able to achieve competitive performance on both tasks. Moreover, to show the merit of jointly learning the shared feature representations, we conduct experiments to validate that introducing stereo and flow pairs improves both performance. We further enforce geometric constraints to construct the spatiotemporal correspondence in the stereo video and show that such constraint improves the performance via our newly proposed warping function. The code and model will be made available for the public. More results are provided in the supplementary material.

4.1. Implementation Details

During training, we use a batch of size 2, each with two adjacent stereo pairs, i.e., 4 stereo pairs and 4 flow pairs. Images are scaled to the size of 512×256. Our model is based on Monodepth [9] using ResNet-50 as the encoder, with modifications of the last layer before output at each scale to generate 2-channel maps including horizontal and vertical correspondence maps. The data augmentation follows Monodepth, containing left-right flipping, color augmentation of random gamma, brightness and color shifts, where each augmentation type has 50% of chances to be selected. Each color augmentation is sampled by uniform distributions in the ranges of [0.8, 1.2], [0.5, 2.0], [0.8, 1.2] respectively. We use Adam as our optimizer with default parameter settings. The learning rate is set as 10^{-4} and we apply learning rate decay which is halved every 3 epochs for 5 times when training on full training set. Our hyper-parameters $\{\alpha, \beta, \lambda_{sm}, \lambda_{lr}, \lambda_{2warp}\}$ are set to $\{0.85, 10,$ 10, 0.5, 0.2\}. When only training on stereo pairs, λ_{lr} would be 1 for balancing the proportion of stereo pairs in a batch. Please note that we use a model variant trained without 2-warp consistency loss (i.e. denoted as Ours (flow+stereo) in Table 1, 2, and 3) for better initializing the learning of our proposed full models.

4.2. Dataset and Setting

The KITTI dataset contains stereo sequences of real road scenes, providing accurate but sparse depth and optical flow ground truth for a small subset. We evaluate our method on the KITTI 2012 and 2015 datasets, in which there are 194 and 200 pairs of flow and stereo with high quality annotations, covering 28 scenes of the KITTI raw dataset. During training, we generate 28968 cycles (i.e., a cycle contains 4 images as in Figure 2) from the remaining 33 scenes.

To compare with other methods on depth estimation from the test set split by Eigen et al. [2], which contains 697 pairs from 29 scenes in the KITTI raw dataset, We use the remaining 32 scenes and sample a subset consisting of 8000 cycles for training. We cap the depth to 0-80 meters with the same crop as in Garg *et al.* [7] during evaluation.

4.3. Results for Depth Estimation

KITTI Split. In Table 1, we compare our results with state-of-the-art approaches [24, 9, 27] categorized by the use of stereo pairs during training and testing. Compared to [9] with the same setting, our models considering both the flow and stereo pairs consistently performs better in all the metric. Note that, we use the same number of training data in training all the models for fair comparisons. With comparing among our variants, adding flow pairs that are jointly learned within the same model with stereo pairs improves the base model (i.e., stereo only) by a significant margin. Further including the proposed 2-warp geometric connections brings additional gains in performance, using either type of the 2-warp operations as in Figure 4.

Eigen Split. In Table 2, we show depth estimation performance compared to state-of-the-art methods [2, 9, 27, 32, 28] on the Eigen split. While existing methods do not have the same setting of using stereo pairs during training/testing, we show that our model significantly improves the performance by adding stereo pairs during testing, in which such data can be obtained as a free resource that only slightly increases the computational cost. Note that, adding flow pairs without 2-warp consistency does not significantly improve the performance here in this Eigen split. The potential cause is due to the nature that flow estimation is considered to be harder than stereo matching, and the Eigen split is much smaller than the KITTI split. Therefore, learning optical flow simultaneously could lead to slower convergence and affect the performance of stereo matching. After we

Table 1. Quantitative evaluation of the depth estimation task on KITTI 2015 stereo set. Our results are capped between 0-80 meters. Our full model includes settings with three types of 2-warp operations from Figure 4 and full-1,2,3 correspond to Figure 4(a), 4(b) and 4(c) respectively. Using stereo pairs during training/testing is also indicated in the table.

Method	Train Stereo	Test Stereo	Lower the better				Higher the better		
			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Wang et al. [24]			0.148	1.187	5.496	0.226	0.812	0.938	0.975
Godard et al. [9]	✓		0.097	0.896	5.093	0.176	0.879	0.962	0.986
Yang et al. [27]	✓		0.099	0.986	6.122	0.194	0.860	0.957	0.986
Godard et al. [9]	✓	✓	0.068	0.835	4.392	0.146	0.942	0.978	0.989
Ours (stereo only)	✓	✓	0.078	0.811	4.700	0.174	0.918	0.965	0.983
Ours (flow + stereo)	✓	✓	0.0653	0.819	4.268	0.151	0.946	**0.979**	**0.990**
Ours (full-1)	✓	✓	0.0631	0.756	4.207	0.147	0.947	**0.979**	**0.990**
Ours (full-2)	✓	✓	**0.0620**	**0.747**	**4.113**	**0.146**	**0.948**	**0.979**	**0.990**
Ours (full-3)	✓	✓	0.0630	0.773	4.195	0.147	0.947	**0.979**	**0.990**

Table 2. Quantitative evaluation of the depth estimation task on the KITTI raw dataset split by Eigen et al. [2]. All results are cropped based on the setting in [7]. Using stereo pairs during training/testing or supervised data is indicated in the table.

Method	Train Stereo	Test Stereo	Super-vised	Lower the better				Higher the better		
				Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen et al. [2]			✓	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Godard et al. [9]	✓			0.114	0.898	4.935	0.206	0.861	0.949	0.976
Yang et al. [27]	✓			0.114	1.074	5.836	0.208	0.856	0.939	0.976
Zhou et al. [32]				0.198	1.836	6.565	0.275	0.718	0.901	0.960
GeoNet [28]				0.153	1.328	5.737	0.232	0.802	0.934	0.972
Ours (stereo only)	✓	✓		0.090	0.844	4.373	0.190	0.900	0.954	0.976
Ours (flow + stereo)	✓	✓		0.094	0.791	4.455	0.188	0.897	0.957	0.978
Ours (full-1)	✓	✓		0.089	0.766	4.369	**0.183**	0.905	**0.959**	**0.979**
Ours (full-2)	✓	✓		0.088	**0.759**	**4.346**	0.184	**0.906**	**0.959**	**0.979**
Ours (full-3)	✓	✓		**0.087**	0.765	4.380	0.184	**0.906**	**0.959**	0.978

advance to include 2-warp consistency objective in our full models, it successfully overcomes the aforementioned issue and improves the performance, as now each stereo pair or temporally adjacent one can contribute multiple times to the same network P by the proposed 2-warp function.

4.4. Results for Flow Estimation

In Table 3, we show our unsupervised flow results compared with state-of-the-art supervised methods [11, 22] and unsupervised approaches [17, 28, 33]. The results suggest that our model without using 2-warp already performs favorably against other unsupervised framework. It demonstrates the benefit of using a single network to jointly learn feature representations shared across two highly co-related tasks (i.e., flow estimation and stereo matching) and help improve both performance. In addition, even optical flow estimation is a harder task, our proposed 2-warp consistency loss is able to encourage the tighter connection across two tasks and thus further boost the performance. We also ob-

serve from the KITTI 2015 dataset that all three variants of our full model achieve similar improvement in a significant margin, in which it shows that our proposed 2-warp consistency loss could benefit the estimation of pixel correspondences regardless the warping directions.

4.5. Results without sharing weights

In order to demonstrate the benefit of using a single network for both flow estimation and stereo matching instead of having separate architectures for each task, we train a model variant of full-2 with untying the weights of both tasks and test it on KITTI 2015, which is denoted as Ours (w/o sharing) in Table 3. We find its performance comparable to our full model in stereo matching but much worse in optical flow estimation. The main reason is that the learning rates for flow and depth networks are now hard to balance without well-tuning, and the performance of optical flow estimation becomes unstable for the 2-warp operation. This shows the advantage of having a single and principled

Table 3. Quantitative evaluation on the optical flow task. EPE means average end-point-error where the postfix "-noc" and "-occ" only accounts for non-occlusion regions and occlusion regions, respectively. Fl means the error rate of the flow map values where one pixel is considered wrong if the EPE is $<$3px or $<$5%.

| Method	Train Stereo	Super-vised	KITTI 2012 train EPE-all	train EPE-noc	KITTI 2015 train EPE-all	train Fl-all	train EPE-noc	train Fl-noc	train EPE-occ	train Fl-occ
Flownet2 [11]		✓	4.09	-	10.06	30.37%	-	-	-	-
Flownet2-CSS [11]		✓	3.55	-	8.94	29.77%	-	-	-	-
PWC-Net [22]		✓	4.14		10.35	33.67%	-	-	-	-
UnFlow-CSS [17]			3.29	**1.26**	8.10	**23.27%**	-	-	-	-
GeoNet [28]			-	-	10.81	-	8.05	-	-	-
DF-net [33]			3.54	-	8.98	26.01%	-	-	-	-
Ours (flow only)			4.29	1.98	9.70	32.77%	5.23	25.89%	26.06	65.08%
Ours (flow + stereo)	✓		2.64	1.45	7.47	28.54%	4.707	22.56%	17.83	56.29%
Ours (w/o sharing)	✓		3.49	1.99	8.78	34.56%	5.33	28.65%	21.38	62.61%
Ours (full-1)	✓		2.59	1.41	**7.021**	27.34%	4.257	21.41%	**17.57**	54.78%
Ours (full-2)	✓		2.61	1.39	7.044	27.73%	**4.229**	21.65%	17.89	55.74%
Ours (full-3)	✓		**2.56**	1.388	7.134	27.13%	4.306	**21.19%**	17.79	**54.09%**
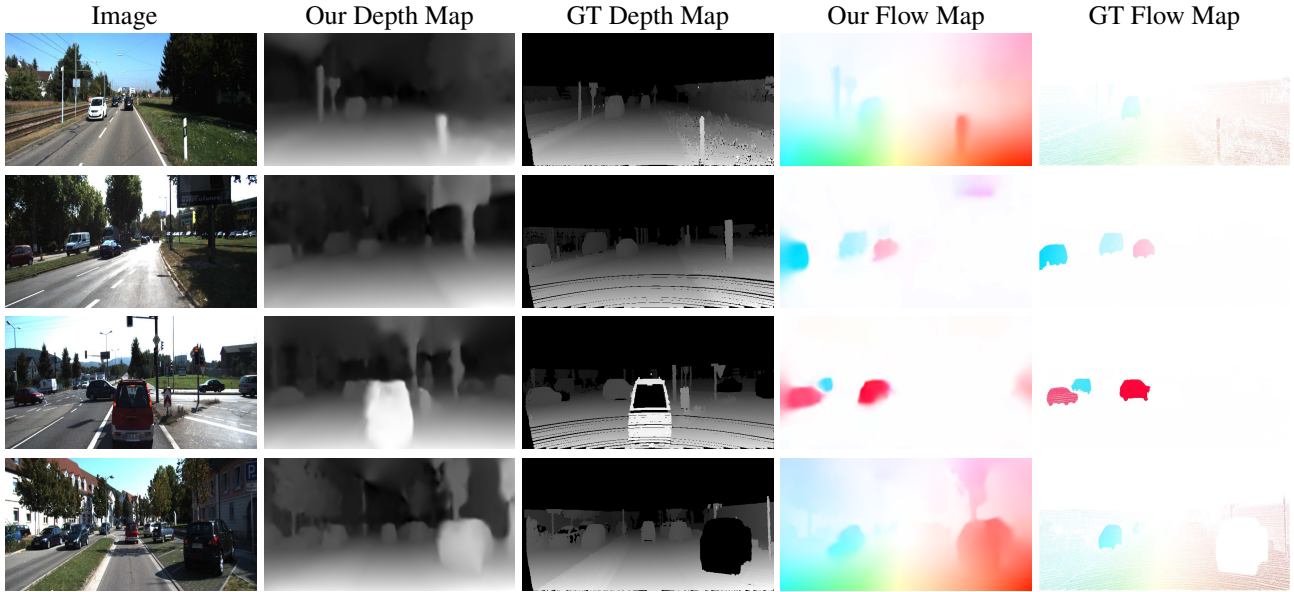
Figure 5. Example results on KITTI. In each row, we sequentially show the left image at time t, our predicted depth map, the ground truth depth, our flow prediction, and the ground truth flow.

network for both tasks. We show some example results in Figure 5.

5. Conclusions

In this paper, we propose to use a single, principled network to perform both stereo matching and flow estimation. The advantage lies in that the feature representations can be jointly learned and shared across two tasks, which all aim to predict pixel correspondences, spatially and temporally. Given a stereo video, we further enforce geometric connections between adjacent stereo pairs, in which a 2-warp consistency term is introduced to optimize the reconstruction loss via the warping functions. Experimental results show that the proposed framework facilitates the information from two tasks and thus improves the performance on both depth and flow estimation.

References

[1] J.-R. Chang and Y.-S. Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[2] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. 1, 2, 6, 7

[3] P. Fischer, A. Dosovitskiy, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 1, 2

[4] D. Fleet and Y. Weiss. Optical flow estimation. In *Handbook of mathematical models in computer vision*, pages 237–257. 2006. 2

[5] D. Fortun, P. Bouthemy, and C. Kervrann. Optical flow modeling and computation: a survey. *Computer Vision and Image Understanding (CVIU)*, 2015. 2

[6] D. Gadot and L. Wolf. Patchbatch: A batch augmented loss for optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[7] R. Garg, B. V. Kumar, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 6, 7

[8] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2, 6

[9] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3, 5, 6, 7

[10] F. Güney and A. Geiger. Deep discrete flow. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2016. 1

[11] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 7, 8

[12] J. Janai, F. Güney, A. Ranjan, M. J. Black, and A. Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *European Conference on Computer Vision (ECCV)*, volume Lecture Notes in Computer Science, vol 11220, pages 713–731. Springer, Cham, Sept. 2018. 2, 3

[13] L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–96, 2014. 1

[14] F. Liu, C. Shen, G. Lin, and I. D. Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016. 1, 2

[15] W. Luo, A. G. Schwing, and R. Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[16] Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, and L. Lin. Single View Stereo Matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[17] S. Meister, J. Hur, and S. Roth. UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 1, 2, 4, 7, 8

[18] M. Menze, C. Heipke, and A. Geiger. Joint 3d estimation of vehicles and scene flow. In *ISPRS Workshop on Image Sequence Analysis (ISA)*, 2015. 2

[19] A. Ranjan and M. J. Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[20] A. Ranjan, V. Jampani, K. Kim, D. Sun, J. Wulff, and M. J. Black. Competitive Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation. *ArXiv:1805.09806*, 2018. 1

[21] Z. Ren, J. Yan, B. Ni, B. Liu, X. Yang, and H. Zha. Unsupervised deep learning for optical flow estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017. 2

[22] D. Sun, X. Yang, M.-Y. Liu, and J. 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 (CVPR)*, 2018. 7, 8

[23] N. Sundaram, T. Brox, and K. Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010. 4

[24] C. Wang, J. Miguel Buenaposada, R. Zhu, and S. Lucey. Learning depth from monocular videos using direct methods. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 6, 7

[25] Y. Wang, Y. Yang, Z. Yang, L. Zhao, P. Wang, and W. Xu. Occlusion Aware Unsupervised Learning of Optical Flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 4, 5

[26] Y. Wang, Z. Yang, P. Wang, Y. Yang, C. Luo, and W. Xu. Joint Unsupervised Learning of Optical Flow and Depth by Watching Stereo Videos. 2018. 5

[27] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia. Every Pixel Counts: Unsupervised Geometry Learning with Holistic 3D Motion Understanding. *ArXiv:1806.10556*, 2018. 1, 3, 6, 7

[28] Z. Yin and J. Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 3, 6, 7, 8

[29] J. J. Yu, A. W. Harley, and K. G. Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *ECCV Workshops*, 2016. 2

[30] J. Zbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research (JMLR)*, 2016. 2

[31] C. Zhou, H. Zhang, X. Shen, and J. Jia. Unsupervised learning of stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2

[32] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised Learning of Depth and Ego-Motion from Video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3, 6, 7

[33] Y. Zou, Z. Luo, and J.-B. Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 3, 4, 7, 8