IEEE Robotics and Automation Letters (RAL) paper presented at the
2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Loop-Net: Joint Unsupervised Disparity and Optical Flow Estimation of Stereo Videos with Spatiotemporal Loop Consistency

Taewoo Kim, Kwonyoung Ryu*, Kyeongseob Song* and Kuk-Jin Yoon

***Abstract*— Most of existing deep learning-based depth and optical flow estimation methods require the supervision of a lot of ground truth data, and hardly generalize to video frames, resulting in temporal inconsistency (flickering). In this paper, we propose a joint framework that estimates disparity and optical flow of stereo videos and generalizes across various video frames by considering the spatiotemporal relation between the disparity and flow without supervision. To improve both accuracy and consistency, we propose a loop consistency loss which enforces the spatiotemporal consistency of the estimated disparity and optical flow. Furthermore, we introduce a video-based training scheme using the c-LSTM to reinforce the temporal consistency. Extensive experiments show our proposed methods not only estimate disparity and optical flow accurately but also further improve spatiotemporal consistency. Our framework outperforms the state-of-the-art unsupervised depth and optical flow estimation models on the KITTI benchmark dataset.**

I. INTRODUCTION

Depth and optical flow estimation have been core tasks in numerous robotics applications. With the recent advances in deep learning, deep learning-based approaches [1]–[5] have shown significant performance improvement on depth and optical flow estimation. However, most existing deep learning-based methods rely on supervised learning with ground truth, requiring expensive annotation costs. Also, it is often assumed that the testing data is similar to the training data in supervised learning, which hinders the network from generalizing to the unfamiliar data in the testing phase. To address these issues, several researches have recently proposed the unsupervised learning frameworks [4], [6]–[8]. However, there is still room for improvement for stereo videos. In supervised learning-based frameworks, the network learns the similarity metric to find correspondences using ground truth disparity and flow. On the other hand, in unsupervised learning-based frameworks, instead of learning the similarity metric, it is commonly assumed that the brightness of the correspondence match is consistent. However, since the stereo cameras are not identical in practice and the brightness can vary due to motion and temporal illumination changes, it is not desirable to strictly assume the brightness consistency for the accurate disparity and flow estimation. In addition, since most disparity and flow estimation networks take a single image pair as input, the temporal relations between consecutive image pairs are not leveraged. For that reason, simply extending existing frameworks for stereo videos results in poor generalization ability and temporal inconsistency (flickering). Thus, multiple consecutive image pairs containing rich temporal information need to be provided as input to the network. To resolve aforementioned problems in estimating disparity and optical flow for stereo videos, we design a joint framework for training in consecutive stereo sequence using the convolutional long short-term memory (c-LSTM) and define a new spatiotemporal loop consistency loss on account of the temporal inconsistency, achieving accurate and temporally consistent disparity and flow estimation. Furthermore, we propose the zero mean normalized cross correlation (ZNCC)-based data loss in consideration of brightness difference between the corresponding pixels. We evaluate our framework on both stereo image and video datasets, successfully verifying that our methods help improve accuracy on both datasets and ensure the spatiotemporal consistency. Our main contributions can be summarized as follows: First, we propose a loop consistency loss to reinforce the spatiotemporal consistency. The loop consistency loss exploits spatiotemporal relation between consecutive frames and improves the accuracy. Second, we design a joint framework for unsupervised estimation of disparity and optical flow of stereo videos. The framework makes better use of the temporal information of the previous input frames provided by the c-LSTM. Finally, our framework generalizes to both a single image pair and video, showing state-of-the-art performance in disparity and optical flow estimation on both datasets.

II. RELATED WORK

Traditional stereo depth estimation methods often measure the similarity between two images by exploiting the low-level features of image patches around each pixel to compute disparity maps. Fookes *et al.* [9] compared several similarity measures such as sum of absolute differences(SAD, ZSAD) and normalized cross correlation(NCC, ZNCC) for window-based stereo matching. Meanwhile, optical flow estimation aims to optimize the sum of a data term based on brightness constancy and a regularization term to obtain smooth displacement fields [10]. By coupling the estimated depth and optical flow maps, several early works [11], [12] recover scene flow given a sequence of stereo images.

A. Supervised Depth and Optical Flow Estimation

With the development of deep learning, CNNs have been utilized to compute matching cost between two sampled

*These authors contributed equally.

Taewoo Kim, Kwonyoung Ryu, Kyeongseob Song, and Kuk-Jin Yoon are with Visual Intelligence Lab, Mechanical Engineering, KAIST, Republic of Korea {intelpro, rky912, kssong, kjyoon}@kaist.ac.kr

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2018R1A2B3008640).

patches in a stereo image pair. For accurate depth estimation in ambiguous regions, Kendall *et al.* [2] leverage context information with the encoder-decoder architecture by incorporating cost aggregation and disparity refinement. Chang *et al.* [3] make better use of context information, utilizing spatial pyramid pooling module and 3D-CNN. In optical flow, Dosovitskiy *et al.* [13] propose FlowNet based on U-Net architecture and is further improved by Ilg *et al.* [14]. Ummenhofer *et al.* [15] make use of [13] to iterate over optical flow, depth, and camera motion estimation. Afterwards, Sun *et al.* [5] introduce a lightweight architecture using the warped feature to construct cost volume.

B. Unsupervised Depth and Optical Flow Estimation

Numerous unsupervised learning methods have been proposed to alleviate the dependency on ground truth data. Early unsupervised approaches in optical flow estimation [6], [16] introduce smoothness loss and image reconstruction loss between the warped reference image and the target image. Some methods [4], [17] take occlusion into account to further improve flow estimation. In depth estimation, some works [7], [8] have proposed to learn disparity maps constrained by left-right consistency without supervision. Later, Chen *et al.* [18] suggest to use a patch-based ZNCC loss rather than a pixel-wise loss as a photometric loss, demonstrating it helps network converge to the global minimum. Zhan *et al.* [19] design a novel feature-based reconstruction loss utilizing the dense features as an alternative to the standard image reconstruction loss. Meanwhile, Zhong *et al.* [20] apply the c-LSTM [21] to the stereo video depth estimation to achieve the generalization ability to the open-world scenarios. Most recently, various approaches have proposed a joint learning framework for depth and optical flow estimation. Wang *et al.* [22] propose a joint framework for estimating camera pose and depth from monocular video using the c-LSTM, achieving temporally coherent results regardless of the length of input video sequences. Lai *et al.* [23] propose 2-Warp operations that warp an input image twice through both spatial and temporal axis. Wang *et al.* [24] further propose a unified framework for optical flow, depth, and camera pose estimation with various training losses to leverage the geometric consistency of each task jointly. Ranjan *et al.* [25] suggest to combine various task to constrain them by introducing a framework where the networks act as a competitor and a moderator.

C. Scene Flow Estimation

Scene flow estimation is a task of estimating dense flow in 3D given a pair of images, by jointly taking account of depths and optical flow given consecutive frames. As an early approach, Huguet *et al.* [11] suggest to jointly estimate both the 3D reconstruction and the scene flow by coupling the depth and optical flow. Later, Vogel *et al.* [26] propose to simultaneously estimate the depth and 3D motion field such that the estimation is view-consistent by enforcing consistency of the scene flow with respect to its all neighboring views rather than a reference frame. Along with the development of deep neural networks, Luo *et al.* [27] estimate scene flow with various consistency loss terms to supervise three parallel networks which estimate optical flow, depth, and camera pose, respectively. Most recently, Saxena *et al.* [28] propose to learn occlusion in a self-supervised manner, pointing out the importance of occlusion in scene flow estimation.

III. DEFINITION

For a stereo image sequence, S, consisting of N pairs of stereo images as $S = \{(I^L(t), I^R(t)) \,|\, 1 \le t \le N\}$, we can compute disparity and forward/backward motion fields as

$$\begin{aligned} &\{(\mathbf{D}^L(t), \mathbf{D}^R(t)) | 1 \le t \le N\}, \\ &\{(\mathbf{W}_F(t), \mathbf{W}_B(t+1)) | 1 \le t \le (N-1)\}, \end{aligned} \tag{1}$$

respectively. We call a set of the disparity and forward/backward motion fields of a stereo image sequence a *total field set*. $\mathbf{D}^L(t)$ and $\mathbf{D}^R(t)$ are the disparity fields of $I^L(t)$ and $I^R(t)$. Here, we assume that the stereo image pairs are all rectified. $\mathbf{W}_F(t)$ and $\mathbf{W}_B(t+1)$ are the forward/backward motion fields of $I(t)$ and $I(t+1)$, representing displacements of corresponding points from $I(t)$ to $I(t+1)$ and from $I(t+1)$ to $I(t)$, respectively. Then, we define a loop in a stereo image sequence with the total field set.

Definition *Loop* A *loop* from one point at $(\mathbf{x}, t, s)$, where $s \in \{L, R\}$[1], to another point at $(\mathbf{x}', t, s)$, denoted as $\mathscr{L}(\mathbf{x}, t, s)$, is an any ordered set of distinct points as $\{(\mathbf{x}_i, t_i, s_i) | 0 \le i \le n\}$, where $(\mathbf{x}_0, t_0, s_0) = (\mathbf{x}, t, s)$ and $(\mathbf{x}_n, t_n, s_n) = (\mathbf{x}', t, s)$, and $(t_i, s_i) \neq (t, s)$ for $1 \le i \le (n-1)$. Two adjacent coordinates in a loop, $(\mathbf{x}_i, t_i, s_i)$ and $(\mathbf{x}_{i+1}, t_{i+1}, s_{i+1})$, represent two corresponding points in different images defined by the disparity or motion field at $(\mathbf{x}_i, t_i, s_i)$.

We call $\mathbf{x}'$ a *terminal point* of $\mathscr{L}(\mathbf{x}, t, s)$ and represent it as $\mathbf{x}' = \mathscr{E}(\mathscr{L}(\mathbf{x}))$[2]. We then define the spatial and temporal consistency of disparity and motion fields, respectively.

Definition Two disparity fields $(\mathbf{D}^L(t), \mathbf{D}^R(t))$ are *spatially consistent up to* ε iff they satisfy the condition,

$$\| \mathscr{E}(\mathscr{L}(\mathbf{x})) - \mathbf{x} \| \le \varepsilon, \tag{2}$$

where $\mathscr{L}(\mathbf{x}, t, L)$ is defined as

$$\begin{aligned} \mathscr{L}(\mathbf{x}, t, L) = \{&(\mathbf{x}, t, L), (\mathbf{x} + \mathbf{D}^L(\mathbf{x}), t, R), \\ &(\mathbf{x} + \mathbf{D}^L(\mathbf{x}) + \mathbf{D}^R(\mathbf{x} + \mathbf{D}^L(\mathbf{x})), t, L)\} \end{aligned} \tag{3}$$

for all points visible from the both images. And two motion fields $(\mathbf{W}_F(t),\ \mathbf{W}_B(t+1))$ are *temporally consistent up to* ε iff they satisfy the condition Eq. (2), by replacing the notations in Eq. (3) with those of motion fields.

We then define the spatiotemporal consistency of a total field set by combining the spatial and temporal consistency.

Definition *Spatiotemporal Consistency* A total field is said to be *spatiotemporally consistent up to* ε iff the disparity and motion fields in it satisfy the condition Eq. (2) for any point at $\mathbf{x}$ with any arbitrary loop of $\mathbf{x}$ in a stereo image sequence.

[1]For example, $(\mathbf{x}, t, L)$ represents a point at $\mathbf{x}$ in $I^L(t)$.

[2]Here, we drop t and s since $\mathbf{x}$ and $\mathbf{x}'$ are in the same image, $I^s(t)$.

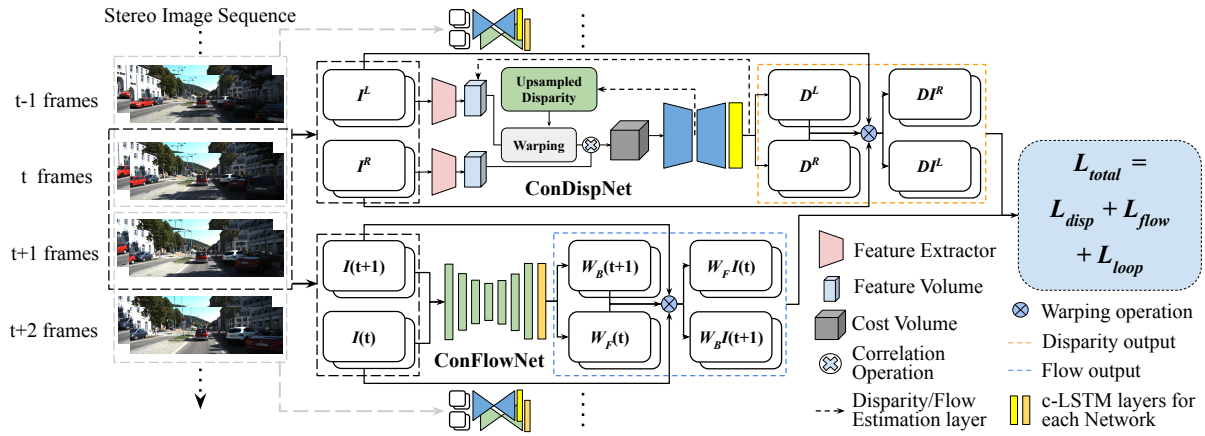


Fig. 1: Overview of our framework. Our network consists of two subnetworks, ConDispNet and ConFlowNet, that estimate the disparity and the optical flow, respectively. Please note that ConDispNet and ConFlowNet share the same architecture, but not weights. Using the estimated dense correspondence map from each subnetwork, the reference image is warped to the target image. DI^s and $WI(t)$ denote the warped image corresponding to the target image I^s and $I(t)$, respectively.

IV. PROPOSED APPROACH

In this section, we first present an overall framework, elaborating on each component enforcing spatiotemporal consistency in the framework. Then, we propose the loss functions for robust joint estimation of *accurate* and *consistent* disparity and optical flow estimation. Finally, we introduce a way to reason the occluded region.

A. Overall Framework

Network Architecture As illustrated in Fig. 1, our network consists of two subnetworks, ConDispNet and ConFlowNet, that estimate the disparity and the optical flow, respectively. Each subnetwork takes consecutive stereo image sequences as input and adopts PWC-Net [5] as the backbone architecture, as it is computationally light and shows good performance, though any other networks can be used. With the PWC-Net, ConDispNet estimates left and right disparity fields for two consecutive stereo pairs, and ConFlowNet calculates forward/backward flow, for left and right views, given a sequence of images. Since the PWC-Net was originally designed for optical flow estimation, we set the dual channel result of the PWC-Net to be the left and right disparity fields respectively for ConDispNet. Then, we modify our subnetwork architecture to be trained in the video domain by adding the c-LSTM module to the final estimation layer to leverage temporal features for consecutive estimation. In our subnetwork architecture, two pairs of images are fed into the siamese feature extractors. Then, the disparity/flow estimation layer estimates the disparity/motion field with the help of the c-LSTM module that is integrated into the final estimation layer.

Training Losses Our network is trained to optimize the total loss L_{total} consisting of three losses; the disparity loss L_{disp}, the flow loss L_{flow}, and the loop consistency loss L_{loop}. The disparity and flow losses are defined as follows:

$$L_{disp} = \lambda_{rec,d} L_{rec,d} + \lambda_{sm,d} L_{sm,d} + \lambda_{lr} L_{lr} \quad (4)$$

$$L_{flow} = \lambda_{rec,w} L_{rec,w} + \lambda_{sm,w} L_{sm,w} + \lambda_{fb} L_{fb} \quad (5)$$

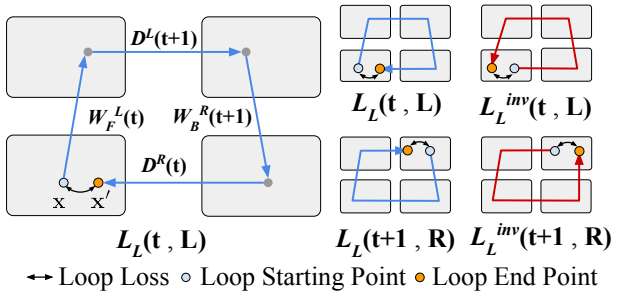


Fig. 2: Concept of the loop consistency loss. $\mathbf{x}$ denotes the starting position of loop and $\mathbf{x}'$ denote the end positions of the loops.

where L_{rec} and L_{sm} terms represent the image reconstruction loss and the edge-aware smoothness loss respectively, and the subscripts d and w indicate whether each term belongs to the disparity and flow loss. The hyperparameters $[\lambda_{rec,d}, \lambda_{rec,w}, \lambda_{sm,d}, \lambda_{sm,w}, \lambda_{lr}, \lambda_{fb}]$ are set to be $[1, 1, 5, 5, 0.5, 0.5]$. L_{lr} and L_{fb} represent the spatial and temporal consistency loss of the disparity and motion field, respectively. Each loss term is detailed in the following subsections. Our total loss is then defined as

$$L_{total} = \lambda_{disp} L_{disp} + \lambda_{flow} L_{flow} + \lambda_{loop} L_{loop} \quad (6)$$

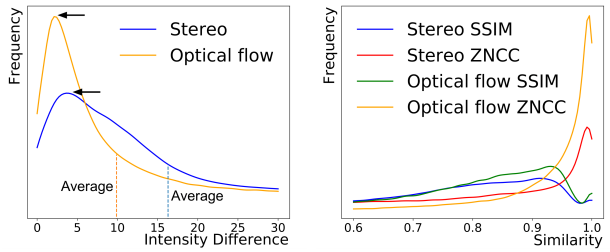
where $[\lambda_{disp}, \lambda_{flow}, \lambda_{loop}]$ are set to be [1, 1.5, 1.5].

B. Spatiotemporal Consistency Loss

Our network learns the spatiotemporal consistency among stereo image sequences and leverages the relations for *more accurate* and *consistent* optical flow and disparity estimation. In order to enforce consistency in each task, we apply temporal and spatial consistency losses. In addition to these two losses, we design a novel loop consistency loss as illustrated in Fig. 2 to enforce the spatiotemporal consistency for both optical flow and disparity estimation.

Spatial and Temporal Consistency Losses We apply temporal consistency loss L_{fb} so that the network estimates a more consistent motion field. It is defined as

$$L_{fb} = \sum_{\mathbf{x}} ||W_F(\mathbf{x}) + W_B(\mathbf{x} + W_F(\mathbf{x}))|| \quad (7)$$



(a) L1 norm (b) Patch similarity

Fig. 3: Distribution of (a) L1 norm and (b) SSIM/ZNCC scores for corresponding pixels, computed by using ground truth correspondence maps of the KITTI 2015 dataset. The arrows indicate the peak values.

By Sec. III, this loss enforces the motion field to be temporally consistent by aiming to make the forward motion fields and the corresponding backward motion fields reversed. By replacing the motion field with a disparity field, the temporal consistency loss L_{fb} becomes a spatial consistency loss L_{lr}. We apply two losses to each subnetwork ConFlowNet and ConDispNet, respectively.

Loop Consistency Loss We design a new loop consistency loss to reinforce the spatiotemporal consistency between consecutive stereo image pairs. By definition, the loop consistency loss $L_L(t,s)$ is defined as the displacement from the starting point $\mathbf{x}$ to the terminal point $\mathbf{x}'$:

$$L_L(t,s) = \sum_{\mathbf{x}} \alpha_x U(\|u' - u\| - T) + \alpha_y U(\|v' - v\| - T) \quad (8)$$

In this equation, the displacement both on the x-axis and y-axis are considered in optical flow estimation while only x-directional displacement is considered in disparity estimation. We alleviate the directional imbalances by setting balancing parameters α_x and α_y to be 1 and 0.2. U denotes the heaviside step function, and we set threshold T as 3 throughout experiments. To compute the loop consistency loss, we pick up the starting point $\mathbf{x}$ among two consecutive stereo image pairs and form a loop in two different directions, clockwise and counter-clockwise. For efficiency, we select the starting points only in two images (e.g., (t,L), $(t+1,R)$). The total loop consistency loss is then defined as

$$L_{loop} = \sum_{(t,s)} (L_L(t,s) + L_L^{inv}(t,s)) \quad (9)$$

where $\forall (t,s) \in \{(t,L),(t+1,R)\}$. $L_L(t,s)$ and $L_L^{inv}(t,s)$ denote loop consistency losses applied to $I^s(t)$, where the superscript *inv* denotes the loop in the inverse direction from the same starting point as illustrated in Fig. 2.

C. ZNCC-based Data Loss and Smoothness Loss

ZNCC-based Data Loss In unsupervised learning-based correspondence search, it is crucial to adopt a proper image reconstruction loss to compute the dissimilarity between the target image and the reconstructed one, since no ground truth correspondence maps are available. The combination of the L1-norm and the structural similarity index (SSIM) [29] has been commonly used as an image reconstruction loss. However, we demonstrate it is desirable to adopt the ZNCC rather than the SSIM as an unsupervised image reconstruction loss by experiments. Figure 3(a) shows the distributions of L1-norm of the brightness differences of stereo and optical flow pairs given by ground truth data. They have average values far from 0 (16.3 for stereo, 9.8 for flow), which is because stereo cameras are not identical in practice and the brightness can vary due to motion and temporal illumination changes. This indicates that correspondences may have intensity differences and the brightness constancy assumption with L1 norm alone can lead to performance degradation. On the other hand, Fig. 3(b) illustrates the distributions of the SSIM and ZNCC scores between stereo and optical flow pairs. The ZNCC score shows better tendency (clear peak closer to 1 and higher average) than the SSIM, since the ZNCC compensates the affine brightness changes and compares structural similarity more strictly. We therefore choose the ZNCC instead SSIM and consider the absolute brightness for the robust image reconstruction by adding the L1 norm to it. Final image reconstruction loss L_{rec} is defined as:

$$L_{ZNCC} = \frac{1}{2}\sum_{\mathbf{x}} 1 - \frac{(Y(\mathbf{x}) - \mu_Y)(\hat{Y}(\mathbf{x}) - \mu_{\hat{Y}})}{(\sigma_Y + \kappa)(\sigma_{\hat{Y}} + \kappa)} \quad (10)$$

$$L_{rec} = \lambda ||\hat{Y} - Y||_1 + (1-\lambda) \cdot L_{ZNCC} \quad (11)$$

where Y and $\hat{Y}$ denote the original and the reconstructed frame, respectively. A small number κ (set to 10^{-5}) is added to each term in the denominator to avoid division by zero. λ is a balancing parameter, empirically set to 0.1 and 0.05 for motion and disparity estimation, respectively.

Edge-aware Smoothness Loss Since the motion boundary usually coincides with the image edge, we apply the edge-aware smoothness loss introduced in [8].

$$L_{sm}(I,W,\alpha) = \sum_{p_i} \sum_{d \in (x,y)} \|\nabla_d^2 W(p_i)\| e^{-\alpha \|\nabla_d I(p_i)\|} \quad (12)$$

I, W, and α denote the image the loss is applied, the motion or disparity fields, and the image edge weight, respectively.

D. Occlusion Estimation

When applying the aforementioned loss functions, occluded or invisible pixels should be excluded during computation. To handle the occlusion and visibility issues during training loss computation, we estimate the possibly occluded pixels by checking the spatial or temporal consistency as in [4], [23]. For example, pixels are considered occluded in the optical flow estimation when they violate the forward-backward constraint as below:

$$\begin{aligned} \|W_F(\mathbf{x}) + W_B(\mathbf{x} + W_F(\mathbf{x}))\|^2 < \alpha_1 (\|W_F(\mathbf{x})\|^2 \\ + \|W_F(\mathbf{x}) + W_B(\mathbf{x} + W_F(\mathbf{x}))\|^2) + \alpha_2 \end{aligned} \quad (13)$$

where $\alpha_1 = 0.005$, $\alpha_2 = 1.0$. Then we define the occlusion mask $O_{\mathbf{x}}$ filtering out the possibly occluded pixels from data loss, which is set to be 1 to the occluded pixels, and 0 otherwise. This can be also regarded as a set of occluded pixels. The left-right consistency check is conducted in the same way by replacing the motion fields with the disparity fields. We perform this occlusion estimation for every consecutive input pair as a baseline for both optical flow and disparity estimation. We apply this occlusion mask to all the losses.

TABLE I: Quantitative evaluation of stereo depth estimation task on the KITTI 2015 stereo dataset. "R" and "P" denote the ResNet and PWC-Net [5] version implementations, respectively, in [23]. The boldface and underscore denote the best and second best performances, respectively. The same notation and typography are applied to the following tables.

Method	Train Stereo	Test Stereo	Use Flow	Use Ego-motion	Lower the better					Higher the better		
					Abs rel	Sq Rel	RMSE	RMSE log	D1-all	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zhou *et al.* [7]	✓	✓			-	-	-	-	9.41%	-	-	-
Monodepth [8]	✓	✓			0.068	0.835	4.392	0.146	9.194%	0.942	0.978	0.989
Zhong *et al.* [30]	✓	✓			0.075	1.726	4.857	0.165	6.424%	0.956	0.976	0.985
Bridging-R [23]	✓	✓	✓		0.062	0.747	4.113	0.146	-	0.948	0.979	0.990
Bridging-P [23]	✓	✓	✓		0.058	0.694	4.020	0.152	-	0.952	0.979	0.990
UnOS [24]	✓	✓	✓	✓	0.049	0.515	**3.404**	0.121	**5.943%**	**0.964**	**0.984**	**0.992**
Ours(Full w/o c-LSTM)	✓	✓	✓		**0.048**	0.451	3.470	0.122	6.484%	0.963	**0.984**	**0.992**
Ours(Full)	✓	✓	✓		0.049	**0.443**	**3.404**	**0.119**	6.721%	0.963	**0.984**	**0.992**

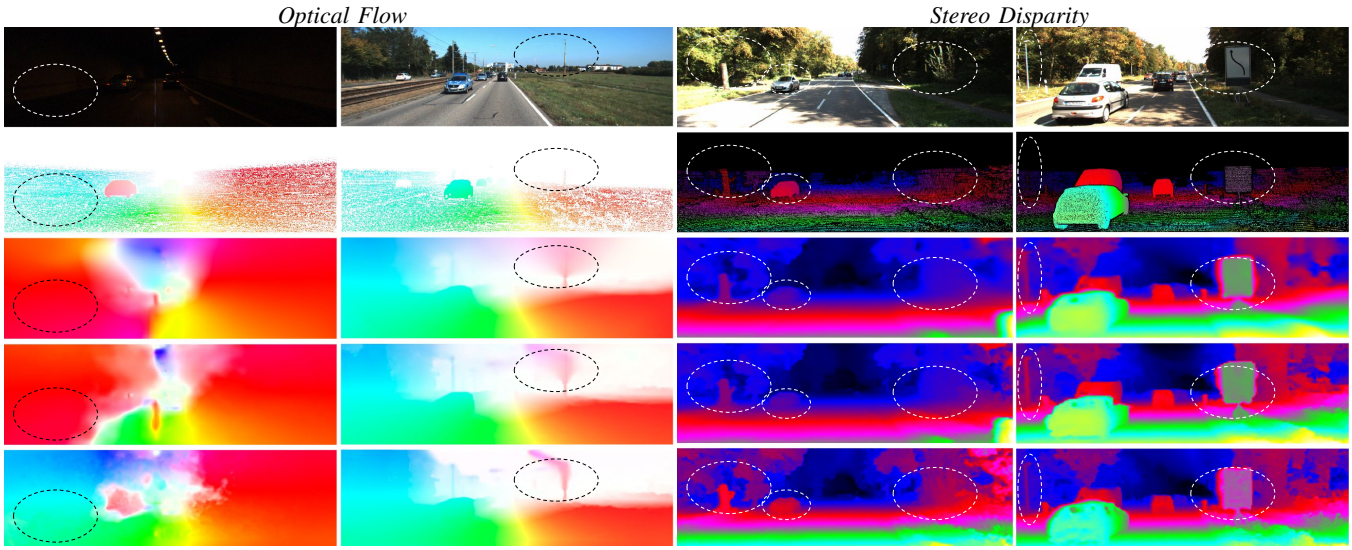


Fig. 4: Qualitative comparison of our method with other SOTA methods on the KITTI 2015 dataset. *First row*: Input image. *Second row*: Ground truth. *Third row*: Estimated result of Bridging [23]. *Fourth row*: Estimated result of UnOS [24]. *Last row*: Estimated result of ours.

TABLE II: Quantitative evaluation of the stereo depth estimation task on the sampled KITTI raw video clips.

Method	Unsupervised	Test Stereo	Lower the better		Higher the better
			Abs Rel	Sq Rel	$\delta < 1.25$
Monodepth	✓	✓	0.0712	0.6877	0.947
Bridging-R [23]	✓	✓	0.0683	0.6885	0.951
Bridging-P [23]	✓	✓	0.0627	0.5857	0.954
UnOS [24]	✓	✓	0.0592	0.4707	0.954
Ours(Full)	✓	✓	**0.0561**	**0.4432**	**0.962**

TABLE III: Quantitative evaluation of the optical flow task on the KITTI 2015 dataset. "noc" denote non-occlusion regions.

Method	Joint learning	Use ego-motion	EPE -all	F1 -all	EPE -noc
UnFlow-C [4]			8.80	28.94%	-
GeoNet [31]	✓	✓	10.81	-	8.05
Wang *et al.* [32]			8.88	-	-
Janai *et al.* [33]			6.59	-	3.22
DFnet [34]	✓	✓	8.98	26.01%	-
CC [25]	✓	✓	6.21	21.50%	-
Bridging-R [23]	✓		7.02	27.34%	4.26
Bridging-P [23]	✓		6.66	21.05%	3.60
UnOS [24]		✓	**5.58**	-	3.79
Ours (Full w/o c-LSTM)	✓		6.55	19.47%	**3.11**
Ours (Full)	✓		6.34	**19.19%**	3.38

V. EXPERIMENTAL RESULTS

We evaluate our framework on the KITTI 2015 and KITTI raw dataset, and compare with state-of-the-art unsupervised learning-based methods both quantitatively and qualitatively.

A. Implementation Details

We implement our framework using PyTorch v.0.4.1. During training, the video clip is randomly sampled with the temporal window size 3 and each frame is fed into the network after scaled to the size of 1280×384. We use Adam optimizer with the parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set as 10^{-4} and we apply learning rate decay. We randomly shift gamma, brightness, and color with 50% probability as a way of data augmentation.

B. Datasets

Datasets for Training We train our network with the KITTI raw dataset [35] to jointly learn both stereo disparity and optical flow. Following an evaluation protocol with previous work [8], [23], [24], we sample the training dataset excluding the scene included in the training set of the KITTI 2015.

Datasets for Single Image Pair Evaluation To evaluate the accuracy of disparity or flow estimation on a pair of correspondence maps, we use the KITTI 2015 dataset that consists of 200 pairs of the ground truth optical flow and disparity maps.

Datasets for Video Evaluation To evaluate the performance in stereo video depth estimation, we randomly sample two video clips from the KITTI raw dataset that are not included in the training dataset.

C. Evaluation Metrics

Accuracy Metrics We evaluate our network using standard evaluation metrics for depth evaluation as follows: Absolute relative error(abs rel), squared relative error(sq rel), root mean squared error(RMSE), root mean squared logarithmic error(RMSE log), D1-all, and $\delta < [1.25, 1.25^2, 1.25^3]$. δ is defined as $max(\frac{d_i}{g_i}, \frac{g_i}{d_i})$ and the percentage of d_i is measured. d_i and g_i are the estimated and ground truth depths of pixel i, respectively. To evaluate optical flow estimation, we use two widely-used metrics: average endpoint error (EPE) and percentage of erroneous pixels(Fl-all).

Spatiotemporal Consistency Metrics Based on the definitions in Sec. III, the spatiotemporal consistency is measured as the displacement from the starting point $\mathbf{x}$ to the terminal point $\mathbf{x}'$ in a loop, while each of spatial and temporal consistency is measured only between the two motion or disparity fields. Therefore, simply combining the two consistencies is far from the general concept of the spatiotemporal consistency itself. To quantify the spatiotemporal consistency, we design a new metric, loop consistency (LC), defined as:

$$LC = \frac{1}{N}\sum_{k=1}^{N}\sum_{\mathbf{x}\in(u,v)} ||\mathbf{x}' - \mathbf{x}||_1 \tag{14}$$

where $\mathbf{x}$ and $\mathbf{x}'$ follow the notation explained in Sec. III. Please note that the occlusion mask is not taken into account in the evaluation since there is no dataset with ground truth occlusion mask between the consecutive stereo sequence.

D. Experimental Evaluation

Stereo Disparity Estimation We evaluate the stereo disparity estimation of our network on both stereo images and stereo videos and compare the results with those of state-of-the-art models. Please note that both UnOS [24] and Bridging-P [23] use the PWC-Net [5] as a baseline architecture, same as ours. As shown in the Table. I, ours outperforms Bridging-P [23] since their 2-warp loss is ineffective in the photometrically homogeneous regions. In contrast, our loop consistency loss, which is based on the displacement on the spatial and temporal axis, helps accurate depth estimation in such regions. In Table. II, ours outperforms other state-of-the-art unsupervised methods on the video dataset since we take advantage of the temporal features from the sampled video clip during training via the c-LSTM and loop consistency loss. The qualitative depth estimation comparison is shown in the third and fourth columns of Fig.4. Our network yields sharper boundaries and more detailed background than other approaches, as indicated by the ellipse. Furthermore, ours yields accurate depth estimation in the shadowed areas since the ZNCC-based data loss compensates for the affine brightness difference in such areas.

Optical Flow Estimation We conduct the evaluation of the optical flow estimation on the KITTI 2015 training dataset and compare it with that of state-of-the-art methods. In Table. III, our model records the lowest F1 score and the second-lowest EPE in non-occluded regions. Ours shows much lower EPE-noc than Bridging-P [23], which is because the proposed loop consistency loss has the significant effects in non-occluded regions. Compared to UnOS [24], ours shows lower EPE-noc but larger EPE-all. We reason the use of ego-motion in UnOS [24] improves the estimation in occluded regions. When qualitatively compared to [23], [24], ours shows a more detailed flow estimation at object boundaries, as shown in the first and second columns of Fig. 4. In particular, the first column indicates other SOTA methods with SSIM as an image reconstruction loss are almost incorrect when the illumination condition is extremely bad. We can infer our network is trained to find correspondence using structural similarity that compensates for abrupt illumination change.

TABLE IV: Ablation study on the KITTI 2015 dataset.

Patch Similarity	Loop loss	c-LSTM	Stereo matching		Optical flow		
			Abs rel	Sq rel	EPE-all	EPE Noc	EPE-occ
SSIM			0.0537	0.6152	7.442	3.837	20.73
ZNCC			0.0528	0.5778	7.308	3.801	20.73
ZNCC	✓		**0.0481**	0.4509	6.551	**3.109**	20.11
ZNCC	✓	✓	0.0489	**0.4432**	**6.338**	3.384	**17.62**

Spatiotemporal Consistency To evaluate the spatiotemporal consistency, we use consecutive frames of stereo pair as mentioned in Sec. V-B. We evaluate the ratio of true-positive/positive case and mean/median of LC. Here, true indicates the error of optical flow and disparity comes within 3 pixels based on the reference image, and the positive case represents the LC comes within a specific threshold. For a fair comparison, we conduct an evaluation with the pre-trained model specified in the original papers [23], [24]. In Table VII, Bridging-P [23] shows higher LC than that of UnOS [24] since it considers spatiotemporal consistency from the 2-Warp consistency loss in table VII. In true-positive case, the overall result of Bridging-P is lower than UnOS [24]. The reason is the ratio of true in UnOS is higher than bridging-P [23] since the accuracy of UnOS is impressive. On the other hand, ours records the highest in all consistency metrics among other methods. It means our framework enforces the spatiotemporal consistency more effectively than [23], [24] through the c-LSTM and loop consistency loss. We can infer our network shows good results when evaluating accuracy and consistency through the evaluation of the true-positive ratio.

E. Ablation Study

Patch Similarity Function We demonstrate the result of ablation study on patch similarity function in Table. IV. Comparing the first two rows in the stereo matching column, we demonstrate the model trained with the ZNCC-based data loss shows more accurate disparity estimation than that with the SSIM. We attribute this improvement to the ZNCC-based data loss compensating for affine brightness changes from the different stereo camera settings. In optical flow, the ZNCC-based data loss also helps resolve the brightness difference arising from the motion and temporal illumination changes.

Loop Consistency Loss Comparison of the second and the third row in stereo matching column indicates the model trained with the loop consistency loss shows an significant performance improvement on square relative error; this suggests that some of the larger disparity errors are reduced. We

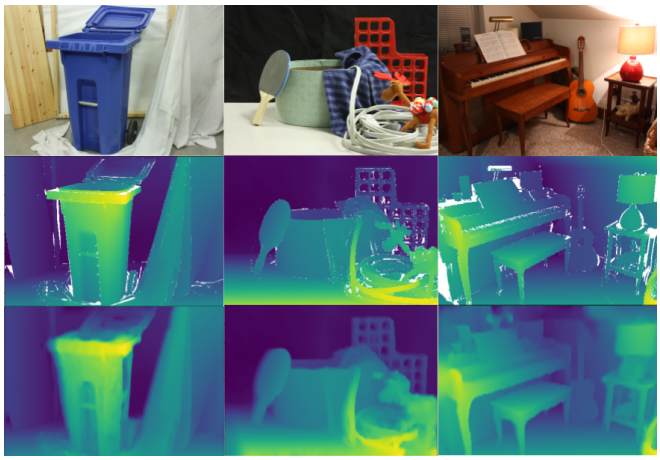


Fig. 5: Our qualitative results of stereo disparity estimation results on Middlebury dataset(2014): First row: left input image, Second row: ground truth disparity, Third row: estimated disparity results

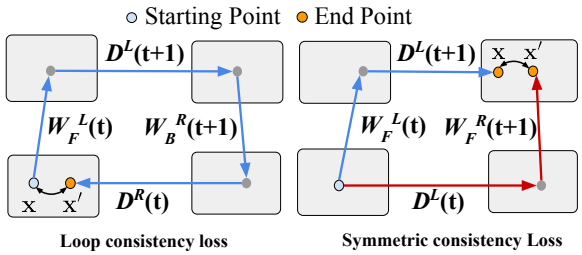


Fig. 6: Two types of consistency loss. The same notation are applied as fig. 2

also see that the loop consistency loss significantly improves the precision of depth estimation from the results. In the case of optical flow estimation, the loop loss corrects flow fields that do not fit in the loop, reducing the overall error range and giving consistency. However, it improves the EPE-occ marginally since it is applied only to the non-occluded region obtained by the method in Sec IV-B.

Video-based Training Scheme and c-LSTM We also conduct ablation study on the c-LSTM module which helps learn temporal dynamics in the video. The effect of the c-LSTM, temporal smoothing, is apparent as the last two rows of the metric "Sq rel" shows, which reflects the significant errors. Similarly, EPE-occ has improved in optical flow estimation thanks to the c-LSTM module. However, pixel accuracy "abs rel" in stereo and "EPE-noc" in flow are not improved. We reason the c-LSTM is effective for alleviating large errors while the performance is dominantly determined by the main loss functions(ZNCC-based loss, loop consistency loss).

Temporal window size At the training phase, we randomly sample the video clips by the given temporal window size. We experiment to find out the aspect of the performance according to the different temporal window sizes. As you can see in the table V, the performances show that feeding batch with large temporal windows(> 10) to network quickly cause over-fitting since generalization ability in various scenarios is lower than without the small number of temporal windows. On the other hand, the result shows our network can successfully handle arbitrary temporal window size less than seven since there is a slight performance difference both in optical flow and in stereo matching.

TABLE V: Ablation study on temporal window size. For a fair comparison, we apply both proposed loop consistency loss and ZNCC-based data loss to all models. The boldface denotes the best performance.

Temporal window size	Optical flow			Stereo matching		
	EPE-all	F1-all(%)	EPE-noc	Abs rel	Sq rel	RMSE
3	**6.338**	**19.19**	3.380	**0.0478**	**0.4432**	**3.404**
7	6.378	19.62	**3.307**	0.0509	0.5461	3.614
10	6.775	19.72	3.639	0.0527	0.5864	3.631
15	7.431	20.73	4.033	0.0520	0.6608	3.485

TABLE VI: Ablation study on two types of consistency losses.

Consistency loss	Stereo matching		Optical flow		
	Abs rel	Sq rel	EPE-all	EPE-noc	F1-all(%)
Loop	0.049	0.443	6.338	3.380	19.19
Symmetric	0.050	0.527	6.762	3.442	19.67

TABLE VII: Quantitative evaluation of spatiotemporal consistency on the KITTI 2015 dataset. ξ is a set of LC(loop consistency). μ, *med* denote mean and median, respectively. The boldface and underscore denote the best performances of positive samples and true-positive samples, respectively.

Method	True Positive	Posi-tive	Higher the better			Lower the better	
			$\xi < 2$	$\xi < 3$	$\xi < 4$	$\mu(\xi)$	$med(\xi)$
Bridging-R [23]	✓		0.5916	0.6300	0.6484	-	-
		✓	0.7027	0.7655	0.8017	4.9295	1.3659
Bridging-P [23]	✓		0.6720	0.6955	0.7085	-	-
		✓	0.7412	0.7868	0.8150	4.4213	0.9652
UnOS [24]	✓		0.6789	0.7088	0.7242	-	-
		✓	0.7156	0.7703	0.8011	6.3182	1.2234
Ours(Symmetric)	✓		0.7091	0.7290	0.7390	-	-
		✓	0.7468	0.7952	0.8262	4.3221	0.9601
Ours(Loop)	✓		0.7129	0.7346	0.7451	-	-
		✓	**0.7526**	**0.8017**	**0.8322**	**3.5098**	**0.8554**

Various type of consistency loss In addition to loop consistency loss, we can formulate various type of spatio-temporal consistency loss for each edge. As illustrated in Fig. 6, we conducted an ablation study on two types(loop, symmetric) of consistency loss for comparison. As table VI indicates, our loop consistency loss helps to estimate both depth and optical flow more accurately although a symmetric consistency loss seems simpler. Moreover, the result of table VII shows our proposed consistency loss is better in terms of consistency as well as accuracy.

F. Generalization ability to different dataset

To experiment with the generalization ability of unfamiliar data, we test conDispNet to open stereo datasets different from the training dataset. Since the training domain of our framework is the driving scene, we performed qualitative evaluation using the Middlebury(2014) [36], not driving scene. Please note that our conDispNet has never seen data from the test set.

Datasets for finetuning For better generalization, we finetuned the network pretrained in the KITTI dataset [35] to the sceneflow dataset(Flyingthing3D) [37] for 1 epoch. We don't use ground truth and only use the consecutive stereo image of the flyingthing3D dataset in the finetuning process.

Middlebury dataset The Middlebury dataset [38] are indoor dataset with various scenes. We illustrate our qualitative examples in Fig. 5. Despite different locations, cameras, conDispNet produced visually good results.

VI. Conclusion

In this paper, we presented unsupervised learning framework for *accurate* and *consistent* optical flow and disparity estimation of stereo videos. We demonstrated that the novel loop consistency loss and the proposed video training scheme using network architecture with the c-LSTM module not only improve accuracy but also maintain spatiotemporal consistencies. Experimental results on the KITTI benchmark datasets show that our framework successfully maintains temporal consistency and also achieves the significant accuracy improvement both in optical flow and disparity estimation.

References

[1] J. Zbontar, Y. LeCun, *et al.*, "Stereo matching by training a convolutional neural network to compare image patches." *Journal of Machine Learning Research*, vol. 17, no. 1-32, p. 2, 2016.

[2] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *ICCV*, 2017, pp. 66–75.

[3] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *CVPR*, 2018, pp. 5410–5418.

[4] S. Meister, J. Hur, and S. Roth, "Unflow: Unsupervised learning of optical flow with a bidirectional census loss," in *AAAI*, 2018.

[5] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *CVPR*, 2018, pp. 8934–8943.

[6] J. Y. Jason, A. W. Harley, and K. G. Derpanis, "Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness," in *ECCV*. Springer, 2016, pp. 3–10.

[7] C. Zhou, H. Zhang, X. Shen, and J. Jia, "Unsupervised learning of stereo matching," in *ICCV*, 2017, pp. 1567–1575.

[8] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*, 2017.

[9] C. Fookes, A. Maeder, S. Sridharan, and J. Cook, "Multi-spectral stereo image matching using mutual information," in *3DPVT*. IEEE, 2004, pp. 961–968.

[10] B. K. Horn and B. G. Schunck, "Determining optical flow," in *Techniques and Applications of Image Understanding*, vol. 281. International Society for Optics and Photonics, 1981, pp. 319–331.

[11] F. Huguet and F. Devernay, "A variational method for scene flow estimation from stereo sequences," in *ICCV*. IEEE, 2007, pp. 1–7.

[12] T. Taniai, S. N. Sinha, and Y. Sato, "Fast multi-frame stereo scene flow with motion segmentation," in *CVPR*, 2017, pp. 3939–3948.

[13] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *ICCV*, 2015, pp. 2758–2766.

[14] 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 *CVPR*, 2017, pp. 2462–2470.

[15] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "Demon: Depth and motion network for learning monocular stereo," in *CVPR*, 2017, pp. 5038–5047.

[16] Z. Ren, J. Yan, B. Ni, B. Liu, X. Yang, and H. Zha, "Unsupervised deep learning for optical flow estimation," in *AAAI*, 2017.

[17] Y. Wang, Y. Yang, Z. Yang, L. Zhao, P. Wang, and W. Xu, "Occlusion aware unsupervised learning of optical flow," in *CVPR*, 2018, pp. 4884–4893.

[18] L. Chen, W. Tang, and N. W. John, "Self-supervised monocular image depth learning and confidence estimation," *CoRR*, vol. abs/1803.05530, 2018. [Online]. Available: http://arxiv.org/abs/1803.05530

[19] H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *CVPR*, 2018, pp. 340–349.

[20] Y. Zhong, H. Li, and Y. Dai, "Open-world stereo video matching with deep rnn," in *ECCV*, 2018, pp. 101–116.

[21] X. Shi, Z. Chen, H. Wang, D.-Y. Yeg, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *NIPS*, 2015, pp. 802–810.

[22] R. Wang, S. M. Pizer, and J.-M. Frahm, "Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth," in *CVPR*, 2019, pp. 5555–5564.

[23] H.-Y. Lai, Y.-H. Tsai, and W.-C. Chiu, "Bridging stereo matching and optical flow via spatiotemporal correspondence," in *CVPR*, 2019, pp. 1890–1899.

[24] Y. Wang, P. Wang, Z. Yang, C. Luo, Y. Yang, and W. Xu, "Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos," in *CVPR*, 2019, pp. 8071–8081.

[25] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, and M. J. Black, "Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation," in *CVPR*, 2019, pp. 12 240–12 249.

[26] C. Vogel, S. Roth, and K. Schindler, "View-consistent 3d scene flow estimation over multiple frames," in *ECCV*. Springer, 2014, pp. 263–278.

[27] C. Luo, Z. Yang, P. Wang, Y. Wang, W. Xu, R. Nevatia, and A. Yuille, "Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding," *arXiv preprint arXiv:1810.06125*, 2018.

[28] R. Saxena, R. Schuster, O. Wasenmuller, and D. Stricker, "Pwoc-3d: Deep occlusion-aware end-to-end scene flow estimation," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 324–331.

[29] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[30] Y. Zhong, Y. Dai, and H. Li, "Self-supervised learning for stereo matching with self-improving ability," *arXiv preprint arXiv:1709.00930*, 2017.

[31] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *CVPR*, 2018, pp. 1983–1992.

[32] R. Wang, M. Schworer, and D. Cremers, "Stereo dso: Large-scale direct sparse visual odometry with stereo cameras," in *ICCV*, 2017, pp. 3903–3911.

[33] J. Janai, F. Guney, A. Ranjan, M. Black, and A. Geiger, "Unsupervised learning of multi-frame optical flow with occlusions," in *ECCV*, 2018, pp. 690–706.

[34] Y. Zou, Z. Luo, and J.-B. Huang, "Df-net: Unsupervised joint learning of depth and flow using cross-task consistency," in *ECCV*, 2018.

[35] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *IJRR*, vol. 32, no. 11, pp. 1231–1237, 2013.

[36] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German conference on pattern recognition*. Springer, 2014, pp. 31–42.

[37] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *CVPR*, 2016, pp. 4040–4048.

[38] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1-3, pp. 7–42, 2002.