DF-Net: Unsupervised Joint Learning of Depth and Flow using Cross-Task Consistency

Yuliang Zou[1][0000−0002−8374−6662], Zelun Luo[2][0000−0003−3597−5046], and Jia-Bin Huang[1][0000−0002−0536−3658]

[1]Virginia Tech [2]Stanford University

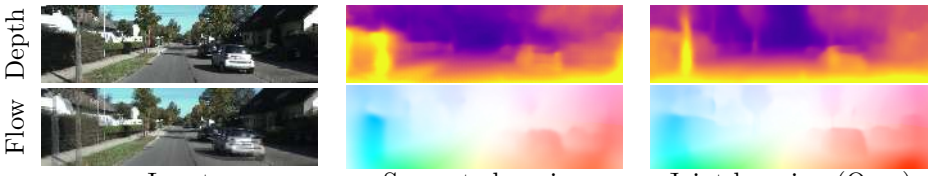


Fig. 1: **Joint learning v.s. separate learning.** Single-view depth prediction and optical flow estimation are two highly correlated tasks. Existing work, however, often addresses these two tasks in isolation. In this paper, we propose a novel cross-task consistency loss to couple the training of these two problems using unlabeled monocular videos. Through enforcing the underlying geometric constraints, we show substantially improved results for both tasks.

Abstract. We present an unsupervised learning framework for simultaneously training single-view depth prediction and optical flow estimation models using unlabeled video sequences. Existing unsupervised methods often exploit brightness constancy and spatial smoothness priors to train depth or flow models. In this paper, we propose to leverage geometric consistency as additional supervisory signals. Our core idea is that for rigid regions we can use the predicted scene depth and camera motion to synthesize 2D optical flow by backprojecting the induced 3D scene flow. The discrepancy between the rigid flow (from depth prediction and camera motion) and the estimated flow (from optical flow model) allows us to impose a cross-task consistency loss. While all the networks are jointly optimized during training, they can be applied independently at test time. Extensive experiments demonstrate that our depth and flow models compare favorably with state-of-the-art unsupervised methods.

1 Introduction

Single-view depth prediction and optical flow estimation are two fundamental problems in computer vision. While the two tasks aim to recover highly correlated information from the scene (i.e., the scene structure and the dense motion field between consecutive frames), existing efforts typically study each problem in isolation. In this paper, we demonstrate the benefits of exploring the geometric

Pixelwise ground truth Unlabeled video sequences (ours)

Fig. 2: **Supervised v.s. unsupervised learning.** Supervised learning of depth or flow networks requires large amount of training data with pixelwise ground truth annotations, which are difficult to acquire in real scenes. In contrast, our work leverages the readily available *unlabeled* video sequences to jointly train the depth and flow models.

relationship between depth, camera motion, and flow for unsupervised learning of depth and flow estimation models.

With the rapid development of deep convolutional neural networks (CNNs), numerous approaches have been proposed to tackle dense prediction problems in an end-to-end manner. However, supervised training CNN for such tasks often involves in constructing large-scale, diverse datasets with dense pixelwise ground truth labels. Collecting such densely labeled datasets in real-world requires significant amounts of human efforts and is prone to error. Existing efforts of RGB-D dataset construction [18,45,53,54] often have limited scope (e.g., in terms of locations, scenes, and objects), and hence are lack of diversity. For optical flow, dense motion annotations are even more difficult to acquire [37]. Consequently, existing CNN-based methods rely on synthetic datasets for training the models [5,12,16,24]. These synthetic datasets, however, do not capture the complexity of motion blur, occlusion, and natural image statistics from real scenes. The trained models usually do not generalize well to unseen scenes without fine-tuning on sufficient ground truth data in a new visual domain.

Several work [17,21,28] have been proposed to capitalize on large-scale real-world videos to train the CNNs in the unsupervised setting. The main idea lies to exploit the brightness constancy and spatial smoothness assumptions of flow fields or disparity maps as supervisory signals. These assumptions, however, often do not hold at motion boundaries and hence makes the training unstable.

Many recent efforts [59,60,65,73] explore the geometric relationship between the two problems. With the estimated depth and camera pose, these methods can produce dense optical flow by backprojecting the 3D scene flow induced from camera ego-motion. However, these methods implicitly assume *perfect* depth and camera pose estimation when "synthesizing" the optical flow. The errors in either depth or camera pose estimation inevitably produce inaccurate flow predictions.

In this paper, we present a technique for *jointly* learning a single-view depth estimation model and a flow prediction model using unlabeled videos as shown in Figure 2. Our key observation is that the predictions from depth, pose, and optical flow should be *consistent* with each other. By exploiting this geometry

cue, we present a novel cross-task consistency loss that provides additional supervisory signals for training both networks. We validate the effectiveness of the proposed approach through extensive experiments on several benchmark datasets. Experimental results show that our joint training method significantly improves the performance of both models (Figure 1). The proposed depth and flow models compare favorably with state-of-the-art unsupervised methods.

We make the following contributions. (1) We propose an unsupervised learning framework to *simultaneously* train a depth prediction network and an optical flow network. We achieve this by introducing a cross-task consistency loss that enforces geometric consistency. (2) We show that through the proposed unsupervised training our depth and flow models compare favorably with existing unsupervised algorithms and achieve competitive performance with supervised methods on several benchmark datasets. (3) We release the source code and pre-trained models to facilitate future research: http://yuliang.vision/DF-Net/

2 Related Work

Supervised learning of depth and flow. Supervised learning using CNNs has emerged to be an effective approach for depth and flow estimation to avoid hand-crafted objective functions and computationally expensive optimization at test time. The availability of RGB-D datasets and deep learning leads to a line of work on single-view depth estimation [13,14,35,38,62,72]. While promising results have been shown, these methods rely on the *absolute* ground truth depth maps. These depth maps, however, are expensive and difficult to collect. Some efforts [8,74] have been made to relax the difficulty of collecting absolute depth by exploring learning from *relative/ordinal* depth annotations. Recent work also explores gathering training datasets from web videos [7] or Internet photos [36] using structure-from-motion and multi-view stereo algorithms.

Compared to ground truth depth datasets, constructing optical flow datasets of diverse scenes in real-world is even more challenging. Consequently, existing approaches [12,26,47] typically rely on synthetic datasets [5,12] for training. Due to the limited scalability of constructing diverse, high-quality training data, fully supervised approaches often require fine-tuning on sufficient ground truth labels in new visual domains to perform well. In contrast, our approach leverages the readily available real-world videos to jointly train the depth and flow models. The ability to learn from unlabeled data enables unsupervised pre-training for domains with limited amounts of ground truth data.

Self-supervised learning of depth and flow. To alleviate the dependency on large-scale annotated datasets, several works have been proposed to exploit the classical assumptions of brightness constancy and spatial smoothness on the disparity map or the flow field [17,21,28,43,71]. The core idea is to treat the estimated depth and flow as latent layers and use them to differentiably warp the source frame to the target frame, where the source and target frames can either be the stereo pair or two consecutive frames in a video sequence. A

photometric loss between the synthesized frame and the target frame can then serve as an unsupervised proxy loss to train the network. Using photometric loss alone, however, is not sufficient due to the ambiguity on textureless regions and occlusion boundaries. Hence, the network training is often unstable and requires careful hyper-parameter tuning of the loss functions. Our approach builds upon existing unsupervised losses for training our depth and flow networks. We show that the proposed cross-task consistency loss provides a sizable performance boost over individually trained models.

Methods exploiting geometry cues. Recently, a number of work exploits the geometric relationship between depth, camera pose, and flow for learning depth or flow models [60,65,68,73]. These methods first estimate the depth of the input images. Together with the estimated camera poses between two consecutive frames, these methods "synthesize" the flow field of rigid regions. The synthesized flow from depth and pose can either be used for flow prediction in rigid regions [60,65,68,48] as is or used for view synthesis to train depth model using monocular videos [73]. Additional cues such as surface normal [67], edge [66], physical constraints [59] can be incorporated to further improve the performance.

These approaches exploit the inherent geometric relationship between structure and motion. However, the errors produced by either the depth or the camera pose estimation propagate to flow predictions. Our key insight is that for rigid regions the estimated flow (from flow prediction network) and the synthesized rigid flow (from depth and camera pose networks) should be consistent. Consequently, coupled training allows both depth and flow networks to learn from each other and enforce geometrically consistent predictions of the scene.

Structure from motion. Joint estimation of structure and camera pose from multiple images of a given scene is a long-standing problem [46,15,64]. Conventional methods can recover (semi-)dense depth estimation and camera pose through keypoint tracking/matching. The outputs of these algorithms can potentially be used to help train a flow network, but not the other way around. Our work differs as we are also interested in learning a depth network to recover dense structure from a single input image.

Multi-task learning. Simultaneously addressing multiple tasks through multi-task learning [52] has shown advantages over methods that tackle individual ones [70]. For examples, joint learning of video segmentation and optical flow through layered models [6,56] or feature sharing [9] helps improve accuracy at motion boundaries. Single-view depth model learning can also benefit from joint training with surface normal estimation [35,67] or semantic segmentation [13,30].

Our approach tackles the problems of learning both depth and flow models. Unlike existing multi-task learning methods that often require *direct supervision* using ground truth training data for each task, our approach instead leverage *meta-supervision* to couple the training of depth and flow models. While our models are jointly trained, they can be applied independently at test time.

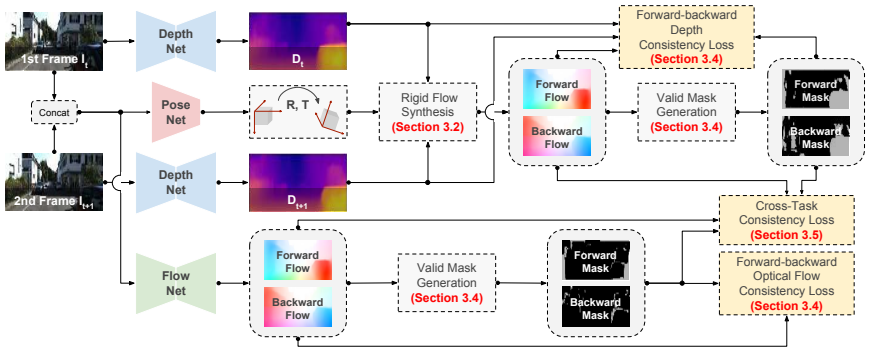


Fig. 3: **Overview of our unsupervised joint learning framework.** Our framework consists of three major modules: (1) a *Depth Net* for single-view depth estimation; (2) a *Pose Net* that takes two stacked input frames and estimates the relative camera pose between the two input frames; and (3) a *Flow Net* that estimates dense optical flow field between the two input frames. Given a pair of input images $\mathbf{I}_t$ and $\mathbf{I}_{t+1}$ sampled from an unlabeled video, we first estimate the depth of each frame, the 6D camera pose, and the dense forward and backward flows. Using the predicted scene depth and the estimated camera pose, we can synthesize 2D forward and backward optical flows (referred as *rigid flow*) by backprojecting the induced 3D forward and backward scene flows (Section 3.2). As we do not have ground truth depth and flow maps for supervision, we leverage standard photometric and spatial smoothness costs to regularize the network training (Section 3.3, not shown in this figure for clarity). To enforce the consistency of flow and depth prediction in both directions, we exploit the forward-backward consistency (Section 3.4), and adopt the valid masks derived from it to filter out invalid regions (e.g., occlusion/dis-occlusion) for the photometric loss. Finally, we propose a novel cross-network consistency loss (Section 3.5) — encouraging the optical flow estimation (from the *Flow Net*) and the rigid flow (from the *Depth and Pose Net*) to be consistent to each other within in valid regions.

3 Unsupervised Joint Learning of Depth and Flow

3.1 Method overview

Our goal is to develop an unsupervised learning framework for *jointly* training the single-view depth estimation network and the optical flow prediction network using *unlabeled* video sequences. Figure 3 shows the high-level sketch of our proposed approach. Given two consecutive frames (I_t, I_{t+1}) sampled from an unlabeled video, we first estimate depth of frame I_t and I_{t+1}, and forward-backward optical flow fields between frame I_t and I_{t+1}. We then estimate the 6D camera pose transformation between the two frames (I_t, I_{t+1}).

With the predicted depth map and the estimated 6D camera pose, we can produce the 3D scene flow induced from camera ego-motion and backproject

them onto the image plane to synthesize the 2D flow (Section 3.2). We refer this synthesized flow as *rigid flow*. Suppose the scenes are mostly static, the synthesized rigid flow should be consistent with the results from the estimated optical flow (produced by the optical flow prediction model). However, the prediction results from the two branches may not be consistent with each other. Our intuition is that the discrepancy between the rigid flow and the estimated flow provides additional supervisory signals for both networks. Hence, we propose a *cross-task consistency loss* to enforce this constraint (Section 3.5). To handle non-rigid transformations that cannot be explained by the camera motion and occlusion-disocclusion regions, we exploit the forward-backward consistency check to identify valid regions (Section 3.4). We avoid enforcing the cross-task consistency for those forward-backward inconsistent regions.

Our overall objective function can be formulated as follows:

$$L = L_{\text{photometric}} + \lambda_s L_{\text{smooth}} + \lambda_f L_{\text{forward-backward}} + \lambda_c L_{\text{cross}}. \tag{1}$$

All of the four loss terms are applied to both depth and flow networks. Also, all of the four loss terms are symmetric for forward and backward directions, for simplicity we only derive them for the forward direction.

3.2 Flow synthesis using depth and pose predictions

Given the two input frames I_t and I_{t+1}, the predicted depth map $\hat{D}_t$, and relative camera pose $\hat{T}_{t\to t+1}$, here we wish to establish the dense pixel correspondence between the two frames. Let p_t denotes the 2D homogeneous coordinate of an pixel in frame I_t and K denotes the intrinsic camera matrix. We can compute the corresponding point of p_t in frame I_{t+1} using the equation [73]:

$$p_{t+1} = K\hat{T}_{t\to t+1}\hat{D}_t(p_t)K^{-1}p_t. \tag{2}$$

We can then obtain the synthesized forward rigid flow at pixel p_t in I_t by

$$F_{\text{rigid}}(p_t) = p_{t+1} - p_t \tag{3}$$

3.3 Brightness constancy and spatial smoothness priors

Here we briefly review two loss functions that we used in our framework to regularize network training. Leveraging the brightness constancy and spatial smoothness priors used in classical dense correspondence algorithms [4,23,40], prior work has used the photometric discrepancy between the warped frame and the target frame as an unsupervised proxy loss function for training CNNs without ground truth annotations.

Photometric loss. Suppose that we have frame I_t and I_{t+1}, as well as the estimated flow $F_{t\to t+1}$ (either from the optical flow predicted from the flow model or the synthesized rigid flow induced from the estimated depth and camera pose), we can produce the warped frame $\bar{I}_t$ with the inverse warping from frame

I_{t+1}. Note that the projected image coordinates p_{t+1} might not lie exactly on the image pixel grid, we thus apply a differentiable bilinear interpolation strategy used in the spatial transformer networks [27] to perform frame synthesis.

With the warped frame $\bar{I}_t$ from I_{t+1}, we formulate the brightness constancy objective function as

$$L_{\text{photometric}} = \sum_{p} \rho\big(I_t(p), \bar{I}_t(p)\big). \tag{4}$$

where $\rho(\cdot)$ is a function to measure the difference between pixel values. Previous work simply choose L_1 norm or the appearance matching loss [21], which is not invariant to illumination changes in real-world scenarios [61]. Here we adopt the ternary census transform based loss [43,55,69] that can better handle complex illumination changes.

Smoothness loss. The brightness constancy loss is not informative in low-texture or homogeneous region of the scene. To handle this issue, existing work incorporates a smoothness prior to regularize the estimated disparity map or flow field. We adopt the spatial smoothness loss as proposed in [21].

3.4 Forward-backward consistency

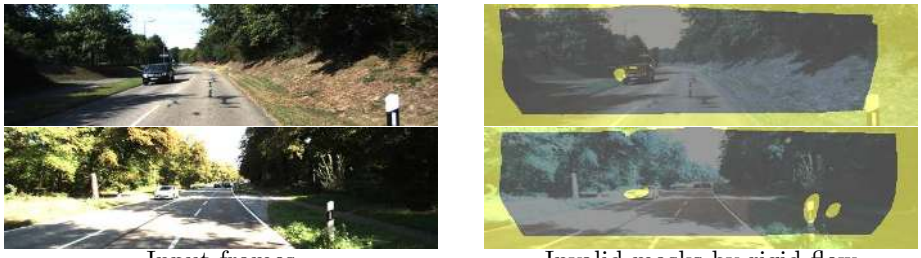
According to the brightness constancy assumption, the warped frame should be similar to the target frame. However, the assumption does not hold for occluded and dis-occluded regions. We address this problem by using the commonly used forward-backward consistency check technique to identify invalid regions and do not impose the photometric loss on those regions.

Valid masks. We implement the occlusion detection based on forward-backward consistency assumption [58] (i.e., traversing flow vector forward and then backward should arrive at the same position). Here we use a simple criterion proposed in [43]. We mark pixels as invalid whenever this constraint is violated. Figure 4 shows two examples of the marked invalid regions by forward-backward consistency check using the synthesized rigid flow (animations can be viewed in Adobe Reader).

Denote the valid region by V (either from rigid flow or estimated flow), we can modify the photometric loss term (4) as

$$L_{\text{photometric}} = \sum_{p \in V} \rho\big(I_t(p), \bar{I}_t(p)\big). \tag{5}$$

Forward-backward consistency loss. In addition to using forward-backward consistency check for identifying invalid regions, we can further impose constraints on the valid regions so that the network can produce consistent predictions for both forward and backward directions. Similar ideas have been exploited in [25,43] for occlusion-aware flow estimation. Here, we apply the forward-backward consistency loss to both flow and depth predictions.



Input frames Invalid masks by rigid flow

Fig. 4: **Valid mask visualization.** We estimate the invalid mask by checking the forward-backward consistency from the synthesized rigid flow, which can not only detect occluded regions, but also identify the moving objects (cars) as they cannot be explained by the estimated depth and pose. Animations can be viewed in Adobe Reader.

For flow prediction, the forward-backward consistency loss is of the form:

$$L_{\text{forward-backward, flow}} = \sum_{p \in V_{\text{flow}}} ||F_{t \to t+1}(p) + F_{t+1 \to t}(p + F_{t \to t+1}(p))| \, |_1 \quad (6)$$

Similarly, we impose a consistency penalty for depth:

$$L_{\text{forward-backward, depth}} = \sum_{p \in V_{\text{depth}}} ||D_t(p) - \bar{D}_t(p)||_1 \quad (7)$$

where $\bar{D}_t$ is warped from D_{t+1} using the synthesized rigid flow from t to $t+1$.

While we exploit robust functions for enforcing photometric loss, forward-backward consistency for each of the tasks, the training of depth and flow networks using unlabeled data remains non-trivial and sensitive to the choice of hyper-parameters [33]. Building upon the existing loss functions, in the following we introduce a novel cross-task consistency loss to further regularize the network training.

3.5 Cross-task consistency

In Section 3.2, we show that the motion of rigid regions in the scene can be explained by the ego-motion of the camera and the corresponding scene depth. On the one hand, we can estimate the rigid flow by backprojecting the induced 3D scene flow from the estimated depth and relative camera pose. On the other hand, we have direct estimation results from an optical flow network. Our core idea is the that these two flow fields should be consistent with each other for non-occluded and static regions. Minimizing the discrepancy between the two flow fields allows us to simultaneously update the depth and flow models.

We thus propose to minimize the endpoint distance between the flow vectors in the rigid flow (computed from the estimated depth and pose) and that in

the estimated flow (computed from the flow prediction model). We denote the synthesized rigid flow as $F_{\text{rigid}} = (u_{\text{rigid}}, v_{\text{rigid}})$ and the estimated flow as $F_{\text{flow}} = (u_{\text{flow}}, v_{\text{flow}})$. Using the computed valid masks (Section 3.4), we impose the cross-task consistency constraints over valid pixels.

$$L_{\text{cross}} = \sum_{p \in V_{\text{depth}} \cap V_{\text{flow}}} ||F_{\text{rigid}}(p) - F_{\text{flow}}(p)||_1 \tag{8}$$

4 Experimental Results

In this section, we validate the effectiveness of our proposed method for unsupervised learning of depth and flow on several standard benchmark datasets. More results can be found in the supplementary material. Our source code and pre-trained models are available on http://yuliang.vision/DF-Net/.

4.1 Datasets

Datasets for joint network training. We use video clips from the train split of KITTI raw dataset [18] for joint learning of depth and flow models. Note that our training does not involve any depth/flow labels.

Datasets for pre-training. To avoid the joint training process converging to trivial solutions, we (unsupervisedly) pre-train the flow network on the SYNTHIA dataset [51]. For pre-training both depth and pose networks, we use either KITTI raw dataset or the CityScapes dataset [11] .

The SYNTHIA dataset [51] contains multi-view frames captured by driving vehicles in different scenarios and traffic conditions. We take all the four-view images of the left camera from all summer and winter driving sequences, which contains around 37K image pairs. The CityScapes dataset [11] contains real-world driving sequences, we follow Zhou et al. [73] and pre-process the dataset to generate around 75K training image pairs.

Datasets for evaluation. For evaluating the performance of our depth network, we use the *test split* of the KITTI raw dataset. The depth maps for KITTI raw are sampled at irregularly spaced positions, captured using a rotating LIDAR scanner. Following the standard evaluation protocol, we evaluate the performance using only the regions with ground truth depth samples (bottom parts of the images). We also evaluate the generalization of our depth network on general scenes using the Make3D dataset [53].

For evaluating our flow network, we use the challenging KITTI flow 2012 [19] and KITTI flow 2015 [44] datasets. The ground truth optical flow is obtained from a 3D laser scanner and thus only covers about 50% of the pixels.

4.2 Implementation details

We implement our approach in TensorFlow [1] and conduct all the experiments on a single Tesla K80 GPU with 12GB memory. We set $\lambda_s = 3.0$, $\lambda_f = 0.2$, and

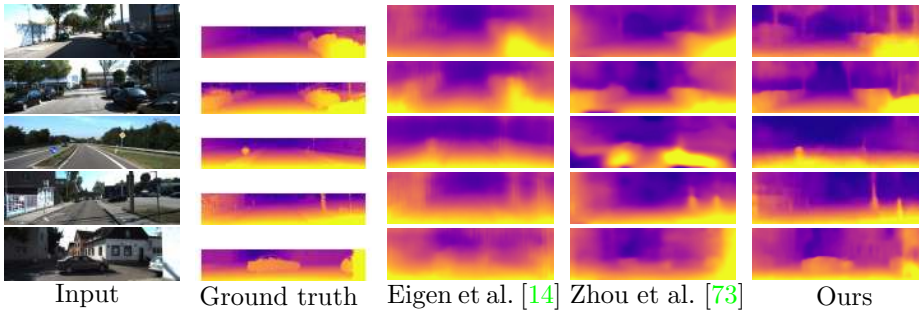


Fig. 5: **Sample results on KITTI raw test set.** The ground truth depth is interpolated from sparse point cloud for visualization only. Compared to Zhou et al. [73] and Eigen et al. [14], our method can better capture object contour and thin structures.

$\lambda_c = 0.2$. For network training, we use the Adam optimizer [31] with $\beta_1 = 0.9$, $\beta_2 = 0.99$. In the following, we provide more implementation details in network architecture, network pre-training, and the proposed unsupervised joint training.

Network architecture. For the pose network, we adopt the architecture from Zhou et al. [73]. For the depth network, we use the ResNet-50 [22] as our feature backbone with ELU [10] activation functions. For the flow network, we adopt the UnFlow-C structure [43] — a variant of FlowNetC [12]. As our network training is *model-agnostic*, more advanced network architectures (e.g., pose [20], depth [36], or flow [57]) can be used for further improving the performance.

Unsupervised depth pre-training. We train the depth and pose networks with a mini-batch size of 6 image pairs whose size is 576×160, from KITTI raw dataset or CityScapes dataset for 100K iterations. We use a learning rate is 2e-4. Each iteration takes around 0.8s (forward and backprop) during training.

Unsupervised flow pre-training. Following Meister et al. [43], we train the flow network with a mini-batch size of 4 image pairs whose size is 1152×320 from SYNTHIA dataset for 300K iterations. We keep the initial learning rate as 1e-4 for the first 100K iterations and then reduce the learning rate by half after each 100K iterations. Each iteration takes around 2.4s (forward and backprop).

Unsupervised joint training. We jointly train the depth, pose, and flow networks with a mini-batch size of 4 image pairs from KITTI raw dataset for 100K iterations. Input size for the depth and pose networks is 576×160, while the input size for the flow network is 1152×320. We divide the initial learning rate by 2 for every 20K iterations. Our depth network produces depth predictions at 4 spatial scales, while the flow network produces flow fields at 5 scales. We enforce the cross-network consistency in the finest 4 scales. Each iteration takes around 3.6s (forward and backprop) during training.

Table 1: **Single-view depth estimation results** on *test split* of KITTI raw dataset [18]. The methods trained on KITTI raw dataset [18] are denoted by K. Models with additional training data from CityScapes [11] are denoted by CS+K. (D) denotes depth supervision, (B) denotes stereo input pairs, (M) denotes monocular video clips . The best and the second best performance in each block are highlighted as bold and underline.

		Error metric ↓				Accuracy metric ↑		
Method	Dataset	Abs Rel	Sq Rel	RMSE	log RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen et al. [14]	K (D)	0.203	1.548	6.307	0.246	0.702	0.890	0.958
Kuznietsov et al. [32]	K (B) / K (D)	**0.113**	**0.741**	**4.621**	**0.189**	**0.862**	**0.960**	**0.986**
Zhan et al. [71]	K (B)	0.144	1.391	5.869	0.241	0.803	0.928	0.969
Godard et al. [21]	K (B)	0.133	1.140	5.527	0.229	0.830	0.936	0.970
Godard et al. [21]	CS+K (B)	0.121	1.032	5.200	0.215	0.854	0.944	0.973
Zhou et al. [73]	K (M)	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Yang et al. [67]	K (M)	0.182	1.481	6.501	0.267	0.725	0.906	0.963
Mahjourian et al. [41]	K (M)	0.163	1.240	6.220	0.250	0.762	0.916	0.968
Yang et al. [66]	K (M)	0.162	1.352	6.276	0.252	-	-	-
Yin et al. [68]	K (M)	0.155	1.296	5.857	0.233	0.793	0.931	**0.973**
Godard et al. [20]	K (M)	0.154	1.218	5.699	0.231	0.798	0.932	**0.973**
Ours (w/o forward-backward)	K (M)	0.160	1.256	5.555	0.226	0.796	0.931	**0.973**
Ours (w/o cross-task)	K (M)	0.160	1.234	5.508	0.225	0.800	0.932	0.972
Ours	K (M)	**0.150**	**1.124**	**5.507**	**0.223**	**0.806**	**0.933**	**0.973**
Zhou et al. [73]	CS+K (M)	0.198	1.836	6.565	0.275	0.718	0.901	0.960
Yang et al. [67]	CS+K (M)	0.165	1.360	6.641	0.248	0.750	0.914	0.969
Mahjourian et al. [41]	CS+K (M)	0.159	1.231	5.912	0.243	0.784	0.923	0.970
Yang et al. [66]	CS+K (M)	0.159	1.345	6.254	0.247	-	-	-
Yin et al. [68]	CS+K (M)	0.153	1.328	5.737	0.232	0.802	0.934	0.972
Ours (w/o forward-backward)	CS+K (M)	0.159	1.716	5.616	0.222	0.805	0.939	0.976
Ours (w/o cross-task)	CS+K (M)	0.155	**1.181**	5.301	0.218	0.805	0.939	0.977
Ours	CS+K (M)	**0.146**	1.182	**5.215**	**0.213**	**0.818**	**0.943**	**0.978**

Image resolution of network inputs/outputs. As the input size of the UnFlow-C network [43] must be divisible by 64, we resize input image pairs of the two KITTI flow datasets to 1280×384 using bilinear interpolation. We then resize the estimated optical flow and rescale the predicted flow vectors to match the original input size. For depth estimation, we resize the input image to the same size of training input to predict the disparity first. We then resize and rescale the predicted disparity to the original size and compute the inverse the obtain the final prediction.

4.3 Evaluation metrics

Following Zhou et al. [73], we evaluate our depth network using several error metrics (absolute relative difference, square related difference, RMSE, log RMSE). For optical flow estimation, we compute the average endpoint error (EPE) on pixels with the ground truth flow available for each dataset. On KITTI flow 2015 dataset [44], we also compute the F1 score, which is the percentage of pixels that have EPE greater than 3 pixels and 5% of the ground truth value.

4.4 Experimental evaluation

Single-view depth estimation. We compare our depth network with state-of-the-art algorithms on the *test split* of the KITTI raw dataset provided by Eigen et al. [14]. As shown in Table 1, our method achieves the state-of-the-art performance when compared with models trained with monocular video sequences. However, our method performs slightly worse than the models that exploit calibrated stereo image pairs (i.e., pose supervision) or with additional ground truth depth annotation. We believe that performance gap can be attributed to the error induced by our pose network. Extending our approach to *calibrated stereo videos* is an interesting future direction.

We also conduct an ablation study by removing the forward-backward consistency loss or cross-task consistency loss. In both cases our results show significant performance of degradation, highlighting the importance the proposed consistency loss. Figure 5 shows qualitative comparison with [14,73], our method can better capture thin structure and delineate clear object contour.

To evaluate the generalization ability of our depth network on general scenes, we also apply our trained model to the Make3D dataset [53]. Table 2 shows that our method achieves the state-of-the-art performance compared with existing unsupervised models and is competitive with respect to supervised learning models (even without fine-tuning on Make3D datasets).

Table 2: **Results on the Make3D dataset** [54]. Our results were obtained by the model trained on Cityscapes + KITTI *without* fine-tuning on the training images in Make3D. Following the evaluation protocol of [21], the errors are only computed where depth is less than 70 meters. The best and the second best performance in each block are highlighted as bold and underline.

		Error metric ↓			
Method	Supervision	Abs Rel	Sq Rel	RMSE	log RMSE
Train set mean	-	0.876	12.98	12.27	0.307
Karsch et al. [29]	depth	0.428	<u>5.079</u>	8.389	0.149
Liu et al. [39]	depth	0.475	6.562	10.05	0.165
Laina et al. [34]	depth	<u>0.204</u>	**1.840**	<u>5.683</u>	<u>0.084</u>
Li et al. [36]	depth	**0.176**	-	**4.260**	**0.069**
Godard et al. [21]	pose	0.544	10.94	11.76	**0.193**
Zhou et al. [73]	none	<u>0.383</u>	<u>5.321</u>	<u>10.47</u>	0.478
Ours	none	**0.331**	**2.698**	**6.89**	<u>0.416</u>

Optical flow estimation. We compare our flow network with conventional variational algorithms, supervised CNN methods, and several unsupervised CNN models on the KITTI flow 2012 and 2015 datasets. As shown in Table 3, our

Table 3: **Quantitative evaluation on optical flow.** Results on KITTI flow 2012 [19] , KITTI flow 2015 [44] datasets. We denote "C" as the FlyingChairs dataset [12], "T" as the FlyingThings3D dataset [42], "K" as the KITTI raw dataset [18], "SYN" as the SYNTHIA dataset [51]. (S) indicates that the model is trained with ground truth annotation, while (U) indicates the model is trained in an unsupervised manner. The best and the second best performance in each block are highlighted as bold and underline.

		KITTI 2012		KITTI 2015		
Method	Dataset	Train EPE	Test EPE	Train EPE	Train F1	Test F1
LDOF [3]	-	10.94	12.4	18.19	38.05%	-
DeepFlow [63]	-	4.58	5.8	10.63	26.52%	29.18%
EpicFlow [50]	-	3.47	**3.8**	9.27	27.18%	**27.10%**
FlowField [2]	-	**3.33**	-	**8.33**	**24.43%**	-
FlowNetS [12]	C (S)	8.26	-	15.44	52.86%	-
FlowNetC [12]	C (S)	9.35	-	12.52	47.93%	-
SpyNet [47]	C (S)	9.12	-	20.56	44.78%	-
SemiFlowGAN [33]	C (S) / K (U)	7.16	-	16.02	38.77%	-
FlowNet2 [26]	C (S) + T (S)	**4.09**	-	**10.06**	**30.37%**	-
UnsupFlownet [28]	C (U) + K (U)	11.3	9.9	-	-	-
DSTFlow [49]	C (U)	16.98	-	24.30	52.00%	-
DSTFlow [49]	K (U)	10.43	12.4	16.79	36.00%	39.00%
Yin et al. [68]	K (U)	-	-	10.81	-	-
UnFlowC [43]	SYN (U) + K (U)	3.78	4.5	**8.80**	28.94%	29.46%
Ours (w/o forward-backward)	SYN (U) + K (U)	3.86	4.7	9.12	26.27%	26.90%
Ours (w/o cross-task)	SYN (U) + K (U)	4.70	5.8	8.95	28.37%	30.03%
Ours	SYN (U) + K (U)	**3.54**	**4.4**	8.98	**26.01%**	**25.70%**
FlowNet2-ft-kitti [26]	C (S) + T (S) + K (S)	(1.28)	1.8	(2.30)	(8.61%)	11.48%
UnFlowCSS-ft-kitti [43]	SYN (U) + K (U) + K (S)	**(1.14)**	**1.7**	**(1.86)**	**(7.40%)**	**11.11%**
UnFlowC-ft-kitti [43]	SYN (U) + K (U) + K (S)	(2.13)	3.0	(3.67)	(17.78%)	24.20%
Ours-ft-kitti	SYN (U) + K (U) + K (S)	(1.75)	3.0	(2.85)	(13.47%)	22.82%

Table 4: **Pose estimation results** on KITTI Odometry datest [19].

	Seq. 09	Seq. 10
ORB-SLAM (full)	0.014±0.008	0.012±0.011
ORB-SLAM (short)	0.064±0.141	0.064±0.130
Mean Odom.	0.032±0.026	0.028±0.023
Zhou et al. [73]	0.021±0.017	0.020±0.015
Mahjourian et al. [41]	0.013±0.010	0.012±0.011
Yin et al. [68]	**0.012±0.007**	**0.012±0.009**
Ours	0.017±0.007	0.015±0.009

Fig. 6: **Visual results on KITTI flow datasets.** All the models are directly applied *without* fine-tuning on KITTI flow annotations. Our model delineates clearer object contours compared to both supervised/unsupervised methods.

method achieves state-of-the-art performance on both datasets. A visual comparison can be found in Figure 6. With optional fine-tuning on available ground truth labels on the KITTI flow datasets, we show that our approach achieves competitive performance sharing similar network architectures. This suggests that our method can serve as an unsupervised pre-training technique for learning optical flow in domains where the amounts of ground truth data are scarce.

Pose estimation. For completeness, we provide the performance evaluation of the pose network. We follow the same evaluation protocol as [73] and use a 5-frame based pose network. As shown in Table 4, our pose network shows competitive performance with respect to state-of-the-art visual SLAM methods or other unsupervised learning methods. We believe that a better pose network would further improve the performance of both depth or optical flow estimation.

5 Conclusions

We presented an unsupervised learning framework for both sing-view depth prediction and optical flow estimation using unlabeled video sequences. Our key technical contribution lies in the proposed cross-task consistency that couples the network training. At test time, the trained depth and flow models can be applied independently. We validate the benefits of joint training through extensive experiments on benchmark datasets. Our single-view depth prediction model compares favorably against existing unsupervised models using unstructured videos on both KITTI and Make3D datasets. Our flow estimation model achieves competitive performance with state-of-the-art approaches. By leveraging geometric constraints, our work suggests a promising future direction of advancing the state-of-the-art in multiple dense prediction tasks using unlabeled data.

Acknowledgement. This work was supported in part by NSF under Grant No. (#1755785). We thank NVIDIA Corporation for the donation of GPUs.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016) 9
2. Bailer, C., Taetz, B., Stricker, D.: Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation. In: ICCV (2015) 13
3. Brox, T., Bregler, C., Malik, J.: Large displacement optical flow. In: CVPR (2009) 13
4. Bruhn, A., Weickert, J., Schnörr, C.: Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. IJCV **61**(3), 211–231 (2005) 6
5. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: ECCV (2012) 2, 3
6. Chang, J., Fisher, J.W.: Topology-constrained layered tracking with latent flow. In: ICCV (2013) 4
7. Chen, W., Deng, J.: Learning single-image depth from videos using quality assessment networks. In: ECCV (2018) 3
8. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In: NIPS (2016) 3
9. Cheng, J., Tsai, Y.H., Wang, S., Yang, M.H.: SegFlow: Joint learning for video object segmentation and optical flow. In: ICCV (2017) 4
10. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). In: ICLR (2016) 10
11. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016) 9, 11
12. Dosovitskiy, A., Fischer, P., Ilg, E., Häusser, P., Hazırbaş, C., Golkov, V., v.d. Smagt, P., Cremers, D., Brox, T.: FlowNet: Learning optical flow with convolutional networks. In: ICCV (2015) 2, 3, 10, 13
13. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV (2015) 3, 4
14. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: NIPS (2014) 3, 10, 11, 12
15. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Towards internet-scale multi-view stereo. In: CVPR (2010) 4
16. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: CVPR (2016) 2
17. Garg, R., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: ECCV (2016) 2, 3
18. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. IJRR (2013) 2, 9, 11, 13
19. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR (2012) 9, 13
20. Godard, C., Mac Aodha, O., Brostow, G.: Digging into self-supervised monocular depth estimation. arXiv preprint arXiv:1806.01260 (2018) 10, 11
21. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR (2017) 2, 3, 7, 11, 12
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 10

23. Horn, B.K., Schunck, B.G.: Determining optical flow. Artificial intelligence **17**(1-3), 185–203 (1981) 6
24. Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B.: DeepMVS: Learning multi-view stereopsis. In: CVPR (2018) 2
25. Hur, J., Roth, S.: MirrorFlow: Exploiting symmetries in joint optical flow and occlusion estimation. In: ICCV (2017) 7
26. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: CVPR (2017) 3, 13
27. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: NIPS (2015) 7
28. Jason, J.Y., Harley, A.W., Derpanis, K.G.: Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In: ECCV Workshop (2016) 2, 3, 13
29. Karsch, K., Liu, C., Kang, S.B.: Depth transfer: Depth extraction from video using non-parametric sampling. TPAMI **36**(11), 2144–2158 (2014) 12
30. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: NIPS (2017) 4
31. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2014) 10
32. Kuznietsov, Y., Stückler, J., Leibe, B.: Semi-supervised deep learning for monocular depth map prediction. In: CVPR (2017) 11
33. Lai, W.S., Huang, J.B., Yang, M.H.: Semi-supervised learning for optical flow with generative adversarial networks. In: NIPS (2017) 8, 13
34. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 3DV (2016) 12
35. Li, B., Shen, C., Dai, Y., van den Hengel, A., He, M.: Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In: CVPR (2015) 3, 4
36. Li, Z., Snavely, N.: MegaDepth: Learning single-view depth prediction from internet photos. In: CVPR (2018) 3, 10, 12
37. Liu, C., Freeman, W.T., Adelson, E.H., Weiss, Y.: Human-assisted motion annotation. In: CVPR (2008) 2
38. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. In: CVPR (2015) 3
39. Liu, M., Salzmann, M., He, X.: Discrete-continuous depth estimation from a single image. In: CVPR (2014) 12
40. Lucas, B.D., Kanade, T., et al.: An iterative image registration technique with an application to stereo vision. In: IJCAI (1981) 6
41. Mahjourian, R., Wicke, M., Angelova, A.: Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In: CVPR (2018) 11, 13
42. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: CVPR (2016) 13
43. Meister, S., Hur, J., Roth, S.: UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In: AAAI (2018) 3, 7, 10, 11, 13
44. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: CVPR (2015) 9, 11, 13
45. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012) 2

46. Newcombe, R.A., Lovegrove, S.J., Davison, A.J.: DTAM: Dense tracking and mapping in real-time. In: ICCV (2011) 4
47. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: CVPR (2017) 3, 13
48. Ranjan, A., Jampani, V., Kim, K., Sun, D., Wulff, J., Black, M.J.: Adversarial Collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. arXiv preprint arXiv:1805.09806 (2018) 4
49. Ren, Z., Yan, J., Ni, B., Liu, B., Yang, X., Zha, H.: Unsupervised deep learning for optical flow estimation. In: AAAI (2017) 13
50. Revaud, J., Weinzaepfel, P., Harchaoui, Z., Schmid, C.: Epicflow: Edge-preserving interpolation of correspondences for optical flow. In: CVPR (2015) 13
51. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: CVPR (2016) 9, 13
52. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017) 4
53. Saxena, A., Chung, S.H., Ng, A.Y.: Learning depth from single monocular images. In: NIPS (2006) 2, 9, 12
54. Saxena, A., Chung, S.H., Ng, A.Y.: 3-d depth reconstruction from a single still image. IJCV **76**(1), 53–69 (2008) 2, 12
55. Stein, F.: Efficient computation of optical flow using the census transform. In: DAGM (2004) 7
56. Sun, D., Wulff, J., Sudderth, E.B., Pfister, H., Black, M.J.: A fully-connected layered model of foreground and background flow. In: CVPR (2013) 4
57. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: PWC-net: Cnns for optical flow using pyramid, warping, and cost volume. In: CVPR (2018) 10
58. Sundaram, N., Brox, T., Keutzer, K.: Dense point trajectories by gpu-accelerated large displacement optical flow. In: ECCV (2010) 7
59. Tung, H.Y.F., Harley, A., Seto, W., Fragkiadaki, K.: Adversarial Inversion: Inverse graphics with adversarial priors. In: ICCV (2017) 2, 4
60. Vijayanarasimhan, S., Ricco, S., Schmid, C., Sukthankar, R., Fragkiadaki, K.: Sfm-net: Learning of structure and motion from video. arXiv preprint arXiv:1704.07804 (2017) 2, 4
61. Vogel, C., Roth, S., Schindler, K.: An evaluation of data costs for optical flow. In: GCPR (2013) 7
62. Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.L.: Towards unified depth and semantic prediction from a single image. In: CVPR (2015) 3
63. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: DeepFlow: Large displacement optical flow with deep matching. In: ICCV (2013) 13
64. Wu, C.: Visualsfm: A visual structure from motion system (2011) 4
65. Wulff, J., Sevilla-Lara, L., Black, M.J.: Optical flow in mostly rigid scenes. In: CVPR (2017) 2, 4
66. Yang, Z., Wang, P., Wang, Y., Xu, W., Nevatia, R.: LEGO: Learning edge with geometry all at once by watching videos. In: CVPR (2018) 4, 11
67. Yang, Z., Wang, P., Xu, W., Zhao, L., Nevatia, R.: Unsupervised learning of geometry with edge-aware depth-normal consistency. In: AAAI (2018) 4, 11
68. Yin, Z., Shi, J.: GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In: CVPR (2018) 4, 11, 13
69. Zabih, R., Woodfill, J.: Non-parametric local transforms for computing visual correspondence. In: ECCV (1994) 7

70. Zamir, A.R., Sax, A., Shen, W., Guibas, L., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: CVPR (2018) 4
71. Zhan, H., Garg, R., Weerasekera, C.S., Li, K., Agarwal, H., Reid, I.: Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In: CVPR (2018) 3, 11
72. Zhang, Z., Schwing, A.G., Fidler, S., Urtasun, R.: Monocular object instance segmentation and depth ordering with cnns. In: ICCV (2015) 3
73. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: CVPR (2017) 2, 4, 6, 9, 10, 11, 12, 13, 14
74. Zoran, D., Isola, P., Krishnan, D., Freeman, W.T.: Learning ordinal relationships for mid-level vision. In: ICCV (2015) 3